Multi-agent deep reinforcement learning: a survey

多智能体深度强化学习: 综述

Sven Gronauer [1]®· Klaus Diepold1
Sven Gronauer [1]®· Klaus Diepold1
Published online: 15 April 2021
在线发表:2021 年 4 月 15 日

Abstract

摘要

The advances in reinforcement learning have recorded sublime success in various domains. Although the multi-agent domain has been overshadowed by its single-agent counterpart during this progress, multi-agent reinforcement learning gains rapid traction, and the latest accomplishments address problems with real-world complexity. This article provides an overview of the current developments in the field of multi-agent deep reinforcement learning. We focus primarily on literature from recent years that combines deep reinforcement learning methods with a multi-agent scenario. To survey the works that constitute the contemporary landscape, the main contents are divided into three parts. First, we analyze the structure of training schemes that are applied to train multiple agents. Second, we consider the emergent patterns of agent behavior in cooperative, competitive and mixed scenarios. Third, we systematically enumerate challenges that exclusively arise in the multi-agent domain and review methods that are leveraged to cope with these challenges. To conclude this survey, we discuss advances, identify trends, and outline possible directions for future work in this research area.

强化学习在各个领域的进展取得了卓越的成果。尽管在此过程中，多智能体领域被其单一智能体 counterpart 抢尽风头，但多智能体强化学习正在迅速获得关注，最新的成果解决了具有现实世界复杂性的问题。本文提供了多智能体深度强化学习领域的当前发展的概述。我们主要关注近年来将深度强化学习方法与多智能体场景相结合的文献。为了调查构成当代景观的作品，主要内容分为三部分。首先，我们分析应用于训练多个智能体的训练方案的结构。其次，我们考虑在合作、竞争和混合场景中智能体行为的涌现模式。第三，我们系统地列举了仅在多智能体领域中出现的挑战，并回顾了用于应对这些挑战的方法。为了总结本综述，我们讨论进展，识别趋势，并概述了该研究领域未来工作的可能方向。

Keywords Multi-agent systems - Multi-agent learning - Machine learning - Reinforcement learning - Deep learning - Survey

关键词多智能体系统 - 多智能体学习 - 机器学习 - 强化学习 - 深度学习 - 综述

1 Introduction

1 引言

A multi-agent system describes multiple distributed entities-so-called agents-which take decisions autonomously and interact within a shared environment (Weiss 1999). Each agent seeks to accomplish an assigned goal for which a broad set of skills might be required to build intelligent behavior. Depending on the task, an intricate interplay between agents can occur such that agents start to collaborate or act competitively to excel opponents. Specifying intelligent behavior a-priori through programming is a tough, if not impossible, task for complex systems. Therefore, agents require the ability to adapt and learn over time by themselves. The most common framework to address learning in an interactive environment is reinforcement learning (RL), which describes the change of behavior through a trial-and-error approach.

多代理系统描述了多个分布式实体，即所谓的代理，它们自主做出决策并在共享环境中互动 (Weiss 1999)。每个代理都试图完成分配给它的目标，这可能需要广泛的技能来构建智能行为。根据任务的不同，代理之间可能会发生复杂的相互作用，从而使代理开始协作或竞争以超越对手。对于复杂系统，通过编程预先指定智能行为是一项艰巨的任务，甚至可能是不可能的。因此，代理需要具备随时间自我适应和

学习的能力。在交互环境中处理学习的最常见框架是强化学习 (RL)，它描述了通过尝试和错误方法来改变行为。

The field of reinforcement learning is currently thriving. Since the breakthrough of deep learning methods, works have been successful at mastering complex control tasks, e.g. in robotics (Levine et al. 2016; Lillicrap et al. 2016) and game playing (Mnih et al. 2015; Silver et al. 2016). The key to these results is based on learning techniques that employ neural networks as function approximators (Arulkumaran et al. 2017). Despite these achievements, the majority of works investigated single-agent settings only, although many real-world applications naturally comprise multiple decision-makers that interact at the same time. The areas of application encompass the coordination of distributed systems (Cao et al. 2013; Wang et al. 2016b) such as autonomous vehicles (Shalev-Shwartz et al. 2016) and multi-robot control (Matignon et al. 2012a), the networking of communication packages (Luong et al. 2019), or the trading on financial markets (Lux and Marchesi 1999). In these systems, each agent discovers a strategy alongside other entities in a common environment and adapts its policy in response to the behavioral changes of others. Carried by the advances of single-agent deep RL, the multi-agent reinforcement learning (MARL) community has been surged with new interest and a plethora of literature has emerged lately (Hernandez-Leal et al. 2019; Nguyen et al. 2020). The use of deep learning methods enabled the community to exceed the historically investigated tabular problems to challenging problems with real-world complexity (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019; Vinyals et al. 2019).

强化学习领域目前正处于繁荣状态。自从深度学习方法取得突破以来，研究工作在掌握复杂控制任务方面取得了成功，例如在机器人学 (Levine et al. 2016；Lillicrap et al. 2016) 和游戏 (Mnih et al. 2015；Silver et al. 2016) 领域。这些成果的关键是基于使用神经网络作为函数逼近的学习技术 (Arulkumaran et al. 2017)。尽管取得了这些成就，但大多数研究仅探讨了单一智能体设置，尽管许多现实世界应用自然包含同时互动的多个决策者。应用领域包括分布式系统的协调 (Cao et al. 2013；Wang et al. 2016b)，如自动驾驶车辆 (Shalev-Shwartz et al. 2016) 和多机器人控制 (Matignon et al. 2012a)，通信包的网络化 (Luong et al. 2019)，或者金融市场的交易 (Lux 和 Marchesi 1999)。在这些系统中，每个智能体在共同环境中与其他实体一起发现策略，并适应其他实体的行为变化。随着单一智能体深度 RL 的进步，多智能体强化学习 (MARL) 社区已经焕发新的兴趣，最近出现了大量的文献 (Hernandez-Leal et al. 2019；Nguyen et al. 2020)。深度学习方法的使用使社区能够超越历史上研究的表格问题，转向具有现实世界复杂性的挑战性问题 (Baker et al. 2020；Berner et al. 2019；Jaderberg et al. 2019；Vinyals et al. 2019)。

In this paper, we provide an extensive review of the recent advances in the area of multi-agent deep reinforcement learning (MADRL). Although multi-agent systems enjoy a rich history (Busoniu et al. 2008; Shoham et al. 2003; Stone and Veloso 2000; Tuyls and Weiss 2012), this survey aims to shed light on the contemporary landscape of the literature in

在本文中，我们对多智能体深度强化学习 (MADRL) 领域的近期进展进行了广泛的回顾。尽管多智能体系统拥有丰富的历史 (Busoniu 等人，2008 年；Shoham 等人，2003 年；Stone 和 Veloso，2000 年；Tuyls 和 Weiss，2012 年)，本次调查旨在揭示当代文献的景观。MADRL.

1.1 Related work

1.1 相关工作

The intersection of multi-agent systems and reinforcement learning holds a long record of active research. As one of the first surveys in the field, Stone and Veloso (2000) analyzed multi-agent systems from a machine learning perspective and classified the reviewed literature according to heterogeneous and homogeneous agent structures as well as communication skills. The authors discussed issues associated with each classification. Shoham et al. (2003) criticized the ill-posed problem statement of MARL which is in the authors' opinion unclear and called for more grounded research. They proposed a coherent research

☒ Sven Gronauer
☒ Sven Gronauer
sven.gronauer@tum.de
Klaus Diepold
Klaus Diepold
kldi@tum.de
1 Department of Electrical and Computer Engineering, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany
1 电子与计算机工程系，慕尼黑工业大学 (TUM)，Arcisstr. 21, 80333 慕尼黑，德国

agenda which includes four directions for future research. Yang and Gu (2004) reviewed algorithms and pointed out that the main difficulty lies in the generalization to continuous action and state spaces and in the scaling to many agents. Similarly, Busoniu et al. (2008) presented selected algorithms and discussed benefits as well as challenges of MARL. Benefits include computational speed-ups and the possibility of experience sharing between agents. In contrast, drawbacks are the specification of meaningful goals, the non-stationarity of the environment, and the need for coherent coordination in cooperative games. In addition to that, they posed challenges such as the exponential increase of computational complexity with the number of agents and the alter-exploration problem where agents must gauge between the acquisition of new knowledge and the exploitation of current knowledge. More specifically, Matignon et al. (2012b) identified challenges for the coordination of independent learners that arise in fully cooperative Markov Games such as non-stationarity, stochasticity, and shadowed equilibria. Further, they analyzed conditions under which algorithms can address such coordination issues. Another work by Tuyls and Weiss (2012) accounted for the historical developments of MARL and evoked non-technical challenges. They criticized that the intersection of RL techniques and game theory dominates multi-agent learning, which may render the scope of the field too narrow and investigations are limited to simplistic problems such as grid worlds. They claimed that the scalability to high numbers of agents and large and continuous spaces are the holy grail of this research domain.

多智能体系统与强化学习的交叉领域拥有长久活跃的研究历史。作为该领域最早的调研之一，Stone 和 Veloso(2000 年) 从机器学习的角度分析了多智能体系统，并根据异构和同构智能体结构以及通信技能对所评文献进行了分类。作者们讨论了与每种分类相关的问题。Shoham 等人 (2003 年) 批评了多智能体强化学习 (MARL) 不明确的问题描述，并呼吁进行更扎实的研究。他们提出了连贯的研究计划，包括未来研究的四个方向。杨和顾 (2004 年) 回顾了算法，并指出主要的困难在于向连续动作和状态空间的泛化以及扩展到许多智能体。同样，Busoniu 等人 (2008 年) 展示了精选算法，并讨论了 MARL 的益处以及挑战。益处包括计算速度提升以及智能体之间可能的经验共享。相比之下，缺点包括指定有意义的目标、环境的非平稳性以及在合作游戏中需要进行协调一致。除此之外，他们还提出了挑战，例如随着智能体数量的增加，计算复杂度呈指数级增长，以及交替探索问题，其中智能体必须在获取新知识与利用现有知识之间进行衡量。更具体地说，Matignon 等人 (2012b 年) 确定了在完全合作马尔可夫博弈中独立学习者协调所面临的挑战，例如非平稳性、随机性和遮蔽均衡。此外，他们分析了算法可以解决这些协调问题的条件。Tuyls 和 Weiss(2012 年) 的另一项工作回顾了 MARL 的历史发展，并提出了非技术性挑战。他们批评了强化学习技术与博弈论的交叉在多智能体学习领域占据主导地位，这可能导致该领域的范围过于狭窄，研究仅限于像格子世界这样的简单问题。他们声称，扩展到大量智能体和大而连续的空间是这一研究领域的终极目标。

Since the advent of deep learning methods and the breakthrough of deep RL, the field of MARL has attained new interest and a plethora of literature has emerged during the last years. Nguyen et al. (2020) presented five technical challenges including nonstationarity, partial observability, continuous spaces, training schemes, and transfer learning. They discussed possible solution approaches alongside their practical applications. Hernandez-Leal et al. (2019) concentrated on four categories including the analysis of emergent behaviors, learning communication, learning cooperation, and agent modeling. Further survey literature focuses on one particular sub-field of MADRL. Oroojlooyjadid and Hajinezhad (2019) reviewed recent works in the cooperative setting while Da Silva and Costa (2019) and Da Silva et al. (2019) focused on knowledge reuse. Lazaridou and Baroni (2020) reviewed the emergence of language and connected two perspectives, which comprise the conditions under which language evolves in communities and the ability to solve problems through dynamic communication. Based on theoretical analysis, Zhang et al. (2019) focused on MARL algorithms and presented challenges from a mathematical perspective.

自从深度学习方法的出现和深度强化学习的突破以来，多智能体强化学习领域获得了新的关注，在过去几年里涌现出了大量的文献。Nguyen 等人 (2020) 提出了五个技术挑战，包括非平稳性、部分可观测性、连续空间、训练方案和迁移学习。他们讨论了可能的解决方案方法及其在实际应用中的伴随。Hernandez-Leal 等人 (2019) 专注于四个类别，包括新兴行为的分析、学习通信、学习合作和智能体建模。进一步的研究文献聚焦于多智能体深度强化学习的一个特定子领域。Oroojlooyjadid 和 Hajinezhad(2019) 回顾了合作环境中的近期工作，而 Da Silva 和 Costa(2019) 以及 Da Silva 等人 (2019) 则关注知识重用。Lazaridou 和 Baroni(2020) 回顾了语言的出现，并连接了两个视角，包括语言在社区中演化的条件以及通过动态通信解决问题的能力。基于理论分析，Zhang 等人 (2019) 专注于多智能体强化学习算法，并从数学角度提出了挑战。

1.2 Contribution and survey structure

1.2 贡献与调查结构

The contribution of this paper is to present a comprehensive survey of the recent research directions pursued in the field of MADRL. We depict a holistic overview of current challenges that arise exclusively in the multi-agent domain of deep RL and discuss state-of-the-art solutions that were proposed to address these challenges. In contrast to the surveys of Hernandez-Leal et al. (2019) and Nguyen et al. (2020), which focus on a subset of topics, we aim to provide a widened and more comprehensive overview of the current investigations conducted in the field of MADRL while recapitulating what has already been accomplished. We identify contemporary challenges and discuss literature that addresses such. We see our work complementary to the theoretical survey of Zhang et al. (2019).

本文的贡献是全面回顾了多智能体深度强化学习领域近期追求的研究方向。我们描绘了当前仅在深度强化学习的多智能体领域中出现的挑战的全貌，并讨论了为解决这些挑战而提出的最先进解决方案。与 Hernandez-Leal 等人 (2019) 和 Nguyen 等人 (2020) 的调查相比，他们专注于主题的一个子集，我们旨在提供一个更广泛、更全面的研究概览，同时概括了在多智能体深度强化学习领域已经完成的工作。我们确定了当代挑战，并讨论了解决这些挑战的文献。我们认为我们的工作是对 Zhang 等人 (2019) 的理论调查的补充。

We dedicate this paper to an audience who wants an excursion to the realm of MADRL. Readers shall gain insights about the historical roots of this still young field and its current developments, but also understand the open problems to be faced by future research. The contents of this paper are organized as follows. We begin with a formal introduction to both single-agent and multi-agent RL and reveal pathologies that are present in MARL in Sect. 2. We then continue with the main contents, which are categorized according to the three-fold taxonomy as illustrated in Fig. 1.

我们将本文献给那些希望涉猎多智能体深度强化学习 (MADRL) 领域的读者。读者将深入了解这一年轻领域的历史根源及其当前发展，同时理解未来研究需要面对的开放性问题。本文内容组织如下。我们从对单一智能体和多元智能体强化学习的正式介绍开始，并在第 2 节揭示多智能体强化学习 (MARL) 中存在的病理现象。然后，我们继续讨论主要内容，这些内容根据图 1 所示的三元分类法进行分类。

We analyze training architectures in Sect. 3, where we categorize approaches according to a centralized or distributed training paradigm and additionally differentiate into execution schemes. Thereafter, we review literature that investigates emergent patterns of agent behavior in Sect. 4. We classify works in terms of the reward structure (Sect. 4.1), the language between multiple agents (Sect. 4.2), and the social context (Sect. 4.3). In Sect. 5, we enumerate current challenges of the multi-agent domain, which include the non-stationarity of the environment due to simultaneously adapting learners (Sect. 5.1), the learning of meaningful communication protocols in cooperative tasks (Sect. 5.2), the need for coherent coordination of agent actions (Sect. 5.3), the credit assignment problem (Sect. 5.4), the ability to scale to an arbitrary number of decision-makers (Sect. 5.5), and non-Markovian environments due to partial observations (Sect. 5.6). We discuss the matter of MADRL, pose trends that we identified in recent literature, and outline possible future work in Sect. 6. Finally, this survey concludes in Sect. 7.

我们在第 3 节分析训练架构，其中我们根据集中式或分布式训练范例对方法进行分类，并额外区分执行方案。之后，我们在第 4 节回顾研究文献，这些文献调查了智能体行为出现的模式。我们根据奖励结构 (第 4.1 节)、多个智能体之间的语言 (第 4.2 节) 和社会背景 (第 4.3 节) 对作品进行分类。在第 5 节，我们列举了多智能体领域的当前挑战，包括由于同时适应的学习者导致的环境的非平稳性 (第 5.1 节)、在合作任务中学习有意义的通信协议的需求 (第 5.2 节)、智能体行动的一致性协调需求 (第 5.3 节)、信用分配问题 (第 5.4 节)、能够扩展到任意数量决策者的能力 (第 5.5 节) 以及由于部分观察导致的非马尔可夫环境 (第 5.6 节)。我们在第 6 节讨论 MADRL 问题，提出我们在近期文献中识别的趋势，并概述未来可能的工作。最后，本调查在第 7 节结束。

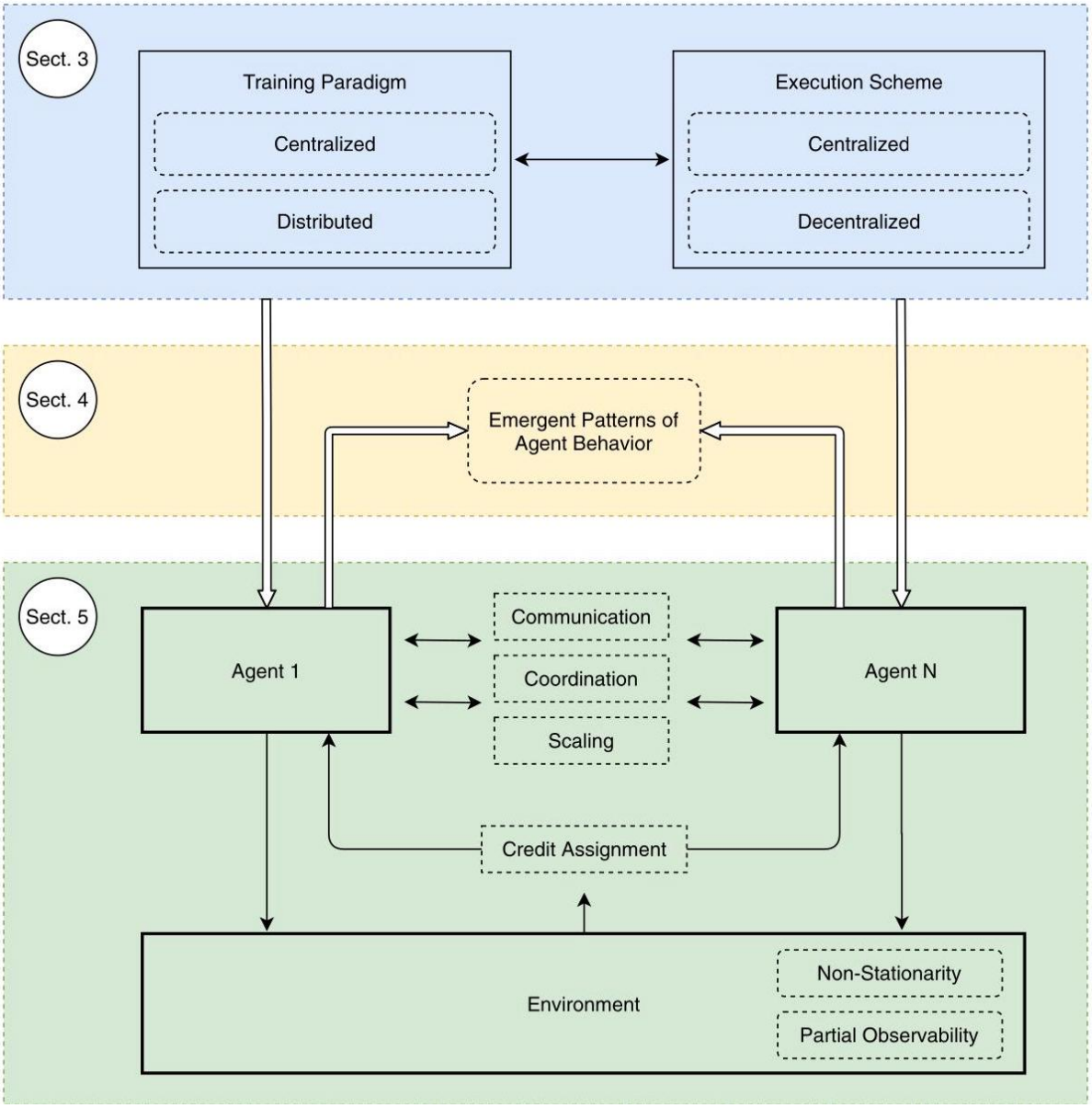

Fig. 1 Schematic structure of the main contents in this survey. In Sect. 3, we review schemes that are applied to train agent behavior in the multi-agent setting. The training of agents can be divided into two paradigms which are namely distributed (Sect. 3.1) and centralized (Sect. 3.2). In Sect. 4, we consider the emergent patterns of agent behavior with respect to the reward structure (Sect. 4.1), the language (Sect. 4.2) and the social context (Sect. 4.3). In Sect. 5, we enumerate current challenges of MADRL which include the non-stationarity of the environment due to co-adapting agents (Sect. 5.1), the learning of communication (Sect. 5.2), the need for a coherent coordination of actions (Sect. 5.3), the credit assignment problem (Sect. 5.4), the ability to scale to an arbitrary number of decision-makers (Sect. 5.5), and non-Markovian environments due to partial observations (Sect. 5.6)

图 1 本调查主要内容示意图。在第 3 节中，我们回顾了应用于多代理环境训练代理行为的方案。代理的训练可以分为两种范例，即分布式 (第 3.1 节) 和集中式 (第 3.2 节)。在第 4 节中，我们考虑了与奖励结构 (第 4.1 节)、语言 (第 4.2 节) 和社会背景 (第 4.3 节) 相关的代理行为的涌现模式。在第 5 节中，我们列举了 MADRL 当前的挑战，包括由于共同适应的代理导致的环境的非平稳性 (第 5.1 节)、通信学习 (第 5.2 节)、行动一致协调的需求 (第 5.3 节)、信用分配问题 (第 5.4 节)、能够扩展到任意数量决策者的能力 (第 5.5 节)，以及由于部分观察导致的非马尔可夫环境 (第 5.6 节)。

2 Background

2 背景

In this section, we provide a formal introduction into the concepts of RL. We start with the Markov decision process as a framework for single-agent learning in Sect. 2.1. We continue with the multi-agent case and introduce the Markov Game in Sect. 2.2. Finally, we pose pathologies that arise in the multi-agent domain such as the non-stationarity of the environment from the perspective of a single learner, relative over-generalization, and the credit assignment problem in Sect. 2.3. We provide the formal concepts behind these MARL pathologies in order to drive our discussion about the state-of-the-art approaches in Sect. 5. The scope of this background section is deliberately focusing on classical MARL works to reveal the roots of the domain and to give the reader insights into the early works on which modern MADRL approaches rest.

在本节中，我们提供了对强化学习 (RL) 概念的正式介绍。我们从第 2.1 节中的马尔可夫决策过程开始，作为单智能体学习的框架。接着在第 2.2 节中，我们讨论了多智能体的情况，并引入了马尔可夫博弈。最后，在第 2.3 节中，我们从单个学习者的角度提出了多智能体领域中出现的病理现象，如环境的非平稳性、相对过度泛化和信用分配问题。我们提供了这些多智能体强化学习 (MARL) 病理背后的正式概念，以便推动我们对第 5 节中最新方法的讨论。本背景章节的范围故意集中在经典 MARL 工作上，以揭示该领域的根源，并给读者洞察现代多智能体深度强化学习 (MADRL) 方法所依赖的早期工作。

2.1 Single-agent reinforcement learning

2.1 单智能体强化学习

The traditional reinforcement learning problem (Sutton and Barto 1998) is concerned with learning a control policy that optimizes a numerical performance by making decisions in stages. The decision-maker called agent interacts with an environment of unknown dynamics in a trial-and-error fashion and occasionally receives feedback upon which the agent wants to improve. The standard formulation for such sequential decision-making is the Markov decision process, which is defined as follows (Bellman 1957; Bertsekas 2012, 2017; Kaelbling et al. 1996).

传统强化学习问题 (Sutton 和 Barto 1998) 关注于学习一个控制策略，通过分阶段做出决策来优化一个数值性能。被称为决策者的智能体与具有未知动态的环境进行试错交互，并偶尔接收到反馈，智能体希望据此改进。这种顺序决策的标准公式是马尔可夫决策过程，其定义如下 (Bellman 1957; Bertsekas 2012, 2017; Kaelbling 等人 1996)。

Definition 1 Markov decision process (MDP) A Markov decision process is formalized by the tuple $(\mathcal{X},\mathcal{U},\mathcal{P},R,\gamma)$ where $\mathcal{X}$ and $\mathcal{U}$ are the state and action space, respectively, $\mathcal{P}:\mathcal{X}\times\mathcal{U}\to P(\mathcal{X})$ is the transition function describing the probability of a state transition, $R:\mathcal{X}\times\mathcal{U}\times\mathcal{X}\to\mathbb{R}$ is the reward function providing an immediate feedback to the agent, and $\gamma\in[0,1)$ describes the discount factor.

定义 1 马尔可夫决策过程 (MDP) 马尔可夫决策过程形式化为元组 $(\mathcal{X},\mathcal{U},\mathcal{P},R,\gamma)$ ，其中 $\mathcal{X}$ 和 $\mathcal{U}$ 分别是状态空间和动作空间，$\mathcal{P}:\mathcal{X}\times\mathcal{U}\to P(\mathcal{X})$ 是描述状态转换概率的转移函数，$R:\mathcal{X}\times\mathcal{U}\times\mathcal{X}\to\mathbb{R}$ 是提供对智能体即时反馈的奖励函数，$\gamma\in[0,1)$ 描述了折扣因子。

The agent's goal is to act in such a way as to maximize the expected performance on a long-term perspective with regard to an unknown transition function $\mathcal{P}$. Therefore, the agent learns a behavior policy $\pi:\mathcal{X}\to P(\mathcal{U})$ that optimizes the expected performance J throughout learning. The performance is defined as the expected value of discounted rewards

代理的目标是采取行动以最大化长期预期表现，这涉及到一个未知的转移函数 $\mathcal{P}$ 。因此，代理学习了一种行为策略 $\pi:\mathcal{X}\to P(\mathcal{U})$ ，以优化整个学习过程中的预期表现 J 。表现被定义为折现奖励的预期值

$$J=\mathbb{E}_{x_0\sim\rho_0,x_{t+1}\sim\mathcal{P},u_t\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R\left(x_t,u_t,x_{t+1}\right)\right] \tag{1}$$

over the initial state distribution ρ_0 while selected actions are governed by the policy π . Here, we regard the infinite-horizon problem where the interaction between agent and environment does not terminate after a countable number of steps. Note that the learning objective can also be formalized for finite-horizon problems (Bertsekas 2012, 2017). As an alternative to the policy performance, which describes the expected performance as a function of the policy, one can define the utility of being in a particular state in terms of a value function. The state-value function $V_\pi:\mathcal{X}\to\mathbb{R}$ describes the utility under policy π when starting from state x , i.e.

在初始状态分布 ρ_0 上，而选择的行为则由策略 π 控制。在这里，我们考虑无限时间视野问题，即代理与环境的交互不会在可数步之后终止。请注意，学习目标也可以为有限时间视野问题形式化 (Bertsekas 2012, 2017)。作为策略表现的替代，即表现预期作为策略的函数，人们可以定义在特定状态下的效用，即价值函数。状态价值函数 $V_\pi : \mathcal{X} \to \mathbb{R}$ 描述了在策略 π 下从状态 x 开始的效用，即

$$V_\pi(x) = \mathbb{E}_{x_{t+1}\sim\mathcal{P}, u_t\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(x_t, u_t, x_{t+1}) \mid x_0 = x\right]. \tag{2}$$

In a similar manner, the action-value function $Q_\pi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ describes the utility of being in state x , performing action u , and following the policy π thereafter, that is

同样地，动作价值函数 $Q_\pi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ 描述了在状态 x 下执行动作 u ，然后遵循策略 π 的效用，即

$$Q_\pi(x, u) = \mathbb{E}_{x_{t+1}\sim\mathcal{P}, u_{t>0}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(x_t, u_t, x_{t+1}) \mid x_0 = x, u_0 = u\right]. \tag{3}$$

In the context of deep reinforcement learning, either the policy, a value function or both are represented by neural networks.

在深度强化学习的背景下，策略、价值函数或两者都由神经网络表示。

2.2 Multi-agent reinforcement learning

2.2 多代理强化学习

When the sequential decision-making is extended to multiple agents, Markov Games [1] are commonly applied as framework. The Markov Game was originally introduced by Littman (1994) to generalize MDPs to multiple agents that simultaneously interact within a shared environment and possibly with each other. The definition is formalized in a discrete-time setting and is denoted as follows (Littman 1994).

当顺序决策扩展到多个代理时，Markov Games [1] 通常作为框架应用。Markov Game 最初由 Littman(1994) 提出，用于将 MDPs 推广到多个代理，这些代理在共享环境中可能同时相互互动。该定义在离散时间设置中形式化，并表示如下 (Littman 1994)。

Definition 2 Markov Games (MG) The Markov Game is an extension to the MDP and is formalized by the tuple $\left(\mathcal{N}, \mathcal{X}, \left\{\mathcal{U}^i\right\}, \mathcal{P}, \left\{R^i\right\}, \gamma\right)$, where $\mathcal{N} = \{1, \ldots, N\}$ denotes the set of $N > 1$ interacting agents and $\mathcal{X}$ is the set of states observed by all agents. The joint action space is denoted by $\mathcal{U} = \mathcal{U}^1 \times \cdots \times \mathcal{U}^N$ which is the collection of individual action spaces from agents $i \in \mathcal{N}$. The transition probability function $\mathcal{P} : \mathcal{X} \times \mathcal{U} \to P(\mathcal{X})$ describes the chance of a state transition. Each agent owns an associated reward function $R^i : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to \mathbb{R}$ that provides an immediate feedback signal. Finally, $\gamma \in [0, 1)$ describes the discount factor.

定义 2 马尔可夫博弈 (MG) 马尔可夫博弈是 MDP 的扩展，形式化为元组 $\left(\mathcal{N}, \mathcal{X}, \left\{\mathcal{U}^i\right\}, \mathcal{P}, \left\{R^i\right\}, \gamma\right)$ ，其中 $\mathcal{N} = \{1, \ldots, N\}$ 表示 $N > 1$ 交互代理的集合，$\mathcal{X}$ 是所有代理观察到的状态集合。联合行动空间表示为 $\mathcal{U} = \mathcal{U}^1 \times \cdots \times \mathcal{U}^N$ ，它是来自代理 $i \in \mathcal{N}$ 的个体行动空间的集合。转移概率函数 $\mathcal{P} : \mathcal{X} \times \mathcal{U} \to P(\mathcal{X})$ 描述了状态转移的机会。每个代理拥有一个关联的奖励函数 $R^i : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to \mathbb{R}$ ，提供即时反馈信号。最后，$\gamma \in [0, 1)$ 描述了折扣因子。

At stage t , each agent $i \in \mathcal{N}$ selects and executes an action depending on the individual policy $\pi^i : \mathcal{X} \to P\left(\mathcal{U}^i\right)$. The system evolves from state x_t under the joint action u_t with respect to the transition probability function $\mathcal{P}$ to the next state x_{t+1} while each agent receives R^i as immediate feedback to the state transition. Akin to the single-agent problem, the aim of each agent is to change its policy in such a way as to optimize the received rewards on a long-term perspective.

在阶段 t 中，每个代理 $i \in \mathcal{N}$ 根据个体策略 $\pi^i : \mathcal{X} \to P\left(\mathcal{U}^i\right)$ 选择并执行一个动作。系统在联合行动 u_t 下，根据转移概率函数 $\mathcal{P}$ 从状态 x_t 演变到下一个状态 x_{t+1} ，同时每个代理接收到 R^i 作为状态转换的即时反馈。类似于单代理问题，每个代理的目标是以一种优化长期接收奖励的方式改变其策略。

A special case of the MG is the stateless setting $\mathcal{X} = \varnothing$ called strategic-form game [2] . Strategic-form games describe one-shot interactions where all agents simultaneously execute an action and receive a reward based on the joint action after which the game ends. Significant progress within the MARL community has been accomplished by studying this simplified stateless setting, which is still under active research to cope with several pathologies as discussed later in this section. These games are also known as matrix games because the reward function is represented by an $N \times N$ matrix. The formalism which extends to multi-step sequential stages is called extensive-form game.

MG 的一个特例是无状态设置 $\mathcal{X}=\varnothing$ ，称为策略形式博弈 [2] 。策略形式博弈描述了一次性交互，其中所有代理同时执行一个动作，并在游戏结束后基于联合行动接收奖励。MARL 社区通过研究这个简化的无状态设置取得了重要进展，该设置目前仍在积极研究中，以应对本节稍后讨论的几种病理现象。这些游戏也被称为矩阵游戏，因为奖励函数由一个 $N \times N$ 矩阵表示。扩展到多步骤顺序阶段的范式称为扩展形式博弈。

In contrast to the single-agent case, the value function $V^i : \mathcal{X} \to \mathbb{R}$ does not only depend on the individual policy of agent i but also on the policies of other agents, i.e. the value function for agent i is the expected sum

与单智能体情况不同，值函数 $V^i : \mathcal{X} \to \mathbb{R}$ 不仅取决于智能体 i 的个体策略，还取决于其他智能体的策略，即智能体 i 的值函数是期望总和

$$V^i_{\pi^i,\pi^{-i}}(x) = \mathbb{E}_{x_{t+1}\sim\mathcal{P},u_t\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R^i(x_t,u_t,x_{t+1}) \mid x_0 = x\right] \tag{4}$$

when the agents behave according to the joint policy π . We denote the joint policy $\pi : \mathcal{X} \to P(\mathcal{U})$ as the collection of all individual policies, i.e. $\pi = \left\{\pi^1,\ldots,\pi^N\right\}$. Further, we make use of the convention that $-i$ denotes all agents except i , meaning for policies that

当智能体根据联合策略 π 行为时。我们用联合策略 $\pi : \mathcal{X} \to P(\mathcal{U})$ 表示所有个体策略的集合，即 $\pi = \left\{\pi^1,\ldots,\pi^N\right\}$ 。此外，我们使用约定 $-i$ 表示除 i 之外的所有智能体，对于策略来说意味着 $\pi^{-i} = \left\{\pi^1,\ldots,\pi^{i-1},\pi^{i+1},\ldots,\pi^N\right\}$.

The optimal policy is determined by the individual policy and the other agents' strategies. However, when other agents' policies are fixed, the agent i can maximize its own utility by finding the best response π^i_* with respect to the other agents' strategies.

最优策略由个体策略和其他智能体的策略决定。然而，当其他智能体的策略固定时，智能体 i 可以通过找到相对于其他智能体策略的最佳响应 π^i_* 来最大化自己的效用。

Definition 3 Best response The agent' s i best response $\pi^i_* \in \Pi^i$ to the joint policy π^{-i} of other agents is

定义 3 最佳响应智能体 i 对其他智能体的联合策略 π^{-i} 的最佳响应是

$$V^i_{\pi^i_*,\pi^{-i}}(x) \geq V^i_{\pi^i,\pi^{-i}}(x)$$

for all states $x \in \mathcal{X}$ and policies $\pi^i \in \Pi^i$.

对于所有状态 $x \in \mathcal{X}$ 和策略 $\pi^i \in \Pi^i$ 。

In general, when all agents learn simultaneously, the found best response may not be unique (Shoham and Leyton-Brown 2008). The concept of best response can be leveraged to describe the most influential solution concept from game theory: the Nash equilibrium.

通常，当所有智能体同时学习时，找到的最佳响应可能不是唯一的 (Shoham 和 Leyton-Brown 2008)。最佳响应的概念可以用来描述来自博弈论的最有影响力的解决方案概念: 纳什均衡。

Definition 4 Nash equilibrium A solution where each agent' s policy π^*_i is the best response to the other agents' policy π^{-i}_* such that the following inequality

translation: 定义 4 纳什均衡每个智能体的策略 π^*_i 是对其他智能体策略 π^{-i}_* 的最佳响应，使得以下不等式

$$V^i_{\pi^i_*,\pi^{-i}_*}(x) \geq V^i_{\pi^i,\pi^{-i}_*}(x)$$

holds true for all states $x \in \mathcal{X}$ and all policies $\pi^i \in \Pi^i \forall i$ is called Nash equilibrium.

对于所有状态 $x \in \mathcal{X}$ 和所有策略 $\pi^i \in \Pi^i \forall i$ 都成立的均衡被称为纳什均衡。

Intuitively spoken, a Nash equilibrium is a solution where one agent cannot improve when the policies of other agents are fixed, that is no agent can improve by unilaterally deviating from π^* . However, a Nash equilibrium may not be unique. Thus, the concept of Pareto-optimality might be useful (Matignon et al. 2012b).

---

1 Markov games are also known as Stochastic Games (Shapley 1953), but we continue to use the term Markov Game to draw a clear distinction between deterministic Markov Games and stochastic Markov Games.

1 Markov 游戏也被称为随机游戏 (Shapley 1953)，但我们继续使用 Markov 游戏这个术语，以明确区分确定性 Markov 游戏和随机 Markov 游戏。

2 The strategic-form game is also known as matrix game or normal-form game. The most commonly studied strategic-form game is the one with $N = 2$ players, the so-called bi-matrix game.

2 战略形式游戏也被称为矩阵游戏或标准形式游戏。最常研究的战略形式游戏是具有 $N = 2$ 个参与者的游戏，所谓的双矩阵游戏。

直观地说，纳什均衡是一个解决方案，其中一个代理在其它代理的策略固定时无法改进，也就是说没有代理可以通过单方面偏离 π^* 来改进。然而，纳什均衡可能不是唯一的。因此，帕累托最优的概念可能是有用的 (Matignon et al. 2012b)。

Definition 5 Pareto-optimality A joint policy π Pareto-dominates a second joint policy $\widehat{\pi}$ if and only if

定义 5 帕累托最优性如果且仅如果第一个联合策略 π 帕累托支配第二个联合策略 $\widehat{\pi}$ 。

$$V_{\pi}^{i}(x) \geq V_{\widehat{\pi}}^{i}(x) \ \forall i, \forall x \in \mathcal{X} \text{ and } V_{\pi}^{j}(x) > V_{\widehat{\pi}}^{j}(x) \ \exists j, \exists x \in \mathcal{X}.$$

A Nash equilibrium is regarded to be Pareto-optimal if no other has greater value and, thus, is not Pareto-dominated.

如果没有其他策略具有更大的价值，那么纳什均衡被认为是帕累托最优的，因此不会被帕累托支配。

Classical MARL literature can be categorized according to different features, such as the type of task and the information available to agents. In the remainder of this section, we introduce MARL concepts based on the taxonomy proposed in Busoniu et al. (2008). For one, the primary factor that influences the learned agent behavior is the type of task. Whether agents compete or cooperate is promoted by the designed reward structure.

经典的多智能体强化学习文献可以根据不同的特征进行分类，例如任务类型和代理可用的信息。在本节的其余部分，我们基于 Busoniu et al. (2008) 提出的分类法介绍多智能体强化学习的概念。首先，影响学习代理行为的主要因素是任务类型。代理是竞争还是合作是由设计的奖励结构所促进的。

(1) Fully cooperative setting All agents receive the same reward $R = R^i = \cdots = R^N$ for state transitions. In such an equally-shared reward setting, agents are motivated to collaborate and try to avoid the failure of an individual to maximize the performance of the team. More generally, we talk about cooperative settings when agents are encouraged to collaborate but do not own an equally-shared reward.

(1) 完全合作设置所有代理在状态转换时获得相同的奖励 $R = R^i = \cdots = R^N$ 。在这样的平等共享奖励设置中，代理有动力合作并尝试避免个体的失败以最大化团队的表现。更一般地说，当代理被鼓励合作但不拥有平等共享的奖励时，我们谈论的是合作设置。

(2) Fully competitive setting Such problem is described as a zero-sum Markov Game where the sum of rewards equals zero for any state transition, i.e. $R = \sum_{i=1}^{N} R^i(x, u, x') = 0$. Agents are prudent to maximize their own individual reward while minimizing the reward of the others. In a loose sense, we refer to competitive games when agents are encouraged to excel against opponents, but the sum of rewards does not equal zero.

(2) 完全竞争设置这种问题被描述为零和马尔可夫游戏，其中任何状态转换的奖励总和为零，即 $R = \sum_{i=1}^{N} R^i(x, u, x') = 0$ 。代理谨慎地最大化自己的个人奖励，同时最小化其他代理的奖励。在宽松的意义上，当代理被鼓励在对手中表现出色，但奖励的总和不等于零时，我们指的是竞争游戏。

(3) Mixed setting Also known as general-sum game, the mixed setting is neither fully cooperative nor fully competitive and, thus, does not incorporate restrictions on agent goals.

(3) 混合设置也称为总和博弈，混合设置既不是完全合作也不是完全竞争，因此，不对代理目标设置限制。

Beside the reward structure, other taxonomy may be used to differentiate between the information available to the agents. Claus and Boutilier (1998) distinguished between two types of learning, namely independent learners and joint-action learners. The former ignores the existence of other agents and cannot observe the rewards and selected actions of others as considered in Bowling and Veloso (2002) and Lauer and Riedmiller (2000). Joint-action learners, however, observe the taken actions of all other actions a-posteriori as shown in Hu and Wellman (2003) and Littman (2001).

除了奖励结构，还可以使用其他分类法来区分代理可用的信息。Claus 和 Boutilier(1998) 区分了两种学习类型，即独立学习者和联合行动学习者。前者忽略了其他代理的存在，不能观察到其他代理的奖励和选择的行为，如 Bowling 和 Veloso(2002) 以及 Lauer 和 Riedmiller(2000) 所考虑的。然而，联合行动学习者观察到所有其他代理的后验采取的行动，如 Hu 和 Wellman(2003) 以及 Littman(2001) 所示。

2.3 Formal introduction to multi-agent challenges

2.3 多代理挑战的正式介绍

In the single-agent formalism, the agent is the only decision-instance that influences the state of the environment. State transitions can be clearly attributed to the agent, whereas everything outside the agent's field of impact is regarded as part of the underlying system dynamics. Even though the environment may be stochastic, the learning problem remains stationary.

在单代理形式主义中，代理是唯一影响环境状态的决策实例。状态转换可以明确地归因于代理，而代理影响范围之外的一切被视为底层系统动态的一部分。尽管环境可能是随机的，但学习问题仍然是静止的。

On the contrary, one of the fundamental problems in the multi-agent domain is that agents update their policies during the learning process simultaneously, such that the environment appears non-stationary from the perspective of a single agent. Hence, the Markov assumption of an MDP no longer holds, and agents face-without further treatment-a moving target problem (Busoniu et al. 2008; Yang and Gu 2004).

相反，多代理领域中一个基本问题是代理在学习过程中同时更新其策略，以至于从单个代理的角度看，环境显得非静止。因此，MDP 的马尔可夫假设不再成立，代理面临一个移动目标问题 (Busoniu 等人，2008；Yang 和 Gu，2004)。

Definition 6 Non-stationarity A single agent faces a moving target problem when the transition probability function changes

定义 6 非静止性当转移概率函数发生变化时，单个代理面临一个移动目标问题

$$\mathcal{P}\left(x' \mid x, u, \pi^1, \ldots, \pi^N\right) \neq \mathcal{P}\left(x' \mid x, u, \bar{\pi}^1, \ldots, \bar{\pi}^N\right),$$

due to the co-adaption $\pi^i \neq \bar{\pi}^i \exists i \in \mathcal{N}$ of agents.

由于代理的共同适应 $\pi^i \neq \bar{\pi}^i \exists i \in \mathcal{N}$ 。

Above, we have introduced the Nash equilibrium as a solution concept where each agent's policy is the best response to the others. However, it has been shown that agents can converge, despite a high degree of randomness in action selection, to sub-optimal solutions or can get stuck between different solutions (Wiegand 2004). Fulda and Ventura (2007) investigated such convergence to solutions and described a Pareto-selection problem called shadowed equilibrium.

如上所述，我们引入了纳什均衡作为解决方案概念，即每个智能体的策略都是对其他智能体的最佳响应。然而，已经证明，尽管在动作选择中存在高度随机性，智能体仍然可以收敛到次优解，或者可能会在不同的解决方案之间陷入僵局 (Wiegand 2004)。Fulda 和 Ventura(2007) 研究了这种收敛到解决方案的情况，并描述了一个被称为遮蔽均衡的帕累托选择问题。

Definition 7 Shadowed equilibrium A joint policy $\bar{\pi}$ is shadowed by another joint policy $\hat{\pi}$ in a state x if and only if

定义 7 遮蔽均衡如果在状态 x 中，一个联合策略 $\bar{\pi}$ 被另一个联合策略 $\hat{\pi}$ 遮蔽，当且仅当

$$V_{\pi^i, \bar{\pi}^{-i}}(x) < \min_{j, \pi_j} V_{\pi^j, \hat{\pi}^{-j}}(x) \ \exists i, \pi_i. \tag{5}$$

An equilibrium is shadowed by another when at least one agent exists who, when unilaterally deviating from $\bar{\pi}$, will see no better improvement than for deviating from $\hat{\pi}$ (Mat-ignon et al. 2012b). As a form of shadowed equilibrium, the pathology of relative overgeneralization describes that a sub-optimal Nash equilibrium in the joint action space is preferred over an optimal solution. This phenomenon arises since each agent's policy performs relatively well when paired with arbitrary actions from other agents (Panait et al. 2006; Wei and Luke 2016; Wiegand 2004).

当至少存在一个智能体，当它单方面偏离 $\bar{\pi}$ 时，发现除了偏离 $\hat{\pi}$ 之外没有更好的改进时，一个均衡被另一个遮蔽 (Mat-ignon et al. 2012b)。作为遮蔽均衡的一种形式，相对过度概括的病理现象描述了在联合动作空间中的次优纳什均衡优于最优解决方案。这种现象产生是因为每个智能体的策略在与其他智能体的任意动作配对时表现相对较好 (Panait et al. 2006; Wei 和 Luke 2016; Wiegand 2004)。

In a Markov Game, we assumed that each agent observes a state x , which encodes all necessary information about the world. However for complex systems, complete information might not be perceivable. In such partially observable settings, the agents do not observe the whole state space but merely a subset $\mathcal{O}^i \subset \mathcal{X}$. Hence, the agents are confronted to deal with sequential decision-making under uncertainty. The partially observable Markov Game (Hansen et al. 2004) is the generalization of both MG and MDP.

在马尔可夫博弈中，我们假设每个智能体观察到一个状态 x ，它包含了关于世界的所有必要信息。然而，对于复杂系统，完整信息可能无法感知。在部分可观察的设置中，智能体并不观察整个状态空间，而

仅仅是一个子集 $\mathcal{O}^i \subset \mathcal{X}$ 。因此，智能体必须面对在不确定性下的顺序决策。部分可观察马尔可夫博弈 (Hansen et al. 2004) 是 MG 和 MDP 的推广。

Definition 8 Partially observable Markov Games (POMG) The POMG is mathematically denoted by the tuple $\left(\mathcal{N}, \mathcal{X}, \left\{\mathcal{U}^i\right\}, \left\{\mathcal{O}^i\right\}, \mathcal{P}, \left\{R^i\right\}, \gamma\right)$, where $\mathcal{N} = \{1, \ldots, N\}$ denotes the set of $N > 1$ interacting agents, $\mathcal{X}$ is the set of global but unobserved system states, and $\mathcal{U}$ is the set of individual action spaces $\mathcal{U}_i$. The observation space $\mathcal{O}$ denotes the collection of individual observation spaces $\mathcal{O}^i$. The transition probability function is denoted by $\mathcal{P}$, the reward function associated with agent i by R^i , and the discount factor is γ .

定义 8 部分可观察马尔可夫博弈 (POMG) POMG 在数学上表示为一个元组 $\left(\mathcal{N}, \mathcal{X}, \left\{\mathcal{U}^i\right\}, \left\{\mathcal{O}^i\right\}, \mathcal{P}, \left\{R^i\right\}, \gamma\right)$ ，其中 $\mathcal{N} = \{1, \ldots, N\}$ 表示 $N > 1$ 交互代理的集合，$\mathcal{X}$ 是全局但不可观察的系统状态的集合，$\mathcal{U}$ 是个体动作空间的集合 $\mathcal{U}_i$ 。观察空间 $\mathcal{O}$ 表示个体观察空间集合 $\mathcal{O}^i$ 。转移概率函数表示为 $\mathcal{P}$ ，与代理 i 相关的奖励函数表示为 R^i ，折扣因子为 γ 。

When agents face a cooperative task with a shared reward function, the POMG is then known as decentralized Partially Observable Markov decision process (dec-POMDP) (Bernstein et al. 2002; Oliehoek and Amato 2016). In partially observable domains, the inference of good policies is extended in complexity since the history of interactions becomes meaningful. Hence, the agents usually incorporate history-dependent policies $\pi_t^i : \left\{\mathcal{O}^i\right\}_{t>0} \rightarrow P\left(\mathcal{U}^i\right)$, which map from a history of observations to a distribution over actions.

当代理面临具有共享奖励函数的合作任务时，POMG 则被称为分布式部分可观察马尔可夫决策过程 (dec-POMDP)(Bernstein 等人 2002 年；Oliehoek 和 Amato 2016 年)。在部分可观察领域中，良好策略的推理在复杂性上得到扩展，因为交互历史变得有意义。因此，代理通常会采用依赖于历史的策略 $\pi_t^i : \left\{\mathcal{O}^i\right\}_{t>0} \rightarrow P\left(\mathcal{U}^i\right)$ ，这些策略将观察历史映射到动作的分布上。

Definition 9 Credit assignment problem In the fully-cooperative setting with joint reward signals, an individual agent cannot conclude the impact of its own action towards the team's success and, thus, faces a credit assignment problem.

定义 9 信用分配问题在完全合作设置中，具有联合奖励信号时，单个代理无法确定其自身动作对团队成功的影响，因此面临信用分配问题。

In cooperative games, agents are encouraged to maximize a common goal through a joint reward signal. However, agents cannot ascertain their contribution to the eventual reward when they do not experience the taken joint action or deal with partial observations. Associating rewards to agents is known as the credit assignment problem (Chang et al. 2004; Weiß 1995; Wolpert and Tumer 1999).

在合作游戏中，代理被鼓励通过联合奖励信号最大化共同目标。然而，当代理没有经历采取的联合行动或处理部分观察时，他们无法确定自己对最终奖励的贡献。将奖励关联到代理被称为信用分配问题 (Chang 等人 2004 年；Weiß 1995 年；Wolpert 和 Tumer 1999 年)。

Some of the above-introduced pathologies occur in all cooperative, competitive, and mixed tasks, whereas some pathologies like relative over-generalization, credit assignment, and miss-coordination are predominant issues in cooperative settings. To cope with these pathologies, still commonly studied settings are tabular worlds such as variations of the climbing game where solutions are not yet found, e.g. when the environment exhibits reward stochasticity (Claus and Boutilier 1998). Thus, simple worlds remain a fertile ground for further research, especially for problems like shadowed equilibria, non-stationarity or alter-exploration problems [3] and continue to matter for modern deep learning approaches.

上述介绍的一些病理现象在所有合作、竞争和混合任务中都会出现，而相对过度泛化、信用分配和协调失误等病理现象在合作环境中尤为突出。为了应对这些病理现象，目前仍然被广泛研究的设置是表格世界，例如攀爬游戏的变体，其中解决方案尚未找到，例如当环境表现出奖励随机性时 (Claus 和 Boutilier 1998)。因此，简单世界仍然是进一步研究的肥沃土壤，特别是对于阴影均衡、非定常性或交替探索问题 [3] ，并且对于现代深度学习方法仍然具有重要意义。

3 Analysis of training schemes

3 训练方案分析

The training of multiple agents has long been a computational challenge (Becker et al. 2004; Nair et al. 2003). Since the complexity in the state and action space grows exponentially with the number of agents, even modern deep learning approaches may reach their limits. In this section, we describe training schemes that are used in practice for learning agent policies in the multi-agent setting similar to the ones described in Bono et al. (2019). We denote training as the process during which agents acquire data to build up experience and optimize their behavior with respect to the received reward signals. In

contrast, we refer test time [4] to the step after the training when the learned policy is evaluated but is no further refined. The training of agents can be broadly divided into two paradigms, namely centralized and distributed (Weiß 1995). If the training of agents is applied in a centralized manner, policies are updated based on the mutual exchange of information during the training. This additional information is then usually removed at test time. In contrast to the centralized scheme, the training can also be handled in a distributed fashion where each agent performs updates on its own and develops an individual policy without utilizing foreign information.

多个代理的训练长期以来一直是计算上的挑战 (Becker 等人 2004; Nair 等人 2003)。由于状态和动作空间的复杂性随着代理数量的增加而指数级增长，即使是现代深度学习方法也可能达到其极限。在本节中，我们描述了在实践中用于多代理设置下学习代理策略的训练方案，类似于 Bono 等人 (2019) 中描述的方案。我们将训练定义为代理在接收到的奖励信号方面获取数据以建立经验并优化其行为的整个过程。相比之下，我们将测试时间 [4] 指的是训练之后的步骤，此时学习的策略被评估但不再进行细化。代理的训练可以大致分为两种范式，即集中式和分布式 (Weiß 1995)。如果代理的训练是以集中方式进行，那么策略的更新是基于训练期间的信息互换来进行的。然后通常在测试时间去除这些额外信息。与集中方案相比，训练也可以以分布式方式进行，其中每个代理都在自己的策略上执行更新，并发展出不需要利用外部信息的个体策略。

In addition to the training paradigm, agents may deviate in the way of how they select actions. We recognize two execution schemes. Centralized execution describes that agents are guided from a centralized unit, which computes the joint actions for all agents. On the contrary, agents determine actions according to their individual policy for decentralized execution. An overview of the training schemes is depicted in Fig. 2 while Table 1 lists the reviewed literature of this section.

除了训练范式之外，代理在选择动作的方式上可能会偏离。我们识别了两种执行方案。集中执行描述了代理从一个集中单元获得指导，该单元为所有代理计算联合动作。相反，代理根据各自的策略确定动作，以实现分散执行。训练方案的概述如图 2 所示，而表 1 列出了本节回顾的文献。

3.1 Distributed training

3.1 分布式训练

In distributed training schemes, agents learn independently of other agents and do not rely on explicit information exchange.

在分布式训练方案中，代理独立于其他代理进行学习，并不依赖于明确的信息交换。

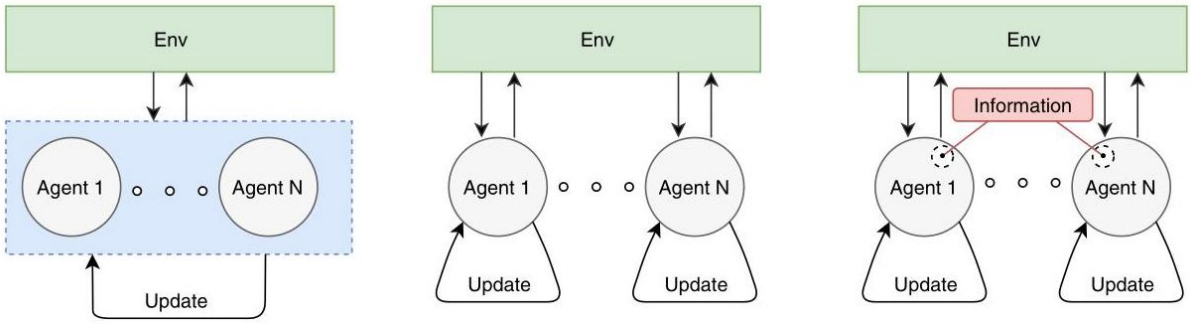

Fig. 2 Training schemes in the multi-agent setting. (Left) CTCE holds a joint policy for all agents. (Middle) Each agent updates its own individual policy in DTDE. (Right) CTDE enables agents to exchange additional information during training which is then discarded at test time

3 The alter-exploration dilemma, also known as the exploration-exploitation problem, describes the tradeoff an agent faces to decide whether to choose actions that extend experience or take decisions that are already optimal according to the current knowledge.

3 替代探索困境，也称为探索-利用问题，描述了代理面临的选择，是选择扩展经验的行为，还是根据当前知识做出已经最优的决策。

[4] Note that test and execution time are often used interchangeably in recent literature. For clarity, we use the term test for the post-training evaluation and the term execution for the action selection with respect to some policy.

[4] 注意，在最近的文献中，测试和执行时间经常可以互换使用。为了清晰起见，我们使用术语“测试”来表示训练后的评估，使用术语“执行”来表示关于某个策略的动作选择。

图 2 多代理环境下的训练方案。(左)CTCE 为所有代理持有一个联合策略。(中) 每个代理在 DTDE 中更新自己的个体策略。(右)CTDE 使得代理在训练期间能够交换额外信息，然后在测试时间丢弃这些信息。

Definition 10 Distributed training decentralized execution (DTDE) Each agent i has an associated policy $\pi^i : \mathcal{O}^i \rightarrow P\left(\mathcal{U}^i\right)$ which maps local observations to a distribution over individual actions. No information is shared between agents such that each agent learns independently.

定义 10 分布式训练分散执行 (DTDE) 每个代理 i 都有一个关联的策略 $\pi^i : \mathcal{O}^i \rightarrow P\left(\mathcal{U}^i\right)$ ，该策略将局部观察映射到个体动作的分布上。代理之间不共享任何信息，使得每个代理独立学习。

The fundamental drawback of the DTDE paradigm is that the environment appears non-stationary from a single agent's viewpoint because agents neither have access to the knowledge of others, nor do they perceive the joint action. The first approaches in this training scheme were studied in tabular worlds. The work by Tan (1993) investigated the question if independently learning agents can match with cooperating agents. The results showed that independent learners learn slower in tabular and deterministic worlds. Based on that, Claus and Boutilier (1998) examined both independent and joint-action learners in cooperative stochastic-form games and empirically showed that both types of learning can converge to an equilibrium in deterministic games. Subsequent works elaborated on the DTDE scheme in discretized worlds (Hu and Wellman 1998; Lauer and Riedmiller 2000).

DTDE 范式的根本缺点在于，从单个智能体的视角来看，环境显得非平稳，因为智能体既无法获取其他智能体的知识，也无法感知联合行动。在这种训练方案中，最早的方法是在表格世界中进行研究的。Tan(1993) 的工作调查了独立学习的智能体是否能够与合作的智能体相匹配的问题。结果显示，在表格和确定性世界中，独立学习者学习速度较慢。基于此，Claus 和 Boutilier(1998) 研究了合作随机形式游戏中的独立学习者和联合行动学习者，并通过实验证明了在确定性游戏中，这两种学习类型都可以收敛到平衡状态。后续的工作在离散化世界中详细阐述了 DTDE 方案 (Hu 和 Wellman 1998；Lauer 和 Riedmiller 2000)。

More recent works report that distributed training schemes scale poorly with the number of agents due to the extra sample complexity, which is added to the learning problem. Gupta et al. (2017) showed that distributed methods have inferior performance compared to policies that are trained with a centralized training paradigm. Similarly, Foerster et al. (2018b) showed that the speed of independently learning actor-critic methods is slower than using centralized training. In further works, DTDE has been applied to cooperative navigation tasks (Chen et al. 2016; Strouse et al. 2018), to partially observable domains (Dobbe et al. 2017; Nguyen et al. 2017b; Srinivasan et al. 2018), and to social dilemmas (Leibo et al. 2017).

更新的研究报道指出，由于额外的样本复杂性被添加到学习问题中，分布式训练方案在智能体数量增加时扩展性较差。Gupta 等人 (2017) 表明，与集中训练范式训练的策略相比，分布式方法的性能较差。同样，Foerster 等人 (2018b) 表明，独立学习演员-评论家方法的速度慢于使用集中训练的速度。在后续的工作中，DTDE 已被应用于合作导航任务 (Chen 等人 2016；Strouse 等人 2018)，部分可观测领域 (Dobbe 等人 2017；Nguyen 等人 2017b；Srinivasan 等人 2018) 以及社会困境 (Leibo 等人 2017)。

Due to limited information in the distributed setting, independent learners are confronted with several pathologies (Matignon et al. 2012b). Besides non-stationarity, environments may exhibit stochastic transitions or stochastic rewards, which further complicates learning. In addition to that, the search for an optimal policy influences the other agents' decision-making, which may lead to action shadowing and impacts the balance between exploration and knowledge exploitation.

由于分布式环境中信息有限，独立学习者面临着多种病理现象 (Matignon 等人，2012b)。除了非平稳性，环境可能表现出随机转换或随机奖励，这进一步使学习复杂化。此外，寻找最优策略会影响其他代理的决策，可能导致行为模仿，并影响探索与知识利用之间的平衡。

A line of recent works expands independent learners with techniques to cope with the aforementioned MARL pathologies in cooperative domains. First, Omidshafiei et al. (2017) introduced a decentralized experience replay extension called Concurrent Experience Replay Trajectories (CERT) that enables independent learners to face a cooperative

近期一系列工作扩展了独立学习者的技术，以应对上述多智能体强化学习 (MARL) 中的病理现象，特别是在合作领域中。首先，Omidshafiei 等人 (2017) 引入了一种去中心化的经验回放扩展，称为并发经验回放轨迹 (CERT)，使得独立学习者能够在合作和部分可观测的环境中面对更加稳定和高效样本。

Table 1 Overview of training schemes applied in recent MADRL works

表 1 近期多智能体深度强化学习 (MADRL) 工作中应用的训练方案概览

Scheme	Approach	Literature
CTCE		Gupta et al. (2017) and Sunehag et al. (2018)
CTDE	Parameter sharing	Altalhi and Dayan (2019), Chu and Ye (2017), Gupta et al. (2017), Peng et al. (2017), Sukhbaatar et al. (2016) and Sunehag et al. (2018)
	Value-based	Castellini et al. (2019), Foerster et al. (2016), Jorge et al. (2016), Rashid et al. (2018), Son et al. (2019) and Sunehag et al. (2018)
	Centralized critic	Boao et al. (2019), Das et al. (2019), Foerster et al. (2018b), Lowe et al. (2017), Iqbal and Sha (2019), Wei et al. (2018) and Wu et al. (2018)
DTDE	Master-slave	Kong et al. (2017) and Kumar et al. (2017)
		Chen et al. (2016), Dobbe et al. (2017), Foerster et al. (2018b), Gupta et al. (2017), Jaderberg et al. (2019), Jaques et al. (2019), Leibo et al. (2017), Liu et al. (2019), Lyu and Amato (2020), Nguyen et al. (2017b) Omidshafiei et al. (2017), Palmer et al. (2018), Palmer et al. (2019), Srinivasan et al. (2018), Strouse et al. (2018) and Zheng et al. (2018a)

方案	方法	文献
CTCE		Gupta 等人 (2017) 和 Sunehag 等人 (2018)
CTDE	参数共享	Altalhi 和 Dayan(2019)、Chu 和 Ye(2017)、Gupta 等人 (2017)、Peng 等人 (2017)、Sukhbaatar 等人 (2016) 和 Sunehag 等人 (2018)
	基于价值的	Castellini 等人 (2019)、Foerster 等人 (2016)、Jorge 等人 (2016)、Rashid 等人 (2018)、Son 等人 (2019) 和 Sunehag 等人 (2018)
	集中式评判	Boao 等人 (2019)、Das 等人 (2019)、Foerster 等人 (2018b)、Lowe 等人 (2017)、Iqbal 和 Sha(2019)、Wei 等人 (2018) 和 Wu 等人 (2018)
DTDE(含集群分散方法)	主从式	Kong 等人 (2017) 和 Kumar 等人 (2017)
		Chen 等人 (2016)、Dobbe 等人 (2017)、Foerster 等人 (2018b)、Gupta 等人 (2017)、Jaderberg 等人 (2019)、Jaques 等人 (2019)、Leibo 等人 (2017)、Liu 等人 (2019)、Lyu 和 Amato(2020)、Nguyen 等人 (2017b)、Omidshafiei 等人 (2017)、Palmer 等人 (2018)、Palmer 等人 (2019)、Srinivasan 等人 (2018)、Strouse 等人 (2018) 和 Zheng 等人 (2018a)

and partially observable setting by rendering samples more stable and efficient. Similarly, Palmer et al. (2018) extended the experience replay of Deep Q-Networks with leniency, which associates stored state-action pairs with decaying temperature values that govern the amount of applied leniency. They showed that this induces optimism in value function updates and can overcome relative over-generalization. Another work by Palmer et al. (2019) proposed negative update intervals double-DQN as an mechanism that identifies and removes generated data from the replay buffer that leads to mis-coordination. Alike, Lyu and Amato 2020 proposed decentralized quantile estimators which identify non-stationary transition samples based on the likelihood of returns. Another work that aims to improve upon independent learners can be found in Zheng et al. (2018a) who used two auxiliary mechanisms, including a lenient reward approximation and a prioritized replay strategy.

同样，Palmer 等人 (2018) 扩展了深度 Q 网络的经验回放，加入了宽容度，这将存储的状态-动作对与衰减的温度值相关联，这些温度值决定了应用宽容度的数量。他们证明这会在价值函数更新中引入乐观主义，并能够克服相对过度泛化。Palmer 等人 (2019) 的另一项工作提出了负更新间隔双重 DQN 机制，该机制识别并从回放缓冲区中移除导致协调失误的生成数据。类似地，Lyu 和 Amato(2020) 提出了去中心化的分位数估计器，基于回报的可能性来识别非平稳转换样本。另一项旨在改进独立学习者的工作可以在 Zheng 等人 (2018a) 中找到，他们使用了两种辅助机制，包括宽容奖励近似和优先回放策略。

A different research direction can be seen in distributed population-based training schemes where agents are optimized through an online evolutionary process such that under-performing agents are substituted by mutated versions of better agents (Jaderberg et al. 2019; Liu et al. 2019).

另一个研究方向可以在分布式基于种群的训练方案中看到，其中代理通过在线进化过程进行优化，使得表现不佳的代理被表现更好的代理的突变版本所替代 (Jaderberg 等人，2019；Liu 等人，2019)。

3.2 Centralized training

3.2 集中式训练

The centralized training paradigm describes agent policies that are updated based on mutual information. While the sharing of mutual information between agents is enabled during the training, this additional information is then discarded at test time. The centralized training can be further differentiated into the centralized and decentralized execution scheme.

集中式训练范式描述了基于互信息更新的代理策略。在训练过程中，代理之间的互信息共享是可行的，但在测试时，这些额外信息会被丢弃。集中式训练可以进一步区分为集中执行和分布式执行的方案。

Definition 11 Centralized training centralized execution (CTCE) The CTCE scheme describes a centralized executor $\pi : \mathcal{O} \rightarrow P(\mathcal{U})$ modeling the joint policy that maps the collection of distributed observations to a set of distributions over individual actions.

定义 11 集中式训练集中式执行 (CTCE)CTCE 方案描述了一个集中执行器 $\pi : \mathcal{O} \rightarrow P(\mathcal{U})$ 模拟联合策略，该策略将分布式观察集合映射到一组关于个体动作的分布。

Some applications assume an unconstrained and instantaneous information exchange between agents. In such a setting, a centralized executor can be leveraged to learn the joint policy for all agents. The CTCE paradigm allows the straightforward employment of single-agent training methods such as actor-critics (Mnih et al. 2016) or policy gradient algorithms (Schulman et al. 2017) to multi-agent problems. An obvious flaw is that state-action spaces grow exponentially by the number of agents. To address the so-called curse of dimensionality, the joint model can be factored into individual policies for each agent. Gupta et al. (2017) represented the centralized executor as a set of independent sub-policies such that agents' individual action distributions are captured rather than the joint action distribution of all agents, i.e. the joint action distribution $P(\mathcal{U}) = \prod_i P\left(\mathcal{U}^i\right)$ is factored into independent action distributions. Next to the policy, the value function can be factored so that the joint value is decomposed into a sum of local value functions, e.g. the joint action-value function can be expressed by $Q_\pi\left(o^1, \ldots, o^N, u^1, \ldots, u^n\right) = \sum_i Q_\pi^i\left(o^i, u^i\right)$ as shown in Russell and Zimdars (2003). A recent approach for the value function factorization is investigated in Sunehag et al. (2018). However, a phenomenon called lazy agents may occur in the CTCE setting when one agent learns a good policy but a second agent has less incentive to learn a good policy, as his actions may hinder the first agent, resulting in a lower reward (Sunehag et al. 2018).

一些应用假设代理之间可以进行无约束且即时的信息交换。在这样的设置中，可以利用集中式执行器来学习所有代理的联合策略。CTCE 范式允许直接将单代理训练方法，如演员-评论家 (Mnih 等人，2016 年) 或策略梯度算法 (Schulman 等人，2017 年) 应用于多代理问题。一个明显的缺陷是，状态-动作空间

随着代理数量的增加而指数级增长。为了解决所谓的维度诅咒问题，可以将联合模型分解为每个代理的个体策略。Gupta 等人 (2017 年) 将集中式执行器表示为一系列独立的子策略，使得代理的个体动作分布被捕获，而不是所有代理的联合动作分布，即联合动作分布 $P(\mathcal{U}) = \prod_i P\left(\mathcal{U}^i\right)$ 被分解为独立的动作分布。除了策略之外，价值函数也可以被分解，使得联合价值被分解为局部价值函数的和，例如，联合动作价值函数可以由 $Q_\pi\left(o^1, \ldots, o^N, u^1, \ldots, u^n\right) = \sum_i Q_\pi^i\left(o^i, u^i\right)$ 表示，如 Russell 和 Zimdars(2003 年) 所示。Sunehag 等人 (2018 年) 研究了一种最近的价值函数分解方法。然而，在 CTCE 设置中，当其中一个代理学到了一个好的策略，而第二个代理由于他的动作可能会阻碍第一个代理，从而导致奖励降低，因此可能发生所谓的“懒惰代理”现象 (Sunehag 等人，2018 年)。

Although CTCE regards the learning problem as a single-agent case, we include the paradigm in this paper because the training schemes presented in the subsequent sections occasionally use CTCE as performance baseline and conduct comparisons.

尽管 CTCE 将学习问题视为单代理案例，但我们在本文中包含了这种范式，因为后续章节中呈现的训练方案有时会使用 CTCE 作为性能基线并进行比较。

Definition 12 Centralized training decentralized execution (CTDE) Each agent i holds an individual policy $\pi^i : \mathcal{O}^i \to P\left(\mathcal{U}^i\right)$ which maps local observations to a distribution over individual actions. During training, agents are endowed with additional information, which is then discarded at test time.

定义 12 集中训练分布式执行 (CTDE) 每个代理 i 持有一个个体策略 $\pi^i : \mathcal{O}^i \to P\left(\mathcal{U}^i\right)$ ，该策略将局部观察映射到个体动作的分布上。在训练过程中，代理会被赋予额外的信息，这些信息在测试时被丢弃。

The CTDE paradigm presents the state-of-the-art practice for learning with multiple agents (Kraemer and Banerjee 2016; Oliehoek et al. 2008). In classical MARL, such setting was utilized as joint action learners which has the advantage that perceiving joint actions a-posteriori discards the non-stationarity in the environment (Claus and Boutilier 1998). As of late, CTDE has been successful in MADRL approaches (Foerster et al. 2016; Jorge et al. 2016). Agents utilize shared computational facilities or other forms of communication to exchange information during training. By sharing mutual information, the training process can be eased and the learning speed can become superior when matched against independently trained agents (Foerster et al. 2018b). Moreover, agents can bypass non-stationarity when extra information about the selected actions is available to all agents during training such that the consequences of actions can be attributed to the respective agents. In what follows, we classify the CTDE literature according to the agent structure.

CTDE 范式是学习多代理的最新实践 (Kraemer 和 Banerjee 2016；Oliehoek 等人 2008)。在经典的多智能体强化学习 (MARL) 中，这样的设置被用作联合动作学习者，其优点在于感知到的事后联合动作可以消除环境中的非平稳性 (Claus 和 Boutilier 1998)。近年来，CTDE 在多智能体深度强化学习 (MADRL) 方法中取得了成功 (Foerster 等人 2016；Jorge 等人 2016)。代理在训练过程中使用共享的计算设施或其他形式的通信来交换信息。通过共享相互信息，训练过程可以简化，并且当与独立训练的代理相比时，学习速度可以提高 (Foerster 等人 2018b)。此外，当训练过程中所有代理都可获得关于所选动作的额外信息时，代理可以绕过非平稳性，这样动作的后果可以归因于各自的代理。接下来，我们根据代理结构对 CTDE 文献进行分类。

Homogeneous agents exhibit a common structure or the same set of skills, e.g. the same learning model or share common goals. Owning the same structure, agents can share parts of their learning model or experience with other agents. These approaches can scale well with the number of agents and may allow an efficient learning of behaviors. Gupta et al. (2017) showed that policies based on parameter sharing can be trained more efficiently and, thus, can outperform independently learned ones. Although agents own the same policy network, different agent behaviors can emerge because each agent perceives different observations at test time. It has been thoroughly demonstrated that parameter sharing can help to accelerate the learning progress (Ahilan and Dayan 2019; Chu and Ye 2017; Peng et al. 2017; Sukhbaatar et al. 2016; Sunehag et al. 2018). Next to parameter sharing, homogeneous agents can employ value-based methods where an approximation of the value function is learned based on mutual information. Agents profit from the joint actions and other agents' policies that are available during training and incorporate this extra information into centralized value functions (Foerster et al. 2016; Jorge et al. 2016). Such information is then discarded at test time. Many approaches consider the decomposition of a joint value function into combinations of individual value functions (Castellini et al. 2019; Rashid et al. 2018; Son et al. 2019; Sunehag et al. 2018). Through decomposition, each agent faces a simplified sub-problem of the original problem. Sunehag et al. (2018) showed that agents learning on local sub-problems scale better with the number of agents than CTCE or independent learners. We elaborate on value function-based factorization more detailed in Sect. 5.4 as an effective approach to tackle credit assignment problems.

同质代理表现出相同的结构或技能集合，例如，相同的学习模型或共享共同目标。由于拥有相同的结

构，代理可以与其他代理共享其学习模型或经验的一部分。这些方法可以随着代理数量的增加而良好地扩展，并可能允许高效的行为学习。Gupta 等人 (2017) 表明，基于参数共享的策略可以更高效地训练，因此，可以超越独立学习的策略。尽管代理拥有相同的策略网络，但在测试时由于每个代理感知到的观察结果不同，可能会出现不同的代理行为。已经充分证明，参数共享可以帮助加速学习进程 (Ahilan 和 Dayan 2019；Chu 和 Ye 2017；Peng 等人 2017；Sukhbaatar 等人 2016；Sunehag 等人 2018)。除了参数共享，同质代理可以采用基于价值的方法，其中价值函数的近似是基于相互信息学习的。代理从训练过程中的联合行动和其他代理的策略中获益，并将这些额外信息纳入集中价值函数中 (Foerster 等人 2016；Jorge 等人 2016)。此类信息在测试时被丢弃。许多方法考虑将联合价值函数分解为个体价值函数的组合 (Castellini 等人 2019；Rashid 等人 2018；Son 等人 2019；Sunehag 等人 2018)。通过分解，每个代理面临的是原始问题的简化子问题。Sunehag 等人 (2018) 表明，在局部子问题上学习的代理比 CTCE 或独立学习者随着代理数量的增加扩展得更好。我们在第 5.4 节详细阐述基于价值函数的分解作为一种有效解决信用分配问题的方法。

Heterogeneous agents, on the contrary, differ in structure and skill. An instance for heterogeneous policies can be seen in the extension of an actor-critic approach with a centralized critic, which allows information sharing to amplify the performance of individual agent policies. These methods can be distinguished from each other based on the representation of the critic. Lowe et al. (2017) utilized one centralized critic for each agent that is augmented with additional information during training. The critics are provided with information about every agent's policy, whereas the actors perceive only local observations. As a result, the agents do not depend on explicit communication and can overcome the non-stationarity in the environment. Likewise, Bono et al. (2019) trained multiple agents with individual policies that share information with a centralized critic and demonstrated that such setup might improve results on standard benchmarks. Besides the utilization of one critic for each agent, Foerster et al. (2018b) applied one centralized critic for all agents to estimate a counterfactual baseline function that marginalizes out a single agent's action. The critic is conditioned on the history of all agents' observations or, if available, on the true global state. Typically, actor-critic methods underlie a variance in the critic estimation that is further exacerbated by the number of agents. Therefore, Wu et al. (2018) proposed an action-dependent baseline which includes information from other agents to reduce the variance in the critic estimation function. Further works that incorporate one centralized critic for distributed policies can be found in Das et al. (2019), Iqbal and Sha (2019) and Wei et al. (2018).

与之相反，异质代理在结构和技能上存在差异。异质策略的一个实例可以在带有集中评估者的演员-评估者方法的扩展中看到，这允许信息共享以放大单个代理策略的性能。这些方法可以根据评估者的表示来相互区分。Lowe 等人 (2017 年) 为每个代理使用一个集中评估者，在训练过程中增加了额外的信息。评估者获得了关于每个代理策略的信息，而演员仅感知局部观察。因此，代理不依赖于显式通信，并且可以克服环境中的非平稳性。同样，Bono 等人 (2019 年) 训练了具有个体策略的多个代理，这些代理与集中评估者共享信息，并证明这种设置可能会在标准基准测试中提高结果。除了为每个代理使用一个评估者之外，Foerster 等人 (2018b 年) 为所有代理应用了一个集中评估者来估计一个反事实基线函数，该函数消除了单个代理的行为。评估者根据所有代理的观察历史或者如果可用的话，根据真实的全局状态进行条件化。通常，演员-评估者方法下，评估者的估计存在方差，而这种方差会随着代理数量的增加而进一步加剧。因此，Wu 等人 (2018 年) 提出了一种动作依赖的基线，它包括来自其他代理的信息以减少评估函数的方差。在 Das 等人 (2019 年)、Iqbal 和 Sha(2019 年) 以及 Wei 等人 (2018 年) 的作品中，可以找到将一个集中评估者纳入分布式策略的进一步研究。

Another way to perform decentralized execution is by employing a master-slave architecture, which can resolve coordination conflicts between multiple agents. Kong et al. (2017) applied a centralized master executor which shares information with decentralized slaves. In each time step, the master receives local information from the slaves and shares its internal state in return. The slaves compute actions conditioned on their local observation and the master's internal state. Similar approaches that make use of different levels of abstraction are hierarchical methods (Kumar et al. 2017) that operate at different time scales or levels of abstraction. We elaborate on hierarchical methods in more detail in Sect. 5.3.

另一种执行去中心化任务的方式是采用主从架构，这可以解决多个代理之间的协调冲突。Kong 等人 (2017 年) 应用了一个集中式的主执行器，该执行器与去中心化的从设备共享信息。在每一个时间步，主执行器接收从设备提供的本地信息，并分享其内部状态作为回报。从设备根据它们的本地观察和主执行器的内部状态计算动作。类似的方法还有使用不同抽象级别的分层方法 (Kumar 等人，2017 年)，它们在不同的时间尺度或抽象级别上运行。我们在第 5.3 节详细阐述分层方法。

4 Emergent patterns of agent behavior

4 代理行为出现的模式

Agents adjust their policy to maximize the task success and react to the behavioral changes of other agents. The dynamic interaction between multiple decision-makers, which simultaneously affects the state of the environment, can cause the emergence of specific behavioral patterns. An obvious way to influence the development of agent behavior is through the designed reward structure. By promoting incentives for cooperation, agents can learn team strategies where they try to collaborate and optimize upon a mutual goal. Agents support other agents since the cumulative reward for cooperation is greater than acting selfishly. On the contrary, if the appeals for maximizing the individual performance are larger than being cooperative, agents can learn greedy strategies and maximize their individual reward. Such competitive attitudes can yield high-level strategies like manipulating adversaries to gain an advantage. However, the boundaries between competition and cooperation can be blurred in the multi-agent setting. For instance, if one agent competes with other agents, it is sometimes useful to cooperate temporarily in order to receive a higher reward in the long run.

代理调整其策略以最大化任务成功并对他代理的行为变化作出反应。多个决策者之间的动态互动，同时影响环境状态，可能导致特定行为模式的出现。影响代理行为发展的一个明显方式是通过设计的奖励结构。通过促进合作的激励，代理可以学习团队策略，在团队策略中，它们尝试协作并针对共同目标进行优化。代理支持其他代理，因为合作的累积奖励大于自私行为的奖励。相反，如果追求个体表现最大化的呼声大于合作的呼声，代理可以学习贪婪策略并最大化其个体奖励。这种竞争态度可能导致操纵对手以获得优势等高级策略。然而，在多代理环境中，竞争和合作之间的界限可能变得模糊。例如，如果一个代理与其他代理竞争，有时暂时合作以在长期内获得更高的奖励是有用的。

In this section, we review the literature that is interested in developed agent behaviors. We differentiate occurring behaviors according to the reward structure (Sect. 4.1), the language between agents (Sect. 4.2), and the social context (Sect. 4.3). Table 2 summarizes the reviewed literature based on this classification. Note that we focus in this section not on works that introduce new methodologies but on literature that analyzes the emergent behavioral patterns.

在本节中，我们回顾了关注开发代理行为的文献。我们根据奖励结构 (第 4.1 节)、代理之间的语言 (第 4.2 节) 和社会背景 (第 4.3 节) 区分行为的发生。表 2 根据这种分类总结了回顾的文献。请注意，在本节中，我们关注的不是引入新方法的作品，而是分析新兴行为模式的文献。

Table 2 Overview of MADRL papers that investigate emergent patterns of agent behavior

表 2 概述了研究代理行为新兴模式的 MADRL 论文

Emergence	Setting	Literature
Reward structure	Cooperative	Diallo et al. (2017), Leibo et al. (2017) and Tampuu et al. (2017)
	Competitive	Bansal et al. (2018), Leibo et al. (2017), Liu et al. (2019) and Tampuu et al. (2017)
	Intrinsic rewards	Baker et al. (2020), Hughes et al. (2018), Jaderberg et al. (2019), Jaques et al. (2019), Jaques et al. (2018), Peysakhovich and Lerer (2018), Sukhbaatar et al. (2017), Wang et al. (2019) and Wang et al. (2020b)
Language	Referential games	Choi et al. (2018), Evtimova et al. (2018), Havrylov and Titov (2017), Jorge et al. (2016), Lazaridou et al. (2017), Lazaridou et al. (2018), Lee et al. (2017) and Mordatch and Abbeel (2018)
Social context	Dialogues	Cao et al. (2018), Das et al. (2017) and Lewis et al. (2017)
	Commons dilemmas	Foerster et al. (2018a), Jaques et al. (2018), Jaques et al. (2019), Leibo et al. (2017) and Lerer and Peysak hovich (2017)
	Public good dilemmas	Pérolat et al. (2017), Hughes et al. (2018) and Zhu and Kirley (2019)

产生	设置	文献
奖励结构	合作	Diallo 等人 (2017 年)，Leibo 等人 (2017 年) 和 Tampuu 等人 (2017 年)
	竞争	Bansal 等人 (2018 年)，Leibo 等人 (2017 年)，Liu 等人 (2019 年) 和 Tampuu 等人 (2017 年)
	内在奖励	Baker 等人 (2020 年)，Hughes 等人 (2018 年)，Jaderberg 等人 (2019 年)，Jaques 等人 (2019 年)，Jaques 等人 (2018 年)，Peysakhovich 和 Lerer(2018 年)，Sukhbaatar 等人 (2017 年)，Wang 等人 (2019 年) 和 Wang 等人 (2020b)
语言	指示性游戏	Choi 等人 (2018)，Evtimova 等人 (2018)，Havrylov 和 Titov(2017)，Jorge 等人 (2016)，Lazaridou 等人 (2017)，Lazaridou 等人 (2018)，Lee 等人 (2017) 和 Mordatch 与 Abbeel(2018)
社会背景	对话	Cao 等人 (2018)，Das 等人 (2017) 和 Lewis 等人 (2017)
	公地困境	Foerster 等人 (2018a)，Jaques 等人 (2018)，Jaques 等人 (2019)，Leibo 等人 (2017) 和 Lerer 与 Peysakhovich(2017)
	公共物品困境	Pérolat 等人 (2017)，Hughes 等人 (2018) 和 Zhu 与 Kirley(2019)

4.1 Reward structure

4.1 奖励结构

The primary factor that influences the emergence of agent behavior is the reward structure. If the reward for mutual cooperation is larger than individual reward maximization, agents tend to learn policies that seek to collaboratively solve the task. In particular, Leibo et al. (2017) compared the magnitude of the team reward in relation to the individual agent reward. They showed that the higher the numerical team reward is compared to the individual reward, the greater is the willingness to collaborate with other agents. The work by Tampuu et al. (2017) demonstrated that punishing the whole team of agents for the failure of a single agent can also cause cooperation. Agents learn policies to avoid the malfunction of an individual, support other agents to prevent failure, and improve the performance of the whole team. Similarly, Diallo et al. (2017) used the Pong video game to investigate the coordination between agents and examined how developed behaviors change regarding the reward function. For a comprehensive

review of learning in cooperative settings, one can consider the article by Panait and Luke (2005) for classical MARL and Oroojlooyjadid and Hajinezhad (2019) for recent MADRL.

影响代理行为出现的主要因素是奖励结构。如果相互合作的奖励大于个体奖励最大化，代理倾向于学习寻求协作解决问题的策略。特别是 Leibo 等人 (2017 年) 比较了团队奖励与个体代理奖励的大小。他们表明，团队奖励的数值相对于个体奖励越高，与其他代理合作的意愿就越大。Tampuu 等人 (2017 年) 的工作表明，对单个代理失败的整个团队进行惩罚也可以导致合作。代理学习避免个体故障的策略，支持其他代理以防止失败，并提高整个团队的表现。同样，Diallo 等人 (2017 年) 使用 Pong 视频游戏来研究代理之间的协调，并检查了根据奖励函数发展的行为如何变化。对于合作环境中学习的全面回顾，可以考虑 Panait 和 Luke(2005 年) 关于经典 MARL 的文章以及 Oroojlooyjadid 和 Hajinezhad(2019 年) 关于近期 MADRL 的文章。

In contrast to the cooperative scenario, one can value individual performance greater than the collaboration among agents. A competitive setting motivates agents to outperform their adversary counterparts. Tampuu et al. (2017) used the video game Pong and manipulated the rewarding structure to examine the emergence of agent behavior. They showed that the higher the reward for competition, the more likely an agent tries to outplay its opponents by using techniques such as wall bouncing or faster ball speed. Employing such high-level strategies to overwhelm the adversary maximizes the individual reward. Similarly, Bansal et al. (2018) investigated competitive scenarios, where agents competed in a 3D world with simulated physics to learn locomotion skills such as running, blocking, or tackling other agents with arms and legs. They argued that adversarial training could help to learn more complex agent behaviors than the environment can exhibit. Likewise, the works of Leibo et al. (2017) and Liu et al. (2019) investigated the emergence of behaviors due to the reward structure in competitive scenarios.

与合作场景相比，人们可能会更重视个体表现而不是代理之间的协作。竞争环境激励代理超越其对手方。Tampuu 等人 (2017 年) 使用了视频游戏 Pong，并操纵了奖励结构来研究代理行为的出现。他们表明，竞争的奖励越高，代理就越有可能尝试通过使用如墙壁反弹或更快球速的技术来超越对手。采用这种高级策略压倒对手可以最大化个体奖励。同样，Bansal 等人 (2018 年) 研究了一些竞争场景，其中代理在一个具有模拟物理的 3D 世界中竞争，以学习如跑步、阻挡或用胳膊和腿拦截其他代理的移动技能。他们认为，对抗性训练可以帮助学习比环境所能展示的更复杂的代理行为。同样，Leibo 等人 (2017 年) 和 Liu 等人 (2019 年) 研究了几种由于竞争场景中的奖励结构而产生的行为。

If the rewards appear in sparse frequency, agents can be equipped with intrinsic reward functions that provide denser feedback signals and, thus, can overcome the sparsity or even the absence of external rewards. One way to realize this is with intrinsic motivation, which is based on the concept of maximizing an internal reinforcement signal by actively discovering novel or surprising patterns (Chentanez et al. 2005; Oudeyer and Kaplan 2007; Schmidhuber 2010). Intrinsic motivation encourages agents to explore states that have been scarcely or never visited and to perform novel actions in those states. Most approaches of intrinsic motivation can be broadly divided into two categories (Pathak et al. 2017). First, agents are encouraged to explore unknown states where the novelty of states is measured by a model that captures the distribution of visited environment states (Bellemare et al. 2016). Second, agents can be motivated to reduce the uncertainty about the consequences of their own actions. The agent builds a model that learns the dynamics of the environment by lowering the prediction error of the follow-up states with respect to the taken actions. The uncertainty indicates the novelty of new experience since the model can only be accurate in states which it has already encountered or can generalize from previous knowledge (Hout-hooft et al. 2016; Pathak et al. 2017). For a recent survey on intrinsic motivation in RL, one can regard the paper by Aubret et al. (2019). The concept of intrinsic motivation was transferred to the multi-agent domain by Sequeira et al. (2011), who studied the motivational impact on multiple agents. Investigations on the emergence of agent behavior based on intrinsic rewards have been abundantly conducted in Baker et al. (2020), Hughes et al. (2018), Jaderberg et al. (2019), Jaques et al. (2018), Jaques et al. (2019), Peysakhovich and Lerer (2018), Sukhbaatar et al. (2017), Wang et al. (2019) and Wang et al. (2020b).

如果奖励出现的频率稀疏，可以给智能体配备内在奖励函数，这些函数提供更密集的反馈信号，从而能够克服外部奖励的稀疏性甚至缺失。实现这一点的途径之一是内在动机，它基于通过主动发现新颖或令人惊讶的模式来最大化内部强化信号的概念 (Chentanez 等人，2005 年；Oudeyer 和 Kaplan，2007 年；Schmidhuber，2010 年)。内在动机鼓励智能体探索很少或从未访问过的状态，并在那些状态下执行新颖的动作。内在动机的大多数方法可以大致分为两类 (Pathak 等人，2017 年)。首先，鼓励智能体探索未知状态，其中状态的 novelty 是通过捕获访问过的环境状态分布的模型来衡量的 (Bellemare 等人，2016 年)。其次，智能体可以被激励去减少对自己行为后果的不确定性。智能体构建一个模型，通过降低相对于采取的动作的后续状态的预测误差来学习环境的动态。不确定性表示新经验的新颖性，因为模型只能在已经遇到或能够从先前知识泛化的状态中准确 (Hout-hooft 等人，2016 年；Pathak 等人，2017 年)。关于强化学习中内在动机的最新调研，可以参考 Aubret 等人 (2019 年) 的论文。内在动机的概念

被 Sequeira 等人 (2011 年) 转移到多智能体领域，他们研究了动机对多个智能体的影响。基于内在奖励的智能体行为出现的调查已经在 Baker 等人 (2020 年)，Hughes 等人 (2018 年)，Jaderberg 等人 (2019 年)，Jaques 等人 (2018 年)，Jaques 等人 (2019 年)，Peysakhovich 和 Lerer(2018 年)，Sukhbaatar 等人 (2017 年)，Wang 等人 (2019 年) 和 Wang 等人 (2020b) 中大量进行。

4.2 Language

4.2 语言

The development of language corpora and communication skills of autonomous agents attracts great attention within the community. For one, the behavior that emerges during the deployment of abstract language as well as the learned composition of multiple words to form meaningful contexts is of interest (Kirby 2002). Deep learning methods have widened the scope of computational methodologies for investigating the development of language between dynamic agents (Lazaridou and Baroni 2020). For building rich behaviors and complex reasoning, communication based on high-dimensional data like visual perception is a widespread practice (Antol et al. 2015). In the following, we focus on works that investigate the emergence of language and analyze behavior. Papers that propose new methodologies for developing communication protocols are discussed in Sect. 5.2. We classify the learning of language according to the performed task and the type of interaction the agents pursue. In particular, we differentiate between referential games and dialogues.

语言语料库的开发和自主代理的通信技能在学术界引起了极大的关注。一方面，抽象语言部署期间出现的行为以及学习多个单词组合形成有意义语境的过程令人感兴趣 (Kirby 2002)。深度学习方法扩大了计算方法的研究范围，用于探究动态代理之间的语言发展 (Lazaridou 和 Baroni 2020)。为了构建丰富的行为和复杂的推理，基于高维数据 (如视觉感知) 的通信是一种普遍的做法 (Antol 等人 2015)。接下来，我们重点关注研究语言出现和分析行为的著作。提出新的通信协议开发方法论的论文在 5.2 节中讨论。我们根据执行的任务和代理追求的交互类型对语言学习进行分类。特别是，我们区分了指代游戏和对话。

The former, referential games, describe cooperative games where the speaking agent communicates an objective via messages to another listening agent. Lazaridou et al. (2017) showed that agents could learn communication protocols solely through interaction. For a meaningful information exchange, agents evolved semantic properties in their language. A key element of the study was to analyze if the agents' interactions are interpretable for humans, showing limited yet encouraging results. Likewise, Mordatch and Abbeel (2018) investigated the emergence of abstract language that arises through the interaction between agents in a physical environment. In their experiments, the agents should learn a discrete set of vocabulary by solving navigation tasks through communication. By involving more than three agents in the conversation and by penalizing an arbitrary size of vocabulary, agents agreed on a coherent set of vocabulary and discouraged ambiguous words. They also observed that agents learned a syntax structure in the communication protocol that is consistent in vocabulary usage. Another work by Li and Bowling (2019) found out that compositional languages are easier to communicate with other agents than languages with less structure. In addition, changing listening agents during the learning can promote the emergence of language grounded on a higher degree of structure. Many studies are concerned with the development of communication in referential games grounded on visual perception as it can be found in Choi et al. (2018), Evtimova et al. (2018), Havrylov and Titov (2017), Jorge et al. (2016), Lazaridou et al. (2018) and Lee et al. (2017). Further works consider the development of communication in social dilemmas (Jaques et al. 2018,

前者，指代游戏，描述的是一种合作游戏，其中发言的代理通过消息与另一个倾听的代理沟通目标。Lazaridou 等人 (2017 年) 展示了代理能够仅仅通过互动学习通信协议。为了进行有意义的信息交换，代理在他们的语言中进化出语义属性。研究的一个关键要素是分析代理的互动是否对人类可解释，展示了有限但令人鼓舞的结果。同样，Mordatch 和 Abbeel(2018 年) 研究了在物理环境中的代理互动中产生的抽象语言的出现。在他们的实验中，代理应该通过通信解决导航任务来学习一组离散的词汇。通过在对话中涉及三个以上的代理并对任意大小的词汇进行惩罚，代理达成了一组连贯的词汇，并阻止了模糊词汇的使用。他们还观察到代理在通信协议中学习了一种语法结构，这种结构在词汇使用上是一致的。Li 和 Bowling(2019 年) 的另一项研究发现，组合语言与其他代理通信比结构较少的语言更容易。此外，在学习过程中更换倾听代理可以促进基于更高结构度的语言的出现。许多研究关注基于视觉感知的指代游戏中通信的发展，如 Choi 等人 (2018 年)、Evtimova 等人 (2018 年)、Havrylov 和 Titov(2017 年)、Jorge 等人 (2016 年)、Lazaridou 等人 (2018 年) 和 Lee 等人 (2017 年) 的研究中可以发现。进一步的工作考虑了在社会困境中通信的发展 (Jaques 等人 2018 年，2019).

As the second category, we describe the emergence of behavioral patterns in communication while

conducting dialogues. One type of dialogue are negotiations in which agents pursue to agree on decisions. In a study about negotiations with natural language, Lewis et al. (2017) showed that agents could master linguistic and reasoning problems. Two agents were both shown a collection of items and were instructed to negotiate about how to divide the objects among both agents. Each agent was expected to maximize the value of the bargained objects. Eventually, the agents learned to use high-level strategies such as deception to accomplish higher rewards over their opponents. Similar studies concerned with negotiations are covered in Cao et al. (2018) and He et al. (2018). Another type of dialogue are scenarios where the emergence of communication is investigated in a question-answering style as shown by Das et al. (2017). One agent received an image as input and was instructed to ask questions about the shown image while the second agent responded, both in natural language.

作为第二类，我们描述了在对话过程中行为模式的出现。一种对话类型是谈判，其中代理追求就决策达成一致。在关于自然语言谈判的研究中，Lewis 等人 (2017 年) 展示了代理能够掌握语言和推理问题。两个代理都看到了一组物品，并被指示就如何在两个代理之间分配物品进行谈判。每个代理都期望最大化谈判物品的价值。最终，代理学会了使用高级策略，如欺骗，以在对手中获得更高的回报。类似关于谈判的研究在 Cao 等人 (2018 年) 和 He 等人 (2018 年) 的文章中有所涉及。另一种对话类型是像 Das 等人 (2017 年) 展示的那样，在问答风格中研究沟通出现的情况。一个代理接收到一个图像作为输入，并被指示就显示的图像提出问题，而第二个代理以自然语言回应。

Many of the above-mentioned papers report that utilizing a communication channel can increase task performance in terms of the cumulative reward. However, numerical performance measurements provide evidence but do not give insights about the communication abilities learned by the agents. Therefore, Lowe et al. (2019) surveyed metrics which are applied to assess the quality of learned communication protocols and provided recommendations about the usage of such metrics. Based on that, Eccles et al. (2019) proposed to incorporate inductive bias into the learning objective of agents, which could promote the emergence of a meaningful communication. They showed that inductive bias could lead to improved results in terms of interpretability.

许多上述论文报告称，利用通信通道可以提高任务表现，即累积奖励。然而，数值性能测量提供了证据，但并没有提供关于代理学到的沟通能力的洞见。因此，Lowe 等人 (2019 年) 调研了用于评估学习到的通信协议质量的指标，并就这些指标的使用提供了建议。基于此，Eccles 等人 (2019 年) 提议将归纳偏置纳入代理的学习目标，这可以促进有意义沟通的出现。他们展示了归纳偏置可以导致在解释性方面得到改进的结果。

4.3 Social context

4.3 社会背景

Next to the reward structure and language, the research community actively investigates the emerging agent behaviors in social contexts. Akin to humans, artificial agents can develop strategies that exploit patterns in complex problems and adapt behaviors in response to others (Baker et al. 2020; Jaderberg et al. 2019). We differentiate the following literature along different dimensions, such as the type of social dilemma and the examined psychological variables.

研究社区积极调查在社会环境中出现的代理行为。类似于人类，人工代理能够发展策略来利用复杂问题中的模式，并对他人做出反应来调整行为 (Baker 等人 2020 年；Jaderberg 等人 2019 年)。我们根据不同的维度区分以下文献，例如社会困境的类型和检查的心理变量。

Social dilemmas have long been studied as conflict scenario in which agents gauge between individualistic and collective profits (Crandall and Goodrich 2011; De Cote et al. 2006). The tension between cooperation and defection is evaluated as an atomic decision according to the numerical values of a pay-off matrix. This pay-off matrix satisfies inequalities in the reward function such that agents must decide between cooperation, to benefit as a whole team, or defection, to maximize selfish performance. To temporally extend matrix games, sequential social dilemmas have been introduced to investigate long-term strategic decisions of agent policies rather than short-term actions (Leibo et al. 2017). The arising behaviors in these dilemmas can be classified along psychological variables known from human interaction (Lange et al. 2013) such as the gain of individual benefits (Lerer and Peysakhovich 2017), the fear of future consequences (Pérolat et al. 2017), the assessment of the impact on another agent's behavior (Jaques et al. 2018, 2019), the trust between agents (Pinyol and Sabater-Mir 2013; Ramchurn et al. 2004; Yu et al. 2013), and the impact of emotions on the decision-making (Moerland et al. 2018; Yu et al. 2013).

社会困境长期以来被研究为冲突场景，其中的代理在个人利益和集体利益之间权衡 (Crandall 和

Goodrich 2011 年；De Cote 等人 2006 年)。合作与背叛之间的紧张关系根据支付矩阵的数值作为原子决策来评估。这个支付矩阵满足奖励函数中的不等式，使得代理必须决定是合作以使整个团队受益，还是背叛以最大化自私表现。为了在时间上扩展矩阵游戏，顺序社会困境被引入，以研究代理策略的长期战略决策，而不仅仅是短期行动 (Leibo 等人 2017 年)。这些困境中产生的行为可以根据已知的人类互动中的心理变量进行分类 (Lange 等人 2013 年)，例如个人利益的获得 (Lerer 和 Peysakhovich 2017 年)，对未来后果的恐惧 (Pérolat 等人 2017 年)，对另一个代理行为影响的评估 (Jaques 等人 2018 年、2019 年)，代理之间的信任 (Pinyol 和 Sabater-Mir 2013 年；Ramchurn 等人 2004 年；Yu 等人 2013 年)，以及情绪对决策的影响 (Moerland 等人 2018 年；Yu 等人 2013 年)。

Kollock (1998) divided social dilemmas into commons dilemmas and public goods dilemmas. The former, commons dilemmas describe the trade-off between individualistic short-term benefits and long-term common interests on a task that is shared by all agents. Recent works on the commons dilemma can be found in Foerster et al. (2018a), Leibo et al. (2017) and Lerer and Peysakhovich (2017). In public goods dilemmas, agents face a scenario where common-pool resources are constrained and oblige a sustainable use of resources. The phenomenon called the tragedy of commons predicts that self-interested agents fail to find socially positive equilibria, which eventually results in the over-exploitation of the common resources (Hardin 1968). Investigations on the trial-and-error learning in common-pool resource scenarios with multiple decision-makers are covered in Hughes et al. (2018), Pérolat et al. (2017) and Zhu and Kirley (2019).

Kollock(1998) 将社会困境分为共有资源困境和公共物品困境。前者，即共有资源困境，描述了个体短期利益与长期共同利益之间的权衡，这个任务是被所有代理共享的。关于共有资源困境的最新研究可以在 Foerster 等人 (2018a)、Leibo 等人 (2017) 以及 Lerer 和 Peysakhovich(2017) 的作品中找到。在公共物品困境中，代理面临的是一个场景，其中共有池资源受限并要求可持续地使用资源。被称为公地悲剧的现象预测，自利的代理无法找到社会正面均衡，最终导致共有资源的过度开发 (Hardin 1968)。关于在多个决策者共同参与下的共有池资源场景中尝试-错误学习的研究，可以在 Hughes 等人 (2018)、Pérolat 等人 (2017) 和 Zhu 与 Kirley(2019) 的作品中找到。

5 Current challenges

5 当前的挑战

In this section, we depict several challenges that arise in the multi-agent RL domain and, thus, are currently under active research. We approach the problem of non-stationarity (Sect. 5.1) due to the presence of multiple learners in a shared environment and review literature regarding the development of communication skills (Sect. 5.2). We further investigate the challenge of learning coordination (Sect. 5.3). Then, we survey the difficulty of attributing rewards to specific agents as the credit assignment problem (Sect. 5.4) and examine scalability issues (Sect. 5.5), which increase with the number of agents. Finally, we consider environments where states are only partially observable (Sect. 5.6). While some challenges are omnipresent in the MARL domain, such as non-stationarity or scal-ability, others like the credit assignment problem or the learning of coordination and communication are prevailing in the cooperative setting.

在本节中，我们描述了多智能体强化学习领域中出现的几个挑战，这些问题目前正处于积极研究之中。我们探讨了由于多个学习者在共享环境中的存在而非平稳性问题 (第 5.1 节)，并回顾了关于发展沟通技能的文献 (第 5.2 节)。我们还进一步研究了学习协调的挑战 (第 5.3 节)。然后，我们调研了将奖励归因于特定智能体的困难，即信用分配问题 (第 5.4 节)，并检查了随智能体数量增加而增大的可扩展性问题 (第 5.5 节)。最后，我们考虑了状态仅部分可观察的环境 (第 5.6 节)。虽然某些挑战在多智能体强化学习领域中无处不在，如非平稳性或可扩展性，但其他挑战，如信用分配问题或协调与沟通的学习，在合作环境中尤为突出。

We aim to provide a holistic overview of the contemporary challenges that constitute the landscape in reinforcement learning with multiple agents and survey treatments that were suggested in recent works. In particular, we focus on those challenges which are currently under active research and where progress has been accomplished recently. There are still open problems that have not been or partially addressed so far. Such problems are discussed in Sect. 6. Deliberately, we do not regard challenges that also persist in the single-agent domain, such as sparse rewards or the exploration-exploitation dilemma. We refer the interested reader for an overview of those topics to the articles of Arulkumaran et al. (2017) and Li (2018). Much of the surveyed literature cannot be assigned to one particular but rather to several of the proposed challenges. Hence, we associate the subsequent literature to the one challenge which we believe best addresses it (Table 3).

我们旨在提供关于多智能体强化学习当前挑战的全面概述，并调研了近期作品中提出的处理方法。特别是，我们关注那些目前正处于积极研究且近期取得进展的挑战。仍有一些尚未解决或部分解决的问题。这些问题在第 6 节中进行了讨论。有意地，我们没有考虑在单智能体领域中同样存在的挑战，如稀疏奖励或探索-利用困境。感兴趣的读者可以查阅 Arulkumaran 等人 (2017) 和 Li(2018) 的文章，以获得这些主题的概述。调研的许多文献不能归入一个特定的挑战，而是与几个提出的挑战相关。因此，我们将后续文献与我们认为最能解决该挑战的问题关联起来 (表 3)。

5.1 Non-stationarity

5.1 非平稳性

One major problem resides in the presence of multiple agents that interact within a shared environment and learn simultaneously. Due to the co-adaption, the environment dynamics appear non-stationary from the perspective of a single agent. Thus, agents face a moving target problem if they are not provided with additional knowledge about other agents. As a result, the Markov assumption is violated, and the learning constitutes an inherently difficult problem (Hernandez-Leal et al. 2017; Laurent et al. 2011). The naïve approach is to neglect the adaptive behavior of agents. One can either ignore the existence of other agents (Matignon et al. 2012b) or discount the adaptive behavior by assuming the others' behavior to be static or optimal (Lauer and Riedmiller 2000). By making such assumptions, the agents are considered as independent learners, and traditional single-agent reinforcement algorithms can be applied. First attempts have been studied in Claus and Boutilier (1998) and Tan (1993), which showed that independent learners could perform well in simple deterministic environments. However, in complex or stochastic environments, independent learners often result in poor performance (Lowe et al. 2017; Matignon et al. 2012b). Moreover, Lanctot et al. (2017) argued that independent learners could over-fit to other agents' policies during the training and, thus, may fail to generalize at test time.

一个主要问题在于存在多个代理在共享环境中相互作用并同时学习。由于共同适应，环境动态对于单个代理来说呈现出非平稳性。因此，如果代理没有获得有关其他代理的额外知识，它们将面临一个移动目标问题。结果，马尔可夫假设被违反，学习成为一个本质上困难的问题 (Hernandez-Leal 等人，2017；Laurent 等人，2011)。一种天真方法是忽略代理的自适应行为。人们可以忽略其他代理的存在 (Matignon 等人，2012b)，或者通过假设其他的行为是静态的或最优的来打折代理的自适应行为 (Lauer 和 Riedmiller，2000)。通过做出这样的假设，代理被视为独立的学习者，传统的单一代理强化算法可以应用。最初的尝试已在 Claus 和 Boutilier(1998) 以及 Tan(1993) 中研究过，这些研究显示独立学习者在简单的确定性环境中表现良好。然而，在复杂或随机环境中，独立学习者往往导致性能不佳 (Lowe 等人，2017；Matignon 等人，2012b)。此外，Lanctot 等人 (2017) 认为独立学习者在训练过程中可能会过度适应其他代理的策略，因此可能在测试时无法泛化。

In the following, we review literature, which addresses the non-stationarity in a multi-agent environment, and categorize the approaches into those with experience replay, centralized units, and meta-learning. A similar categorization proposed Papoudakis et al. (2019). We identify further approaches which cope with non-stationarity by establishing

接下来，我们回顾了处理多代理环境中非平稳性的文献，并将方法分类为使用经验回放、集中单元和元学习的方法。Papoudakis 等人 (2019) 提出了一个类似分类。我们识别了其他处理非平稳性的方法，通过建立

Table 3 Overview of MADRL challenges and approaches proposed in recent literature

表 3 MADRL 挑战和近期文献中提出的方法概述

Challenge	Approach	Literature
Non-stationarity	Experience replay	Foerster et al. (2017), Palmer et al. (2018), Tang et al. (2018), and Zheng et al. (2018a)
	Centralized training	Bono et al. (2019), Foerster et al. (2016), Foerster et al. (2018b), Iqbal and Sha (2019), Jorge et al. (2016), Lowe et al. (2017), Rashid et al. (2018), and Wei et al. (2018)
	Meta-learning	Al-Shedivat et al. (2018), and Rabinowitz et al. (2018)
Communication	Broadcasting	Foerster et al. (2016), Peng et al. (2017), and Sukhbaatar et al. (2016)
	Targeted	Das et al. (2019), Hoshen (2017), Jain et al. (2019), Jiang and Lu (2018), and Singh et al. (2019)
	Networked	Chu et al. (2020), Chu et al. (2020), Qu et al. (2020), Zhang et al. (2018), and Zhang et al. (2019)
Coordination	Extensions	Celikyilmaz et al. (2018), Jaques et al. (2018), Jaques et al. (2019), Kim et al. (2019), Li et al. (2019b), Singh et al. (2019), and Wang et al. (2020c)
	Independent learners	Foerster et al. (2018b), Lyu and Amato (2020), Omidshafiei et al. (2017), Palmer et al. (2018), Palmer et al. (2019), Sunehag et al. (2018), and Zheng et al. (2018a)
	Constructing models	Barde et al. (2019), Everett and Roberts (2018), Foerster et al. (2018a), Foerster et al. (2019), Grover et al. (2018), He et al. (2016), Hong et al. (2017), Hoshen (2017), Jaques et al. (2019), Le et al. (2017), Letcher et al. (2019), Raileanu et al. (2018), Tacchetti et al. (2019), Yang et al. (2018a), and Zheng et al. (2018b)
	Hierarchical methods	Ahilan and Dayan (2019), Cai et al. (2013), Han et al. (2019), Jaderberg et al. (2019), Kumar et al. (2017), Lee et al. (2020), Ma and Wu (2020), Tang et al. (2018), and Vezhnevets et al. (2019)
	Decomposition	Castellini et al. (2019), Chen et al. (2018), Nguyen et al. (2017b), Rashid et al. (2018), Son et al. (2019) Sunehag et al. (2018), Wang et al. (2020a), Wang et al. (2020c), Yang et al. (2018b)
Credit assignment	Marginalization	Foerster et al. (2018b), Nguyen et al. (2018), and Wu et al. (2018)
	Inverse RL	Barrett et al. (2017), Le et al. (2017), Lin et al. (2018), Song et al. (2018), and Yu et al. (2019)
Scalability	Knowledge Reuse	Baker et al. (2020), Da Silva et al. (2017), Da Silva and Costa (2017), Gupta et al. (2017), Hernandez-Leal et al. (2019), Jiang and Lu (2018), Long et al. (2020), Loketaa et al. (2019), Narvekar et al. (2016), Omidshafiei et al. (2019), Peng et al. (2017), Sukhbaatar et al. (2016), Sukhbaatar et al. (2017), Sunehag et al. (2018), and Svetlik et al. (2017)
	Complexity reduction	Chen et al. (2018), Liu et al. (2018), Nguyen et al. (2017a), Nguyen et al. (2017b), and Yang et al. (2018b)
	Robustness	Baker et al. (2020), Bansal et al. (2018), Berner et al. (2019), Glorot et al. (2020), Heinrich and Silver (2016), Lanctot et al. (2017), Li et al. (2019a), Lin et al. (2020), Lowe et al. (2017), Pinto et al. (2017), Raghu et al. (2018), Silver et al. (2016), Silver et al. (2018), Spooner and Savani (2020), and Sukhbaatar et al. (2017)

挑战	方法	文献
非平稳性	经验回放	Foerster 等人 (2017 年)，Palmer 等人 (2018 年)，Tang 等人 (2018 年)，以及 Zheng 等人 (2018a)
	集中训练	Bono 等人 (2019 年)，Foerster 等人 (2016 年)，Foerster 等人 (2018b)，Iqbal 和 Sha(2019 年)，Jorge 等人 (2016 年)，Lowe 等人 (2017 年)，Rashid 等人 (2018 年)，以及 Wei 等人 (2018 年)
	元学习	Al-Shedivat 等人 (2018 年)，以及 Rabinowitz 等人 (2018 年)
通信	广播	Foerster 等人 (2016 年)，Peng 等人 (2017 年)，以及 Sukhbaatar 等人 (2016 年)
	定向	Das 等人 (2019)，Hoshen(2017)，Jain 等人 (2019)，Jiang 和 Lu(2018)，以及 Singh 等人 (2019)
	网络化	Chu 等人 (2020)，Chu 等人 (2020)，Qu 等人 (2020)，Zhang 等人 (2018)，以及 Zhang 等人 (2019)
协调	扩展	Celikyilmaz 等人 (2018)，Jaques 等人 (2018)，Jaques 等人 (2019)，Kim 等人 (2019)，Li 等人 (2019b)，Singh 等人 (2019)，以及 Wang 等人 (2020c)
	独立学习者	Foerster 等人 (2018b)，Lyu 和 Amato(2020)，Omidshafiei 等人 (2017)，Palmer 等人 (2018)；Palmer 等人 (2019)，Sunehag 等人 (2018)，以及 Zheng 等人 (2018a)
	构建模型	Barde 等人 (2019)，Everett 和 Roberts(2018)，Foerster 等人 (2018a)，Foerster 等人 (2019)，Grover 等人 (2018)，He 等人 (2016)，Hong 等人 (2017)，Hoshen(2017)，Jaques 等人 (2019)，Le 等人 (2017)，Letcher 等人 (2019)，Raileanu 等人 (2018)，Tacchetti 等人 (2019)，Yang 等人 (2018a)，以及 Zheng 等人 (2018b)
	分层式方法	Ahilan 和 Dayan(2019)，Cai 等人 (2013)，Han 等人 (2019)，Jaderberg 等人 (2019)，Kumar 等人 (2017)，Lee 等人 (2020)，Ma 和 Wu(2020)，Tang 等人 (2018) 以及 Vezhnevets 等人 (2019)
	分解	Castellini 等人 (2019)，Chen 等人 (2018)，Nguyen 等人 (2017b)，Rashid 等人 (2018)，Son 等人 (2019)，Sunehag 等人 (2018)，Wang 等人 (2020a)，Wang 等人 (2020c)，Yang 等人 (2018b)
信用分配	边缘化	Foerster 等人 (2018b)，Nguyen 等人 (2018) 和 Wu 等人 (2018)
	逆强化学习	Barrett 等人 (2017)，Le 等人 (2017)，Lin 等人 (2018)，Song 等人 (2018) 和 Yu 等人 (2019)
可扩展性	知识重用	Baker 等人 (2020)；Da Silva 等人 (2017)，Da Silva 和 Costa(2017)，Gupta 等人 (2017)；Hernandez Leal 等人 (2019)，Jiang 和 Lu(2018)，Long 等人 (2020)，Loketaa 等人 (2019)，Narvekar 等人 (2016)，Omidshafiei 等人 (2019)，Peng 等人 (2017)，Sukhbaatar 等人 (2016)；Sukhbaatar 等人 (2017)；Sunehag 等人 (2018) 和 Svetlik 等人 (2017)
	复杂性降低	陈等人 (2018)，聂等人 (2018)；拉克等人 (2017a)，聂等人 (2017b)，和等人 (2018b)
	鲁棒性	贝克等人 (2020)，班萨尔等人 (2018)，伯纳尔等人 (2019)，格莱夫等人 (2020)，海因里希和银 (2016)，兰科特等人 (2017)，李等人 (2019a)，刘等人 (2020)，洛等人 (2017)，平托等人 (2017)；拉古等人 (2018)，银等人 (2016)，银等人 (2018)，斯波纳和萨瓦尼尼 (2020)，苏克巴塔尔等人 (2017)

Table 3 (continued)

表 3(续)

Challenge	Approach	Literature
Partial observability	Memory mechanism	Dibangoye and Buffet (2018), Foerster et al. (2018b), Foerster et al. (2019), Gupta et al. (2017), and Omid- shafiei et al. (2017)

挑战	方法	文献
部分可观测性	记忆机制	迪班戈耶和布菲特 (2018)，福尔斯特等人 (2018b)，福尔斯特等人 (2019)，古普塔等人 (2017)，奥米德-沙菲伊等人 (2017)

communication between agents (Sect. 5.2) or building models (Sect. 5.3). However, we discuss these topics separately in the respective sections.

代理间的通信 (第 5.2 节) 或构建模型 (第 5.3 节)。然而，我们在相应的章节中分别讨论这些主题。

Experience replay mechanism Recent successes with reinforcement learning methods such as deep Q-networks (Mnih et al. 2015) rest upon an experience replay mechanism. However, it is not straightforward to employ experience replays to the multi-agent setting because past experience becomes obsolete with the adaption of agent policies over time. To encounter this, Foerster et al. (2017) proposed two approaches. First, they decay outdated transition samples from the replay memory to stabilize targets and then use importance sampling to incorporate off-policy samples. Since the agents' policies are known during the training, off-policy updates can be corrected with importance-weighted policy likelihoods. Second, the state space of each agent is enhanced with estimates of the other agents' policies, so-called fingerprints [5] , to prevent non-stationarity. The value functions can then be conditioned on a fingerprint, which clears the age of data sampled from the replay memory. Another extension for experience replays was proposed by Palmer et al. (2018) who applied leniency to every stored transition sample. Leniency associates each sample of the experience memory with a temperature value, which gradually decays by the number of state-action pair visits. Further utilization of the experience replay mechanism to cope with non-stationarity can be found in Tang et al. (2018) and Zheng et al. (2018a). Nevertheless, if the contemporary dynamics of the learners are neglected, algorithms can utilize short-term buffers as applied in Baker et al. (2020) and Leibo et al. (2017).

经验回放机制近期深度 Q 网络 (Mnih 等人，2015 年) 等强化学习方法的成功基于经验回放机制。然而，在多代理环境中应用经验回放并不简单，因为随着代理策略随时间的适应，过去的经验会变得过时。为了应对这个问题，Foerster 等人 (2017 年) 提出了两种方法。首先，他们从回放记忆中衰减过时的转换样本以稳定目标，然后使用重要性采样来合并异策略样本。由于在训练期间已知代理的策略，因此可以用重要性加权的策略似然性来纠正异策略更新。其次，每个代理的状态空间通过其他代理策略的估计值进行了增强，所谓的指纹 [5] ，以防止非平稳性。然后可以将价值函数对指纹进行条件化，这消除了从回放记忆中抽取数据的时效性。Palmer 等人 (2018 年) 提出了经验回放的另一种扩展，他们为存储的每个转换样本应用了宽容度。宽容度将经验记忆中的每个样本与一个温度值相关联，该值随着状态-动作对访问次数的增加而逐渐衰减。在 Tang 等人 (2018 年) 和 Zheng 等人 (2018a) 的研究中，可以找到更多关于经验回放机制应对非平稳性的应用。尽管如此，如果忽略了学习者的当代动态，算法可以利用短期缓冲区，如 Baker 等人 (2020 年) 和 Leibo 等人 (2017 年) 所应用的。

Centralized Training Scheme As already discussed in Sect. 3.2, the CTDE paradigm can be leveraged to share mutual information between learners to ease training. The availability of information during the training can loosen the non-stationarity of the environment since agents are augmented with information about others. One approach is to enhance actor-critic methods with centralized critics over which mutual information is shared between agents during the training (Bono et al. 2019; Iqbal and Sha 2019; Wei et al. 2018). Lowe et al. (2017) embedded each agent with one centralized critic that is augmented with all agents' observations and actions. Based on this additional information, agents face a stationary environment during the training while acting decentralized on local observations at test time. Next to the equipment of one critic per agent, all agents can share one global centralized critic. Foerster et al. (2018b) applied one centralized critic conditioned on the joint action and observations of all agents. The critic computes an agent's individual advantage through estimating the value of the joint action based on a counterfactual baseline, which marginalizes out single agents' influence. Another approach to the CTDE scheme can be seen in value-based methods. Rashid et al. (2018) learned a joint action-value function conditioned on the joint observation-action history. The joint action-value function is then divided into agent individual value functions based on monotonic non-linear composition. Foerster et al. (2016) used action-value functions that share information through a communication channel during the training but then discarded it at test time. Similarly, Jorge et al. (2016) employed communication during training to promote information exchange for optimizing action-value functions.

集中式训练方案如已在第 3.2 节中讨论的那样，CTDE 范式可以用来在学习者之间共享相互信息以简化训练。训练过程中信息的可用性可以减轻环境的非平稳性，因为代理获得了关于其他代理的信息。一种方法是增强演员-评论家方法，在训练过程中，代理之间共享关于集中评论家的相互信息 (Bono 等人 2019 年；Iqbal 和 Sha 2019 年；Wei 等人 2018 年)。Lowe 等人 (2017 年) 将每个代理嵌入一个集中评论家，该评论家增加了所有代理的观察和动作。基于这些额外信息，代理在训练过程中面临一个平稳环境，而在测试时基于本地观察进行去中心化行动。除了为每个代理配备一个评论家外，所有代理还可以共享

一个全局集中评论家。Foerster 等人 (2018b) 应用了一个基于所有代理的联合动作和观察的集中评论家。评论家通过估计基于反事实基线的联合动作的价值来计算代理的个体优势，该基线消除了单个代理的影响。另一种 CTDE 方案的途径可以在基于价值的方法中看到。Rashid 等人 (2018 年) 学习了一个基于联合观察-动作历史的联合动作价值函数。然后，根据单调非线性组合，将联合动作价值函数划分为代理个体价值函数。Foerster 等人 (2016 年) 使用动作价值函数，在训练过程中通过通信通道共享信息，但在测试时丢弃这些信息。同样，Jorge 等人 (2016 年) 在训练过程中使用通信来促进信息交换，以优化动作价值函数。

Meta-Learning Sometimes, it can be useful to learn how to adapt to the behavioral changes of others. This learning-to-learn approach is known as meta-learning (Finn and Levine 2018; Schmidhuber et al. 1996). Recent works in the single-agent domain have shown promising results (Duan et al. 2016; Wang et al. 2016a). Al-Shedivat et al. (2018) transferred this approach to the multi-agent domain and developed a meta-learning based method to tackle the consecutive adaptation of agents in non-stationary environments. Regarding non-stationarity as a sequence of stationary tasks, agents learn to exploit dependencies between successive tasks and generalize over co-adapting agents at test time. They evaluated the resulting behaviors in a competitive multi-agent setting where agents fight in a simulated physics environment. Meta-learning can also be utilized to construct agent models (Rabinowitz et al. 2018). By learning how to model other agents and make inferences on them, agents learn to predict the other agent's future action sequences. They embedded this principle into how one agent learns to capture the behavioral patterns of other agents efficiently.

元学习有时，学会如何适应他人的行为变化可能是有用的。这种学习-学习的方法被称为元学习 (Finn 和 Levine 2018; Schmidhuber 等人 1996)。最近在单一智能体领域的研究已经显示出有希望的结果 (Duan 等人 2016; Wang 等人 2016a)。Al-Shedivat 等人 (2018) 将这种方法转移到多智能体领域，并开发了一种基于元学习的方法来解决非平稳环境中智能体的连续适应问题。将非平稳性视为一系列平稳任务，智能体学会利用连续任务之间的依赖关系，并在测试时泛化到共同适应的智能体。他们在竞争性多智能体环境中评估了由此产生的行为，其中智能体在模拟物理环境中战斗。元学习也可以用来构建智能体模型 (Rabinowitz 等人 2018)。通过学习如何建模其他智能体并对它们进行推理，智能体学会了预测其他智能体的未来动作序列。他们将这一原则嵌入到一个智能体如何高效地学习捕捉其他智能体的行为模式中。

5.2 Learning communication

5.2 学习交流

Agents capable of developing communication and language corpora pose one of the vital challenges in machine intelligence (Kirby 2002). Intelligent agents must not only decide on what to communicate but also when and with whom. It is indispensable that the developed language is grounded on a common consensus such that all agents understand the spoken language, including its semantics. The research efforts in learning to communicate have intensified because many pathologies can be overcome by incorporating communication skills into agents, including non-stationarity, coherent coordination among agents, and partial observability. For instance, when an agent knows the actions taken by others, the learning problem becomes stationary again from a single agent's perspective in a fully observable environment. Even partial observability can be loosened by messaging local observations to other participants through communication, which helps compensate for limited knowledge (Goldman and Zilberstein 2004).

能够开发通信和语言语料库的智能体在机器智能领域提出了一个至关重要的挑战 (Kirby 2002)。智能体不仅要决定传达什么内容，还要决定何时以及与谁进行交流。开发的语言必须基于共识，以确保所有智能体都能理解所使用的语言，包括其语义。由于将通信技能融入智能体可以克服许多病理问题，比如非定常性、智能体之间的协调一致性和部分可观测性，因此学习通信的研究工作得到了加强。例如，当一个智能体知道其他智能体的行动时，从单个智能体的角度看，在完全可观测的环境中学习问题又变得定常了。即使是部分可观测性也可以通过将本地观察通过通信传递给其他参与者来放宽，这有助于补偿知识的局限性 (Goldman 和 Zilberstein 2004)。

The common framework to investigate communication is the dec-POMDP (Oliehoek and Amato 2016) which is a fully cooperative setting where agents perceive partial observations of the environment and try to improve upon an equally-shared reward. In such distributed systems, agents must not only learn how to cooperate but also how to communicate in order to optimize the mutual objective. Early MARL

5 Fingerprints draw their inspiration from Tesauro (2004) who eluded non-stationarity by conditioning each agent's policy on estimates of other agents' policies.

5 指纹灵感来源于 Tesauro(2004)，他通过在每个智能体的策略上对其他智能体策略的估计进行条件化，来避免非平稳性。

works investigated communication rooted in tabular worlds with limited observability (Kasai et al. 2008). Since the spring of deep learning methods, the research of learning communication has witnessed great attention because advanced computational methods provide new opportunities to study highly complex data.

研究通信的通用框架是 dec-POMDP(Oliehoek 和 Amato 2016)，这是一个完全合作的设置，其中智能体感知环境的部分观察，并尝试提高共享的奖励。在这样的分布式系统中，智能体不仅要学习如何合作，还要学习如何通信，以优化共同目标。早期的多智能体强化学习 (MARL) 作品研究了在具有有限可观测性的表格世界中的通信 (Kasai 等人，2008)。自从深度学习方法兴起以来，学习通信的研究受到了极大的关注，因为先进的计算方法提供了研究高度复杂数据的新机会。

In the following, we categorize the surveyed literature according to the message addressing. First, we describe the broadcasting scenario where sent messages are received by all agents. Second, we look into works that use targeted messages to decide on the recipients by using an attention mechanism. Third and last, we review communication in networked settings where agents communicate only with their local neighborhood instead of the whole population. Figure 3 shows a schematic illustration of this categorization. Another taxonomy may be based on the discrete or continuous nature of messages and the frequency of passed messages.

在以下内容中，我们根据消息寻址方式对所调查的文献进行分类。首先，我们描述了广播场景，其中发送的消息被所有代理接收。其次，我们研究了一些使用有针对性的消息并通过注意力机制决定接收者的工作。第三，也是最后，我们回顾了在网络环境中代理仅与它们的局部邻居而不是整个群体进行通信的情况。图 3 展示了这种分类的示意图。另一种分类法可能基于消息的离散或连续性质以及传递消息的频率。

Broadcasting Messages are addressed to all participants of the communication channel. Foerster et al. (2016) studied how agents learn discrete communication protocols in dec-POMDPs in order to accomplish a fully-cooperative task. Being in a CTDE setting, the communication is not restricted during the training but bandwidth-limited at test time. To discover meaningful communication protocols, they proposed two methods. The first, reinforced inter-agent learning (RIAL), is based on deep recurrent Q-networks combined with independent Q-learning where each agent learns an action-value function conditioned on the observation history as well as messages from other agents. Additionally, they applied parameter sharing so that all agents share and update common features from only one Q-network. The second method, differentiable inter-agent learning (DIAL), combines the centralized learning paradigm with deep Q-networks. Messages are delivered over discrete connections, which are based on a relaxation to become differentiable. In contrast, Sukh-baatar et al. (2016) proposed CommNet as an architecture that allows the learning of communication between agents purely based on continuous protocols. They showed that each agent learns the joint-action and a sparse communication protocol that encodes meaningful information. The authors emphasized that the decreased observability of vicious states encourages the importance of communication between agents. To foster scalable communication protocols that also facilitate heterogeneous agents, Peng et al. (2017) introduced the bidirectionally-coordinated network (BiCNet) where agents learn in a vectorized actor-critic framework to communicate. Through communication, they were able to coordinate heterogeneous agents in a combat game of StarCraft.

广播消息是针对通信通道中的所有参与者进行的。Foerster 等人 (2016 年) 研究了在部分可观察马尔可夫决策过程 (dec-POMDPs) 中，智能体如何学习离散的通信协议以完成完全合作的任务。在 CTDE 设置中，训练过程中的通信不受限制，但在测试时受到带宽限制。为了发现有意义的通信协议，他们提出了两种方法。第一种方法，强化智能体间学习 (RIAL)，基于深度循环 Q 网络结合独立 Q 学习，其中每个智能体学习一个基于观察历史以及其他智能体消息的动作价值函数。此外，他们应用了参数共享，以便所有智能体共享并更新来自单一 Q 网络的共同特征。第二种方法，可微分智能体间学习 (DIAL)，将集中式学习范式与深度 Q 网络相结合。消息通过基于放松以变得可微分的离散连接传递。相比之下，Sukhbaatar 等人 (2016 年) 提出了 CommNet 架构，该架构允许智能体纯粹基于连续协议学习通信。他们展示了每个智能体学习联合动作和一个稀疏的通信协议，该协议编码了有意义的信息。作者强调，恶劣状态的可观测性降低增强了智能体之间通信的重要性。为了促进可扩展的通信协议，同时也便于异构智能体使用，Peng 等人 (2017 年) 引入了双向协调网络 (BiCNet)，其中智能体在向量化演员-评论家框架中学习通信。通过通信，他们能够在《星际争霸》的战斗游戏中协调异构智能体。

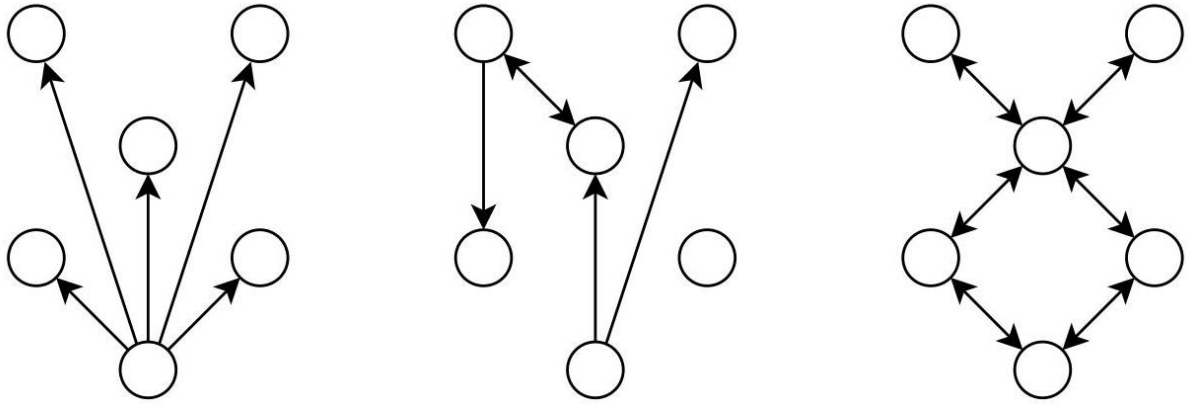

Fig. 3 Schematic illustration of communication types. Unilateral arrows represent unidirectional messages, while bilateral arrows symbolize bidirectional message passing. (Left) In broadcasting, messages are sent to all participants of the communication channel. For better visualization, the broadcasting of only one agent is illustrated but each agent can broadcast messages to all other agents. (Middle) Agents can target the communication through an attention mechanism that determines when, what and with whom to communicate. (Right) Networked communication describes the local connection to neighborhood agents

图 3 通信类型的示意图。单向箭头表示单向消息，而双向箭头象征双向消息传递。(左) 在广播中，消息被发送到通信通道的所有参与者。为了更好地可视化，这里仅说明了单个代理的广播，但每个代理都可以向所有其他代理广播消息。(中) 代理可以通过注意力机制有针对性地进行通信，该机制决定了何时、什么内容以及与谁通信。(右) 网络化通信描述了与邻近代理的局部连接。

Targeted communication When agents are endowed with targeted communication protocols, they utilize an attention mechanism to determine when, what and with whom to communicate. Jiang and Lu (2018) introduced ATOC as an attentional communication model that enables agents to send messages dynamically and selectively so that communication takes place among a group of agents only when required. They argued that attention is essential for large-scale settings because agents learn to decide which information is most useful for decision-making. Selective communication is the reason why ATOC outperforms CommNet and BiCNet on the conducted navigation tasks. A similar conclusion was drawn by Hoshen (2017) who introduced the vertex attention interaction network (VAIN) as an extension to the CommNet. The baseline approach is extended with an attention mechanism that increases performance due to the focus on only relevant agents. The work by Das et al. (2019) introduced targeted multi-agent communication (TarMAC) that uses attention to decide with whom and what to communicate by actively addressing other agents for message passing. Jain et al. (2019) proposed TBONE for visual navigation in cooperative tasks. In contrast to former works, which are limited to the fully-cooperative setting, Singh et al. (2019) considered mixed settings where each agent owns an individual reward function. They proposed the individualized controlled continuous communication model (IC3Net), where agents learn when to exchange information using a gating mechanism that blocks incoming communication requests if necessary.

有针对性的通信当代理被赋予有针对性的通信协议时，它们利用注意力机制来确定何时、何事以及与谁通信。Jiang 和 Lu(2018) 介绍了 ATOC 作为一种注意力通信模型，它使得代理能够动态和有选择地发送消息，从而仅在需要时在代理群组之间进行通信。他们认为，在大型场景中，注意力是必不可少的，因为代理学会了决定哪些信息对于决策最有用。选择性通信是 ATOC 在执行的导航任务中优于 CommNet 和 BiCNet 的原因。Hoshen(2017) 得出了类似的结论，他介绍了顶点注意力交互网络 (VAIN) 作为对 CommNet 的扩展。基线方法通过增加注意力机制来提高性能，该机制仅关注相关代理。Das 等人 (2019) 介绍了有针对性的多代理通信 (TarMAC)，它使用注意力来决定与谁以及通信什么内容，通过主动向其他代理发送消息。Jain 等人 (2019) 为协作任务中的视觉导航提出了 TBONE。与之前仅限于完全协作设置的工作不同，Singh 等人 (2019) 考虑了每个代理拥有个体奖励函数的混合设置。他们提出了个体化控制连续通信模型 (IC3Net)，在该模型中，代理使用门控机制学习何时交换信息，如果需要，该机制会阻止传入的通信请求。

Networked communication Another form of communication is a networked communication protocol where agents can exchange information with their neighborhood (Nedic and Ozdaglar 2009; Zhang et al. 2018). Agents act decentralized based on local observations and received messages from network neighbors. Zhang et al. (2018) used an actor-critic framework where agents share their critic information

with their network neighbors to promote global optimality. Chu et al. (2020) introduced the neural communication protocol (NeurComm) to enhance communication efficiency by reducing queue length and intersection delay. Further, they showed that a spatial discount factor could stabilize training when only the local vicinity is regarded to perform policy updates. For theoretical contributions, one may consider the works of Qu et al. (2020), Zhang et al. (2018) and Zhang et al. (2019) whereas the paper of Chu et al. (2020) provides an application perspective in the domain of traffic light control.

网络化通信另一种通信形式是网络化通信协议，在该协议中，代理可以与它们的邻居 (Nedic 和 Ozdaglar 2009；Zhang 等人 2018) 交换信息。代理基于本地观察和网络邻居收到的消息进行去中心化行动。Zhang 等人 (2018) 使用了一个演员-评论家框架，其中代理与它们的网络邻居共享评论家信息以促进全局最优性。Chu 等人 (2020) 引入了神经通信协议 (NeurComm)，通过减少队列长度和交叉延迟来提高通信效率。此外，他们展示了当仅考虑局部区域进行策略更新时，空间折扣因子可以稳定训练。在理论贡献方面，可以考虑 Qu 等人 (2020)、Zhang 等人 (2018) 和 Zhang 等人 (2019) 的工作，而 Chu 等人 (2020) 的论文在交通灯控制领域提供了应用视角。

Extensions Further methods approach the improvement of coordination skills by applying intrinsic motivation (Jaques et al. 2018, 2019), by making the communication protocol more robust or scalable (Kim et al. 2019; Singh et al. 2019), and maximizing the utility of the communication through efficient encoding (Celikyilmaz et al. 2018; Li et al. 2019b; Wang et al. 2020c).

扩展进一步的方法通过应用内在动机 (Jaques 等人 2018, 2019)，通过使通信协议更加健壮或可扩展 (Kim 等人 2019；Singh 等人 2019)，以及通过有效的编码最大化通信的效用 (Celikyilmaz 等人 2018；Li 等人 2019b；Wang 等人 2020c) 来提高协调技能。

The above-reviewed papers focus on new methodologies about communication protocols. Besides that, a bulk of literature considers the analysis of emergent language and the occurrence of agent behavior, which we discuss in Sect. 4.2.

上述回顾的论文专注于关于通信协议的新方法。除此之外，大量文献考虑了新兴语言的分析和代理行为的出现，我们将在第 4.2 节中讨论这些内容。

5.3 Coordination

5.3 协调

Successful coordination in multi-agent systems requires agents to agree on a consensus (Wei Ren et al. 2005). In particular, accomplishing a joint goal in cooperative settings demands a coherent action selection such that the joint action optimizes the mutual task performance. Cooperation among agents is complicated when stochasticity is present in system transitions and rewards or when agents observe only partial information of the environment's state. Mis-coordination may arise in the form of action shadowing when exploratory behavior influences the other agents' search space during learning and, as a result, sub-optimal solutions are found.

多智能体系统中的成功协调需要智能体就共识达成一致 (Wei Ren et al. 2005)。特别是，在合作环境中实现共同目标需要选择一致的行动，以使联合行动优化相互的任务表现。当系统转换和奖励存在随机性，或者智能体仅观察到环境状态的局部信息时，智能体之间的合作变得复杂。在学习过程中，探索性行为可能会影响其他智能体的搜索空间，导致出现动作阴影形式的不协调，从而找到次优解。

Therefore, the agreement upon a mutual consensus necessitates the sharing and collection of information about other agents to derive optimal decisions. Finding such a consensus in the decision-making may happen explicitly through communication or implicitly by constructing models of other agents. The former requires skills to communicate with others so that agents can express their purpose and align their coordination. For the latter, agents need the ability to observe other agents' behavior and reason about their strategies to build a model. If the prediction model is accurate, an agent can learn the other agents' behavioral patterns and direct actions towards a consensus, leading to coordinated behavior. Besides explicit communication and constructing agent models, the CTDE scheme can be leveraged to build different levels of abstraction, which are applied to learn high-level coordination while independent skills are trained at low-level.

因此，就共同共识达成一致需要共享和收集有关其他智能体的信息以得出最优决策。在决策中找到这样的共识可能通过通信明确发生，或者通过构建其他智能体的模型隐式发生。前者需要与其他人通信的技能，以便智能体能够表达其目的并调整其协调。对于后者，智能体需要能够观察其他智能体的行为并推理它们的策略来构建模型。如果预测模型准确，智能体可以学习其他智能体的行为模式并将行动指向共识，从而导致协调行为。除了显式通信和构建智能体模型之外，CTDE 方案可以利用来构建不同级别的抽象，这些抽象应用于学习高级协调，同时在低级别训练独立技能。

In the remainder of this section, we focus on methods that solve coordination issues without establishing communication protocols between agents. Although communication may ease coordination, we discuss this topic separately in Sect. 5.2.

在本节的剩余部分，我们关注解决协调问题的方法，这些方法不需要在智能体之间建立通信协议。尽管通信可能会简化协调，但我们将在第 5.2 节单独讨论这个话题。

Independent learners The naïve approach to handle multi-agent problems is to regard each agent individually such that other agents are perceived as part of the environment and, thus, are neglected during learning. Opposed to joint action learners, where agents experience the selected actions of others a-posteriori, independently learning agents face the main difficulty of coherently choosing actions such that the joint action becomes optimal concerning the mutual goal (Matignon et al. 2012b). During the learning of good policies, agents influence each other's search space, which can lead to action shadowing. The notion of coordination among several autonomously and independently acting agents enjoys a long record, and a bulk of research was conducted in settings with non-communicative agents (Fulda and Ventura 2007; Matignon et al. 2012b). Early works investigated the convergence of independent learners and showed that the convergence to solutions is feasible under certain conditions in deterministic games but fails in stochastic environments (Claus and Boutilier 1998; Lauer and Riedmiller 2000). Stochasticity, relative over-generalization, and other pathologies such as non-stationarity and the alter-exploration problem led to new branches of research including hysteretic learning (Matignon et al. 2007) and leniency (Potter and De Jong 1994). Hysteretic Q-learning was introduced to encounter the over-estimation of the value function evoked by stochasticity. Two learning rates are used to increase and decrease the value function updates while relying on an optimistic form of learning. A modern approach to hysteretic learning can be seen in Palmer et al. (2018) and Omidshafiei et al. (2017). An alternative method to adjust the degree of applied optimism during learning is leniency (Panait et al. 2006; Wei and Luke 2016). Leniency associates selected actions with decaying temperature values that govern the amount of applied leniency. Agents are optimistic during the early phase when exploration is still high but become less lenient for frequently visited state-action pairs over the training so that value estimations become more accurate towards the end of learning.

独立学习者处理多代理问题的天真方法是单独考虑每个代理，使得其他代理被视为环境的一部分，并在学习过程中被忽视。与联合行动学习者不同，独立学习者在选择行动以达到共同目标时，面临的主要困难是一致性地选择行动，使得联合行动变得最优 (Matignon 等人，2012b)。在良好策略的学习过程中，代理会影响彼此的搜索空间，这可能导致行动模仿。多个自主且独立行动的代理之间的协调概念有着悠久的历史，大量研究在非通信代理的设置中进行 (Fulda 和 Ventura 2007；Matignon 等人，2012b)。早期的工作研究了独立学习者的收敛性，并表明在确定性游戏中，在特定条件下收敛到解决方案是可行的，但在随机环境中则失败 (Claus 和 Boutilier 1998；Lauer 和 Riedmiller 2000)。随机性、相对过度泛化以及其他病态现象，如非平稳性和交替探索问题，导致了包括滞后学习 (Matignon 等人，2007) 和宽容度 (Potter 和 De Jong 1994) 在内的新研究分支的产生。滞后 Q 学习被引入以应对随机性引起的价值函数过度估计问题。使用两个学习率来增加和减少价值函数的更新，同时依赖一种乐观的学习形式。滞后学习的现代方法可以在 Palmer 等人 (2018) 和 Omidshafiei 等人 (2017) 的作品中看到。在学习过程中调整应用乐观程度的一种替代方法是宽容度 (Panait 等人 2006；Wei 和 Luke 2016)。宽容度将选择的行为与控制应用宽容度的衰减温度值相关联。代理在早期阶段仍然高度探索时持乐观态度，但随着训练过程中频繁访问的状态-动作对的增加，代理变得不那么宽容，从而使价值估计在学习的末期变得更加准确。

Further works expanded independent learners with enhanced techniques to cope with the MARL pathologies mentioned above. Extensions to the deep Q-network can be seen in additional mechanisms used for the experience replay (Palmer et al. 2019), the utilization of specialized estimators (Zheng et al. 2018a) and the use of implicit quantile networks (Lyu and Amato 2020). Further literature investigated independent learners as benchmark reference but reported limited success in cooperative tasks of various domains when no other techniques are applied to alleviate the issue of independent learners (Foerster et al. 2018b; Sunehag et al. 2018).

进一步的工作扩展了独立学习者，通过增强的技术来应对上述提到的多智能体强化学习 (MARL) 的病理现象。深度 Q 网络 (DQN) 的扩展可以在经验回放中使用的额外机制 (Palmer 等，2019 年) 中看到，还有专用估计器的应用 (Zheng 等，2018a)，以及隐式分位数网络的运用 (Lyu 和 Amato，2020 年)。更多的文献将独立学习者作为基准参考，但在各种领域的合作任务中，如果没有其他技术缓解独立学习者的问题时，报告的成功有限 (Foerster 等，2018b；Sunehag 等，2018 年)。

Constructing models An implicit way to achieve coordination among agents is to capture the behavior of others by constructing models. Models are functions that take past interaction data as input and output predictions about the agents of interest. This can be very important to render the learning process robust against the decision-making of other agents in the environment (Hu and Wellman 1998). The constructed models and the predicted behavior vary widely depending on the approaches and the assumptions being

made (Albrecht and Stone 2018).

构建模型实现代理之间协调的一种隐式方式是通过构建模型来捕捉其他代理的行为。模型是接受过去的交互数据作为输入并输出关于感兴趣代理的预测的函数。这对于使学习过程能够抵御环境中其他代理的决策制定非常重要 (Hu 和 Wellman，1998 年)。构建的模型和预测的行为根据采用的方法和所做的假设而有很大的差异 (Albrecht 和 Stone，2018 年)。

One of the first works based on deep learning methods was conducted by He et al. (2016) in an adversarial setting. They proposed an architecture that utilizes two neural networks. One neural network captures the opponents' strategies, and the second network estimates the opponents' Q-values. These networks jointly learn models of opponents by encoding observations into a deep Q-network. Another work by Foerster et al. (2018a) introduced a learning method where the policy updates rely on the impact on other agents. The opponent's policy parameters can be inferred from the observed trajectory by using a maximum likelihood technique. The arising non-stationarity is tackled by accounting only recent data. An additional possibility is to address the information gain about other agents through Bayesian methods. Raileanu et al. (2018) employed a model where agents estimate the other agents' hidden states and embed these estimations into their own policy. Inferring other agents' hidden states from their behavior allows them to choose appropriate actions and promotes eventual coordination. Foerster et al. (2019) used all publicly available observations in the environment to calculate a public belief over agents' local information. Another work by Yang et al. (2018a) used Bayesian techniques to detect opponent strategies in competitive games. A particular challenge is to learn agent models in the presence of fast adapting agents, which amplifies the problem of non-stationarity. As a countermeasure, Everett and Roberts (2018) proposed the switching agent model (SAM), which learns a set of opponent models and a switching mechanism between models. By tracking and detecting the behavioral adaption of other agents, the switching mechanism learns to select the best response from the learned set of opponent models and, thus, showed superior performance over single model learners.

He 等人 (2016) 在对抗性环境中基于深度学习方法进行了早期的工作。他们提出了一种架构，该架构利用两个神经网络。一个神经网络捕捉对手的策略，另一个神经网络估计对手的 Q 值。这些网络通过将观察编码到深度 Q 网络中，共同学习对手的模型。Foerster 等人 (2018a) 介绍了另一种学习方法，其中策略更新依赖于对其他代理的影响。可以通过使用最大似然技术从观察到的轨迹中推断出对手的策略参数。通过仅考虑最近的数据来应对出现的非平稳性。另一种可能性是通过贝叶斯方法来解决关于其他代理的信息增益问题。Raileanu 等人 (2018) 使用了一种模型，其中代理估计其他代理的隐藏状态并将这些估计嵌入到自己的策略中。从其他代理的行为推断其隐藏状态使他们能够选择适当的行为并促进最终的协调。Foerster 等人 (2019) 使用环境中所有公开可用的观察结果来计算关于代理本地信息的公共信念。杨等人 (2018a) 使用贝叶斯技术在竞争游戏中检测对手策略。一个特殊的挑战是在快速适应的代理存在的情况下学习代理模型，这放大了非平稳性问题。作为一种对策，Everett 和 Roberts(2018) 提出了切换代理模型 (SAM)，该模型学习一组对手模型以及模型之间的切换机制。通过跟踪和检测其他代理的行为适应，切换机制学习从学到的对手模型集中选择最佳响应，因此，在单一模型学习者之上展示了优越的性能。

Further works on constructing models can be found in cooperative tasks (Barde et al. 2019; Tacchetti et al. 2019; Zheng et al. 2018b) with imitation learning (Grover et al. 2018; Le et al. 2017), in social dilemmas (Jaques et al. 2019; Letcher et al. 2019), and by predicting behaviors from observations (Hong et al. 2017; Hoshen 2017). For a comprehensive survey on constructing models in multi-agent systems, one may consider the work of Albrecht and Stone (2018).

在合作任务 (Barde 等人 2019；Tacchetti 等人 2019；Zheng 等人 2018b) 中使用模仿学习 (Grover 等人 2018；Le 等人 2017)，在社会困境 (Jaques 等人 2019；Letcher 等人 2019) 中，以及通过观察预测行为 (Hong 等人 2017；Hoshen 2017) 的构建模型的工作中，可以找到更多关于构建模型的研究。对于多代理系统中构建模型的全面调查，可以考虑 Albrecht 和 Stone(2018) 的工作。

Besides resolving the coordination problem, building models of other agents can cope with the non-stationarity in the environment. As soon as one agent has knowledge about others' behavior, previously unexplainable transition dynamics can be attributed to the responsible agents, and the environment becomes stationary again from the viewpoint of an individual agent.

除了解决协调问题之外，构建其他代理的模型可以应对环境中的非平稳性。一旦一个代理了解其他代理的行为，之前无法解释的转换动态可以归因于负责的代理，从单个代理的角度来看，环境又变得平稳。

Hierarchical methods Learning to coordinate can be challenging if multiple decision-makers are involved due to the increasing complexity (Bernstein et al. 2002). An approach to deal with the coordination problem is by abstracting low-level coordination to higher levels. The idea originated in the single-agent domain where hierarchies for temporal abstraction are employed to ease long-term reward assignments (Dayan and Hinton 1993; Sutton et al. 1999). Lower levels entail only partial information of the higher levels so that the learning task becomes simpler the lower the level of abstraction. First attempts for

hierarchical multi-agent RL can be found in the tabular case (Ghavamzadeh et al. 2006; Makar et al. 2001). A deep approach was proposed by Kumar et al. (2017), where a higher-level controller guides the information exchange between decentralized agents. Grounded on the high-level controller, the agents communicate with only one other agent at each time step, which allows the exploration of distributed policies. Another work by Han et al. (2019) is built upon the options framework (Sutton et al. 1999) where they embedded a dynamic termination criterion for Q-learning. By adding a termination criterion, agents could flexibly quit the option execution and react to the behavioral changes of other agents. Related to the idea of feudal networks (Dayan and Hinton 1993), Ahilan and Dayan (2019) applied a two-level abstraction of agents to a cooperative multi-agent setting where, in contrast to other methods, the hierarchy relied on rewards instead of state goals. They showed that this approach could be well suited for decentralized control problems. Jaderberg et al. (2019) used hierarchical representations that allowed agents to reason at different time scales. The authors demonstrated that agents are capable of solving mixed cooperative and competitive tasks in simulated physics environments. Another work by Lee et al. (2020) proposed a hierarchical method to coordinate two agents on robotic manipulation and locomotion tasks to accomplish collaboration such as object pick and placement. They learned primitive skills on the low-level, which are guided by a higher-level policy. Further works cover hierarchical methods in cooperation tasks (Cai et al. 2013; Ma and Wu 2020; Tang et al. 2018) or social dilemmas (Vezhnevets et al. 2019). An open challenge for hierarchical methods is the autonomous creation and discovery of abstract goals from data (Schaul et al. 2015; Vezhnevets et al. 2017).

如果涉及多个决策者，由于复杂性增加，学习协调可能会很具挑战性 (Bernstein 等人，2002 年)。处理协调问题的一种方法是通过对低级别协调进行抽象，将其提升到更高层次。这个想法起源于单智能体领域，其中时间抽象的层次结构被用于简化长期奖励分配 (Dayan 和 Hinton，1993 年；Sutton 等人，1999 年)。较低层次只包含较高层次的局部信息，因此抽象层次越低，学习任务就越简单。在表格案例中可以找到分层多智能体强化学习的首次尝试 (Ghavamzadeh 等人，2006 年；Makar 等人，2001 年)。Kumar 等人 (2017 年) 提出了一种深度方法，其中高层控制器指导去中心化智能体之间的信息交换。基于高层控制器，智能体在每个时间步骤只与另一个智能体通信，这允许探索分布式策略。Han 等人 (2019 年) 的工作是基于选项框架 (Sutton 等人，1999 年)，他们为 Q 学习嵌入了一个动态终止标准。通过添加终止标准，智能体可以灵活地退出选项执行并对他智能体的行为变化作出反应。与封建网络的想法 (Dayan 和 Hinton，1993 年) 相关，Ahilan 和 Dayan(2019 年) 在合作多智能体设置中应用了智能体的两层抽象，与其他方法不同，这种层次结构依赖于奖励而不是状态目标。他们展示了这种方法非常适合去中心化控制问题。Jaderberg 等人 (2019 年) 使用了分层表示，使智能体能够在不同的时间尺度上进行推理。作者证明智能体能够在模拟物理环境中解决混合合作与竞争任务。Lee 等人 (2020 年) 的另一项工作提出了一种分层方法，用于协调两个智能体在机器人操作和移动任务上进行协作，例如物体的抓取和放置。他们在低层次上学习基本技能，这些技能受到高层策略的指导。其他研究涵盖了合作任务 (Cai 等人，2013 年；Ma 和 Wu，2020 年；Tang 等人，2018 年) 或社会困境 (Vezhnevets 等人，2019 年) 中的分层方法。分层方法的一个开放挑战是从数据中自主创建和发现抽象目标 (Schaul 等人，2015 年；Vezhnevets 等人，2017 年)。

5.4 Credit assignment problem

5.4 信用分配问题

In the fully-cooperative setting, agents are encouraged to maximize an equally-shared reward signal. Even in a fully-observable state space, it is difficult to determine which agents and actions contributed to the eventual reward outcome when agents do not have access to the joint action. Claus and Boutilier (1998) showed that independent learners could not differentiate between the teammate's exploration and the stochasticity in the environment even in a simple bi-matrix game. This can render the learning problem difficult because agents should be ideally provided with feedback corresponding to the task performance to enable sufficient learning. Associating rewards to agents is known as the credit assignment problem (Weiß 1995; Wolpert and Tumer 1999). This problem is intensified by the sequential nature of reinforcement learning where agents must understand not only the impact of single actions but also the entire action sequences that eventually lead to the reward outcome (Sen and Weiss 1999). An additional challenge arises when agents have only access to local observations of the environment, which we discuss in Sect. 5.6. In the remainder of this section, we consider three actively investigated approaches that deal with how to determine the contribution of agents jointly-shared reward settings.

在完全合作的设置中，鼓励智能体最大化一个平等共享的奖励信号。即使在完全可观察的状态空间中，当智能体无法访问到联合行动时，确定哪些智能体和行动对最终奖励结果有所贡献也是困难的。Claus 和

Boutilier(1998) 表明，即使在简单的双矩阵游戏中，独立的学习者也无法区分队友的探索和环境中的随机性。这可能会使学习问题变得困难，因为理想情况下，智能体应该得到与任务表现相对应的反馈，以实现足够的学习。将奖励与智能体关联起来被称为信用分配问题 (Weiß 1995；Wolpert 和 Tumer 1999)。由于强化学习的顺序性质，这个问题变得更加复杂，智能体必须理解不仅单个行动的影响，还要理解最终导致奖励结果的整体行动序列 (Sen 和 Weiss 1999)。当智能体只能访问环境的局部观察时，还会出现额外的挑战，我们将在第 5.6 节中讨论。在本节的其余部分，我们考虑三种积极研究的方法，这些方法涉及如何确定在智能体共同分享的奖励设置中智能体的贡献。

Decomposition Early works approached the credit assignment problem by applying filters (Chang et al. 2004) or modifying the reward function such as reward shaping (Ng et al. 1999). Recent approaches focus on exploiting dependencies between agents to decompose the reward among the agents with respect to their actual contribution towards the global reward (Kok and Vlassis 2006). The learning problem is simplified by dividing the task into smaller and, hence, easier sub-problems through decomposition. Sunehag et al. (2018) introduced the value decomposition network (VDN) which factorizes the joint action-value function into a linear combination of individual action-value functions. The VDN learns how to optimally assign an individual reward according to the agent's performance. The neural network helps to disambiguate the joint reward signal concerning the impact of the agent. Rashid et al. (2018) proposed QMIX as an improvement over VDN. QMIX learns a centralized action-value function that is decomposed into agent individual action-value functions through non-linear combinations. Under the assumption of monotonic relationships between the centralized Q-function and the individual Q-functions, decentralized policies can be extracted by individual argmax operations. As an advancement over both VDN and QMIX, Son et al. (2019) proposed QTRAN, which discards the assumption of linearity and monotonicity in the factorization and allows any non-linear combination of value functions. Further approaches about the factorization of value functions can be found in Castellini et al. (2019), Chen et al. (2018), Nguyen et al. (2017b), Wang et al. (2020a), Wang et al. (2020c) and Yang et al. (2018b).

分解早期工作通过应用滤波器 (Chang et al. 2004) 或修改奖励函数，如奖励塑形 (Ng et al. 1999)，来处理信用分配问题。最近的方法专注于利用代理之间的依赖关系来分解奖励，根据代理对全局奖励的实际贡献，在代理之间分配奖励 (Kok and Vlassis 2006)。通过分解将任务划分为更小、因此更易于处理子问题，简化了学习问题。Sunehag et al. (2018) 引入了价值分解网络 (VDN)，该网络将联合动作价值函数分解为个体动作价值函数的线性组合。VDN 学习如何根据代理的表现最优地分配个体奖励。神经网络有助于消除关于代理影响的联合奖励信号的模糊性。Rashid et al. (2018) 提出了 QMIX，作为对 VDN 的改进。QMIX 学习一个中心化的动作价值函数，该函数通过非线性组合分解为代理的个体动作价值函数。在中心化 Q 函数与个体 Q 函数之间存在单调关系的假设下，可以通过个体 argmax 操作提取去中心化策略。作为对 VDN 和 QMIX 的进一步改进，Son et al. (2019) 提出了 QTRAN，该算法放弃了分解中的线性和单调性假设，允许价值函数的任何非线性组合。关于价值函数分解的进一步方法可以在 Castellini et al. (2019)，Chen et al. (2018)，Nguyen et al. (2017b)，Wang et al. (2020a)，Wang et al. (2020c) 和 Yang et al. (2018b) 中找到。

Marginalization Next to the decomposition into simpler sub-problems, one can apply an extra function that marginalizes out the effect of agent individual actions. Nguyen et al. (2018) introduced a mean collective actor-critic framework which marginalizes out the actions of agents by using an approximation of the critic and reduces the variance of the gradient estimation. Similarly, Foerster et al. (2018b) marginalized out the individual actions of agents by applying a counterfactual baseline function. The counterfactual baseline function uses a centralized critic, which calculates the advantage of a single agent by comparing the estimated return of the current joint-action to the counterfactual baseline. The impact of a single agent's action is determined and can be attributed to the agent itself. Another work by Wu et al. (2018) used a marginalized action-value function as a baseline to reduce the variance of critic estimates. The marginalization approaches are closely related to the difference rewards proposed by Tumer and Wolpert (2004) who determine the impact of an agent's individual action compared to the average reward of all agents.

边缘化除了将问题分解为更简单的子问题之外，还可以应用一个额外的函数来边缘化出代理个体行动的影响。Nguyen 等人 (2018 年) 引入了一种平均集体演员-评价者框架，该框架通过使用评价者的近似来边缘化出代理的行动，并减少梯度估计的方差。同样，Foerster 等人 (2018b) 通过应用反事实基线函数边缘化出代理的个体行动。反事实基线函数使用集中式评价者，通过将当前联合行动的估计回报与反事实基线进行比较来计算单个代理的优势。单个代理行动的影响被确定，并且可以归因于代理本身。Wu 等人 (2018 年) 的另一项工作使用边缘化的行动价值函数作为基线来减少评价者估计的方差。边缘化方法与 Tumer 和 Wolpert(2004 年) 提出的差异奖励密切相关，他们确定了一个代理的个体行动相对于所有代理的平均奖励的影响。

Inverse reinforcement learning Credit assignment problems can be evoked by a bad design of the reinforcement learning problem. Misinterpretations of the agents can lead to failure because unintentional

strategies are explored, e.g. if the reward function does not capture all important aspects of the underlying task (Amodei et al. 2016). Therefore, an important step in the problem design is the reward function. However, designing a reward function can be challenging for complex problems (Hadfield-Menell et al. 2017) and becomes even more complicated for multi-agent systems since different agents may accomplish different goals. Another approach to address the credit assignment problem is by inverse reinforcement learning (Ng and Russell 2000) that describes how an agent learns a reward function that explains the demonstrated behavior of an expert without having access to the reward signal. The learned reward function can then be used to build strategies. The work of Lin et al. (2018) applied the principle of inverse reinforcement learning to the multi-agent setting. They showed that multiple agents could recover reward functions that are correlated with the ground truths. Related to inverse RL, imitation learning can be used to learn from expert knowledge. Yu et al. (2019) imitated expert behaviors to learn high-dimensional policies in both cooperative and competitive environments. They were able to recover the expert policies for each individual agent from the provided expert demonstrations. Further works on imitation learning consider the fully cooperative setting (Barrett et al. 2017; Le et al. 2017) and Markov Games with mixed settings (Song et al. 2018).

逆强化学习信用分配问题可能由强化学习问题的不良设计引起。对代理的错误解释可能导致失败，因为会探索无意中的策略，例如，如果奖励函数没有捕捉到 underlying task 的所有重要方面 (Amodei 等人 2016)。因此，问题设计的一个重要步骤是奖励函数。然而，对于复杂问题设计奖励函数可能是具有挑战性的 (Hadfield-Menell 等人 2017)，并且在多代理系统中变得更加复杂，因为不同的代理可能实现不同的目标。解决信用分配问题的另一种方法是逆强化学习 (Ng 和 Russell 2000)，它描述了代理如何学习一个奖励函数，该函数能够解释专家展示的行为，而不需要访问奖励信号。学到的奖励函数然后可以用来构建策略。Lin 等人 (2018) 将逆强化学习的原则应用于多代理场景。他们展示了多个代理能够恢复与 ground truths 相关的奖励函数。与逆 RL 相关，模仿学习可以用来从专家知识中学习。Yu 等人 (2019) 模仿专家行为，在合作和竞争环境中学习高维策略。他们能够从提供的专家演示中恢复每个单独代理的专家策略。进一步的模仿学习工作考虑了完全合作设置 (Barrett 等人 2017；Le 等人 2017) 和混合设置的 Markov Games(Song 等人 2018)。

5.5 Scalability

5.5 可扩展性

Training a large number of agents is inherently difficult. Every agent involved in the environment adds extra complexity to the learning problem such that the computational effort grows exponentially by the number of agents. Besides complexity concerns, sufficient scaling also demands agents to be robust towards the behavioral adaption of other agents. However, agents can leverage the benefit of distributed knowledge shared and reused between agents to accelerate the learning process. In the following, we review approaches that address the handling of many agents and discuss possible solutions. We broadly classify the surveyed works into those that apply some form of knowledge reuse, reduce the complexity of the learning problem, and develop robustness against the policy adaptions of other agents.

训练大量代理本质上是非常困难的。环境中涉及的每个代理都会给学习问题增加额外的复杂性，以至于计算工作随着代理数量的增加而指数级增长。除了复杂性方面的考虑，足够的扩展还要求代理对其他代理的行为适应具有鲁棒性。然而，代理可以利用分布式知识共享和重用之间的优势来加速学习过程。接下来，我们回顾了处理多个代理的方法，并讨论可能的解决方案。我们将调查的作品广泛分类为应用某种形式的知识重用、减少学习问题的复杂性以及开发对其他代理策略适应的鲁棒性。

Knowledge reuse The training of individual learning models does scale poorly with the increasing number of agents because the computational effort increases due to the combinatorial possibilities. Knowledge reuse strategies are employed to ease the learning process and scale RL to complex problems by reutilizing previous knowledge into new tasks. Knowledge reuse can be applied in many facets (Silva et al. 2018).

知识重用单个学习模型的训练随着代理数量的增加而扩展性差，因为计算工作由于组合可能性而增加。知识重用策略被应用来简化学习过程，并通过将先前知识重用于新任务来扩展强化学习到复杂问题。知识重用可以在许多方面应用 (Silva et al. 2018)。

First, agents can make use of a parameter sharing technique if they exhibit homogeneous structures, e.g. the weights in a neural network for sharing parts or the whole learning model with others. Sharing the parameters of a policy enables an efficient training process that can scale up to an arbitrary number of agents and, thus, can boost the learning process (Gupta et al. 2017). Parameter sharing has proven to be useful in various applications such as learning to communicate (Foerster et al. 2016; Jiang and Lu 2018; Peng et al. 2017; Sukhbaatar et al. 2016), modeling agents (Hernandez-Leal et al. 2019), and in

partially observable cooperative games (Sunehag et al. 2018). For a discussion on different parameter sharing strategies, one may consider the paper by Chu and Ye (2017).

首先，如果代理表现出同质结构，例如在共享部分或整个学习模型时神经网络的权重，它们可以利用参数共享技术。共享策略的参数能够实现一个有效的训练过程，可以扩展到任意数量的代理，从而可以提升学习过程 (Gupta et al. 2017)。参数共享在各种应用中已被证明是有用的，例如学习交流 (Foerster et al. 2016; Jiang and Lu 2018; Peng et al. 2017; Sukhbaatar et al. 2016)、建模代理 (Hernandez-Leal et al. 2019) 以及在部分可观察的合作游戏中 (Sunehag et al. 2018)。关于不同参数共享策略的讨论，可以考虑 Chu 和 Ye(2017) 的论文。

As the second approach, knowledge reuse can be applied in form of transfer learning (Da Silva et al. 2019; Da Silva and Costa 2019). Experience obtained in learning to perform one task may also improve the performance in a related but different task (Taylor and Stone 2009). Da Silva and Costa (2017) used a knowledge database from which an agent can extract previous solutions of related tasks and embed such information into the current task's training. Likewise, Da Silva et al. (2017) applied expert demonstrations where the agents take the role of students that ask a teacher for advice. They demonstrated that simultaneously learning agents could advise each other through knowledge transfer. Further works on transfer learning can be found in the cooperative multi-agent setting (Omid-shafiei et al. 2019) and in natural language applications (Luketina et al. 2019). In general multi-agent systems, the works of (Boutsioukis et al. 2012; Taylor et al. 2013) substantiate that transfer learning can speed up the learning process.

作为第二种方法，知识重用可以以迁移学习的形式应用 (Da Silva et al. 2019；Da Silva 和 Costa 2019)。在一个任务学习过程中获得的经验也可能提高在相关但不同任务中的表现 (Taylor 和 Stone 2009)。Da Silva 和 Costa(2017) 使用了一个知识数据库，代理可以从该数据库中提取相关任务的先前解决方案，并将这些信息嵌入到当前任务的训练中。同样，Da Silva et al.(2017) 应用了专家演示，其中代理扮演学生角色，向教师寻求建议。他们证明，同时学习的代理可以通过知识转移相互提供建议。关于迁移学习的进一步研究可以在合作多代理环境中找到 (Omid-shafiei et al. 2019) 以及自然语言应用中 (Luketina et al. 2019)。在一般的多代理系统中，(Boutsioukis et al. 2012；Taylor et al. 2013) 的研究证实迁移学习可以加速学习过程。

Besides parameter sharing and transfer learning, curriculum learning may be applied for the scaling to many agents. Since tasks become more challenging to master and more time consuming to train as the number of agents increases, it is often challenging to learn from scratch. Curriculum learning starts with a small number of agents and then gradually enlarges the number of agents over the training course. Through the steady increase within the curriculum, trained policies can perform better than without a curriculum (Gupta et al. 2017; Long et al. 2020; Narvekar et al. 2016). Curriculum learning schemes can also cause improved generalization and faster convergence of agent policies (Bengio et al. 2009). Further works show that agents can generate learning curricula automatically (Sukhbaatar et al. 2017; Svetlik et al. 2017) or can create arms races in competitive settings (Baker et al.

除了参数共享和迁移学习，课程学习也可能适用于多智能体的扩展。因为随着智能体数量的增加，任务变得更加难以掌握和更加耗时的训练，所以从头开始学习往往具有挑战性。课程学习从少量智能体开始，然后在训练过程中逐渐增加智能体的数量。通过课程中的稳步增加，训练的策略可以比没有课程时表现得更好 (Gupta et al. 2017; Long et al. 2020; Narvekar et al. 2016)。课程学习方案还可以导致智能体策略的泛化改进和更快收敛 (Bengio et al. 2009)。进一步的研究表明，智能体可以自动生成学习课程 (Sukhbaatar et al. 2017; Svetlik et al. 2017)，或者在竞争环境中创造军备竞赛 (Baker et al.2020).

Complexity reduction Many real-world applications naturally encompass large numbers of simultaneously interacting agents (Nguyen et al. 2017a, b). As the quantity of agents increases, the requirement to contain the curse of dimensionality becomes inevitable. Yang et al. (2018b) addressed the issue of scalability with a mean-field method. The interactions between large numbers of agents are estimated by the impact of a single agent compared to the mean impact of the whole or local agent population. The complexity reduces as the problem is broken down into pairwise interactions between an agent and its neighborhood. Regarding the average effect to its neighbors, each agent learns the best response towards its proximity. Another approach to constrain the explosion in complexity is by factorizing the problem into smaller sub-problems (Guestrin et al. 2002). Chen et al. (2018) decomposed the joint action-value function into independent components and used pairwise interactions between agents to render large-scale problems computationally tractable. Further works studied large-scale MADRL problems with graphical models (Nguyen et al. 2017a) and the CTDE paradigm (Lin et al. 2018).

复杂度降低许多现实世界应用自然包含大量同时相互作用的代理 (Nguyen et al. 2017a, b)。随着代理数量的增加，控制维度诅咒的需求变得不可避免。Yang et al. (2018b) 使用均值场方法解决了可扩展性问题。大量代理之间的交互通过单个代理相对于整体或局部代理群体平均影响的效应来估计。当问题被分解为代理及其邻域之间的成对交互时，复杂度降低。关于对邻居的平均效应，每个代理学习对其邻近区

域的最优响应。另一种限制复杂性爆炸的方法是通过将问题分解为更小的子问题 (Guestrin et al. 2002)。Chen et al. (2018) 将联合动作价值函数分解为独立组件，并使用代理之间的成对交互使大规模问题在计算上可行。进一步的工作研究了使用图形模型 (Nguyen et al. 2017a) 和 CTDE 范式 (Lin et al. 2018) 的大规模多智能体深度强化学习问题。

Robustness Another desired property is the robustness of learned policies to perturbations in the environment caused by other agents. Perturbations are fortified by the number of agents and the resulting growth of the state-action space. In supervised learning, a common problem is that models can over-fit to the data set. Similarly, over-fitting can occur in RL frameworks if environments provide little or no deviation (Bansal et al. 2018). To maintain robustness over the training process and to the other agents' adaption, several methods have been proposed.

鲁棒性另一个期望的性质是学习到的策略对于由其他代理引起的环境扰动的鲁棒性。扰动因代理数量和状态-动作空间的增长而加剧。在监督学习中，一个常见问题是模型可能会过拟合到数据集。同样，如果环境提供很少或没有偏差，强化学习框架中也可能出现过拟合 (Bansal et al. 2018)。为了在训练过程中以及对于其他代理的适应保持鲁棒性，已经提出了几种方法。

First, regularization techniques can be used to prevent over-fitting to other agents' behavior. Examples can be seen in policies ensembles (Lowe et al. 2017), where a collection of different sub-policies is trained for each agent, or can be found in best responses to policy mixtures (Lanctot et al. 2017).

首先，正则化技术可以被用来防止过度拟合其他智能体的行为。例如，在策略集成 (Lowe 等人，2017 年) 中可以看到，每个智能体都训练了一组不同的子策略，或者在策略混合的最佳响应中 (Lanctot 等人，2017 年) 也可以找到。

Second, adversarial training can be applied to mitigate the vulnerability of polices towards perturbations. Pinto et al. (2017) added an adversarial agent to the environment that applied targeted disturbances to the learning process. By hampering the training, the agents were compelled to encounter these disturbances and develop robust policies. Similarly, Li et al. (2019a) used an adversarial setting to reduce the sensitivity of agents towards the environment. Bansal et al. (2018) demonstrated that policies, which are trained in a competitive setting, could yield behaviors that are far more complex than the environment itself. From an application perspective, Spooner and Savani (2020) studied robust decision-making in market making.

其次，对抗训练可以应用于减轻策略对扰动的脆弱性。Pinto 等人 (2017 年) 在环境中添加了一个对抗性智能体，该智能体对学习过程应用了有针对性的干扰。通过阻碍训练，智能体被迫遭遇这些干扰并开发出鲁棒的策略。同样，Li 等人 (2019a 年) 使用了一个对抗性设置来降低智能体对环境的敏感性。Bansal 等人 (2018 年) 证明了在竞争环境中训练的策略，其行为可能比环境本身复杂得多。从应用的角度来看，Spooner 和 Savani(2020 年) 研究了市场制作中的鲁棒决策。

The observations from above are in accordance with the findings of related studies about the impact of self-play (Raghu et al. 2018; Sukhbaatar et al. 2017). Heinrich and Silver (2016) used self-play to learn approximate Nash equilibria of imperfect-information games and showed that self-play could be used to obtain better robustness in the learned policies. Similarly, self-play was used to compete with older versions of policies to render the learned behaviors more robust (Baker et al. 2020; Berner et al. 2019; Silver et al. 2018). Silver et al. (2016) adapted self-play as a regularization technique to prevent the policy network from over-fitting by playing against older versions of itself. However, Gleave et al. (2020) studied the existence of adversarial policies in competitive games and showed that complex policies could be fooled by comparably easy strategies. Although agents trained through self-play proved to be more robust, allegedly random and uncoordinated strategies caused agents to fail at the task. They argued that the vulnerability towards adversarial attacks increases with the dimensionality of the observation space. A further research direction for addressing robustness is to render the learning representation invariant towards permutations, as shown in Liu et al. (2020).

上述观察与关于自我游戏影响的相关研究结论相符 (Raghu 等人，2018 年；Sukhbaatar 等人，2017 年)。Heinrich 和 Silver(2016 年) 使用自我游戏来学习不完整信息游戏的近似纳什均衡，并展示了自我游戏可以用来提高学习策略的鲁棒性。同样，自我游戏被用来与策略的旧版本竞争，使学习到的行为更加鲁棒 (Baker 等人，2020 年；Berner 等人，2019 年；Silver 等人，2018 年)。Silver 等人 (2016 年) 将自我游戏作为一种正则化技术，通过与自己旧版本对抗来防止策略网络过拟合。然而，Gleave 等人 (2020 年) 研究了竞争游戏中对抗策略的存在，并展示了复杂策略可能被相对简单的策略欺骗。尽管通过自我游戏训练的代理证明了更加鲁棒，但据称随机且不协调的策略导致代理在任务上失败。他们认为，观察空间的维度越大，对对抗攻击的脆弱性就越高。解决鲁棒性的一个进一步研究方向是使学习表征对置换不变，正如 Liu 等人 (2020 年) 所示。

5.6 Partial observability

5.6 部分可观测性

Outside an idealized setting, agents neither can observe the global state of the environment, nor do they have access to the internal knowledge of other agents. By perceiving only partial observations, a single observation does not capture all relevant information about the environment and its history. Hence, the Markov property is not fulfilled, and the environment appears non-Markovian. An additional difficulty elicited by partial observability is the lazy agent problem which can occur in cooperative settings (Sunehag et al. 2018). As introduced in Sect. 2.2, the common frameworks that deal with partial observability are POMPDs for general settings and dec-POMDPs for cooperative settings with a shared reward function. Dec-POMDPs are computationally challenging (Bernstein et al. 2002) and still intractable when solving problems with real-world complexity (Amato et al. 2015). However, recent work accomplished promising results in video games with imperfect information (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019; Vinyals et al.

在非理想化的环境中，智能体既无法观察到环境的全局状态，也无法获取其他智能体的内部知识。由于只能感知到部分观察，单个观察并不能捕捉到关于环境和其历史的所有相关信息。因此，马尔可夫性质没有得到满足，环境表现为非马尔可夫性。部分可观测性引起的另一个困难是懒惰智能体问题，这在合作环境中可能会出现 (Sunehag et al. 2018)。如第 2.2 节所述，处理部分可观测性的常见框架包括适用于一般环境的 POMPDs 和适用于具有共享奖励函数的合作环境的 dec-POMDPs。Dec-POMDPs 在计算上具有挑战性 (Bernstein et al. 2002)，并且在解决具有现实世界复杂性的问题时仍然不可处理 (Amato et al. 2015)。然而，最近在具有不完整信息的电子游戏方面取得了有希望的结果 (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019; Vinyals et al.2019).

A natural way to deal with non-Markovian environments is through information exchange between the decision-makers (Goldman and Zilberstein 2004). Agents that are able to communicate can compensate for their limited knowledge by propagating information and fill the lack of knowledge about other agents or the environment (Foerster et al. 2016). As we already discussed in Sect. 5.2, there are several ways to incorporate communication capabilities into agents. A primary example is Jiang and Lu (2018) who used an attention mechanism to establish communication under partial observations. Rather than having a fixed frequency for the information exchange, they learned to communicate on-demand. Further approaches under partial observability have been investigated in cooperative tasks (Das et al. 2019; Sukhbaatar et al. 2016) or mixed settings (Singh et al. 2019).

处理非马尔可夫环境的自然方式是通过决策者之间的信息交换 (Goldman 和 Zilberstein 2004)。能够进行通信的代理可以通过传播信息来补偿他们有限的知识，并填补对其他代理或环境知识的不足 (Foerster 等人 2016 年)。正如我们已在第 5.2 节中讨论的那样，有几种方法可以将通信能力整合到代理中。一个主要的例子是 Jiang 和 Lu(2018 年)，他们使用注意力机制在部分观察下建立通信。他们不是固定信息交换的频率，而是学会了按需通信。在部分可观察性下的其他方法已经在合作任务 (Das 等人 2019 年；Sukhbaatar 等人 2016 年) 或混合设置 (Singh 等人 2019 年) 中进行了研究。

In the following, we review papers that cope with partial observability by incorporating a memory mechanism. Agents, which have the capability of memorizing past experiences, can compensate for the lack of information.

在接下来的内容中，我们回顾了通过整合记忆机制来处理部分可观察性的论文。具有记忆过去经验能力的代理能够补偿信息不足的问题。

Memory mechanism A common way to tackle partial observability is the usage of deep recurrent neural networks, which equip agents with a memory mechanism to store information that can be relevant in the future (Hausknecht and Stone 2015). However, long-term dependencies render the decision-making difficult since experiences that were observed in the further past may have been forgotten (Hochreiter and Schmidhuber 1997). Approaches involving recurrent neural networks to deal with partial observability can be realized with value-based approaches (Omidshafiei et al. 2017) or actor-critic methods (Dibangoye and Buffet 2018; Foerster et al. 2018b; Gupta et al. 2017). Foerster et al. (2019) used a Bayesian method to tackle partial observability in cooperative settings. They used all publicly available features of the environment and agents to determine a public belief over the agents' internal states. A severe concern in MADRL is that the memorization of past information is exacerbated by the number of agents involved during the learning process.

记忆机制处理部分可观测性的常见方法是使用深度循环神经网络，这为智能体配备了一种记忆机制，用于存储可能在将来相关的信息 (Hausknecht 和 Stone 2015)。然而，长期依赖性使得决策变得困难，因为更早之前观察到的经验可能已经被遗忘 (Hochreiter 和 Schmidhuber 1997)。使用循环神经网络处理部分可观测性的方法可以通过基于价值的方法 (Omidshafiei 等人 2017) 或演员-评论家方法 (Dibangoye 和

Buffet 2018；Foerster 等人 2018b；Gupta 等人 2017) 来实现。Foerster 等人 (2019) 使用贝叶斯方法来解决合作环境中的部分可观测性问题。他们使用环境中所有公开可用的特征和智能体的特征来确定智能体内部状态的公共信念。在多智能体深度强化学习 (MADRL) 中的一个严重问题是，过去信息的记忆化会随着学习过程中涉及智能体的数量增加而加剧。

6 Discussion

6 讨论

In this section, we discuss findings from previous sections. We enumerate trends that we have identified in recent literature. Since these trends are useful for addressing current challenges, they may also be an avenue for upcoming research. To the end of our discussion, we point out possible future work. We elaborate on problems where only a minority of research has been conducted and pose two problems which we find the toughest ones to overcome.

在本节中，我们讨论了前几节的发现。我们列举了我们在最近文献中识别的趋势。由于这些趋势对于解决当前的挑战很有用，它们也可能是未来研究的途径。在我们讨论的结尾，我们指出了可能未来的工作。我们详细讨论了只有少数研究进行过的问题，并提出了我们认为是克服起来最困难的两个问题。

Despite the recent advances in many directions, many pathologies such as relative overgeneralization combined with reward stochasticity are not yet solved, even in allegedly simple tabular worlds. MADRL has taken profit from the history of MARL by scaling up the insights to more complex problems. Approaches where strong solutions exist in simplified MARL settings may be transferable to the MADRL domain. Thus by enhancing older methods with new deep learning approaches, unsolved problems and concepts from MARL continue to matter in MADRL. An essential point for MADRL is that reproducibility is taken conscientiously. Well-known papers from the single-agent domain underline the significance of hyper-parameters, the number of independent random seeds, and chosen code-base towards the eventual task performance (Henderson et al. 2018; Islam et al. 2017). To maintain steady progress, the reporting of all used hyper-parameters and a transparent conduction of experiments is crucial. We want to make the community aware that these findings may also be valid for the multi-agent domain. Therefore, it is inevitable that standardized frameworks are created in which different algorithms can be compared along with their merits and demerits. Many individual environments have been proposed which exhibit intricate structure and real-world complexity (Baker et al. 2020; Beattie et al. 2016; Johnson et al. 2016; Juliani et al. 2018; Song et al. 2019; Vinyals et al. 2017). However, no consistent benchmark yet exists that provides a unified interface and allows a fair comparison between different kinds of algorithms grounded on a great variety of tasks like the OpenAI Gym (Brockman et al. 2016) for single-agent problems.

尽管在许多方向上已经有了最近的进展，但许多病理现象，如相对过度泛化与奖励随机性的结合，即使在所谓的简单表格世界中尚未解决。多智能体强化学习 (MADRL) 从多智能体强化学习 (MARL) 的历史中获益，通过将见解扩展到更复杂的问题。在简化的 MARL 环境中存在强解决方案的方法可能可以迁移到 MADRL 领域。因此，通过用新的深度学习方法增强旧方法，未解决的问题和 MARL 的概念继续在 MADRL 中具有重要意义。对于 MADRL 来说，一个基本点是可重复性被认真对待。来自单一智能体领域的知名论文强调了超参数、独立随机种子的数量以及选择的代码库对最终任务性能的重要性 (Henderson 等人，2018；Islam 等人，2017)。为了保持稳定的进步，报告所有使用的超参数以及透明地 Conduct 实验是至关重要的。我们希望让社区意识到这些发现可能也适用于多智能体领域。因此，创建标准化的框架是不可避免的，其中不同的算法可以进行比较，以及它们的优点和缺点。已经提出了许多具有复杂结构和现实世界复杂性的个体环境 (Baker 等人，2020；Beattie 等人，2016；Johnson 等人，2016；Juliani 等人，2018；Song 等人，2019；Vinyals 等人，2017)。然而，还没有一个统一的接口提供一致的基准，允许在 OpenAI Gym(Brockman 等人，2016) 这样的基于各种任务的不同算法之间进行公平比较，用于单一智能体问题。

Table 4 Our identified trends in MADRL and the addressed challenges

表 4 我们在 MADRL 中识别的趋势和解决的挑战

Trend	Addressed challenge(s)
Curriculum learning	Scalability
Memory	Non-stationarity, partial observability
Communication	Non-stationarity, coordination, partial observability
CTDE	Non-stationarity, coordination, partial observabil- ity, credit assignment, scalability

趋势	解决的挑战
课程学习	可扩展性
内存	非平稳性，部分可观测性
通信	非平稳性，协调，部分可观测性
CTDE(连续时间决策过程)	非平稳性，协调，部分可观测性，信用分配，可扩展性

6.1 Trends

6.1 趋势

Over the last years, approaches in the multi-agent domain achieved successes based on recurring patterns of good practice. We have identified four trends in state-of-the-art literature that have been frequently applied to address current challenges (Table 4).

在过去的几年里，多智能体领域的相关方法基于反复出现的良好实践模式取得了成功。我们在最新的文献中识别出了四种经常应用于应对当前挑战的趋势 (见表 4)。

As the first trend, we observe curriculum learning as an approach to divide the learning process into stages to deal with scalability issues. By starting with a small quantity, the number of agents is gradually enlarged over the learning course so that large-scale training becomes feasible (Gupta et al. 2017; Long et al. 2020; Narvekar et al. 2016). Alternatively, curricula can also be employed to create different stages of difficulty, where agents face relatively easy tasks at the beginning and gradually more complex tasks as their skills increase (Vinyals et al. 2019). Besides that, curriculum training is used to investigate the emergence of agent behavior. Curricula describe engineered changes in the dynamics of the environment. Agents adapt their behaviors over time in response to the strategic changes of others, which can yield arms races between agents. This process of continual adaption is referred to autocurricula (Leibo et al. 2019), which have been reported in several works (Baker et al. 2020; Sukhbaatar et al. 2017; Svetlik et al. 2017).

作为第一种趋势，我们观察到课程学习是一种将学习过程划分为阶段来处理可扩展性问题的方法。通过从较小的数量开始，学习过程中智能体的数量逐渐增加，从而使大规模训练变得可行 (Gupta 等人，2017；Long 等人，2020；Narvekar 等人，2016)。或者，课程也可以用来创建不同难度的阶段，智能体在开始时面对相对简单的任务，随着技能的提升逐渐面对更复杂的任务 (Vinyals 等人，2019)。除此之外，课程训练还用于研究智能体行为的出现。课程描述了环境动态的工程化变化。智能体随着时间的推移适应其他人的策略变化，这可能导致智能体之间的军备竞赛。这种持续的适应过程被称为自动课程 (Leibo 等人，2019)，这在几项工作中有所报道 (Baker 等人，2020；Sukhbaatar 等人，2017；Svetlik 等人，2017)。

Second, we recognize a trend towards deep neural networks embedded with recurrent units to memorize experience. By having the ability to track the history of state transitions and the decisions of other agents, the non-stationarity of the environment due to multiple decision-makers and partially observable states can be addressed in small problems (Omid-shafiei et al. 2017), and can be managed sufficiently well in complex problems (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019).

第二，我们注意到一种趋势，即使用嵌入循环单元的深度神经网络来记忆经验。通过能够追踪状态转换的历史和其他智能体的决策，可以解决由于多个决策者和部分可观察状态导致的环境非平稳性问题，在小规模问题中 (Omid-shafiei 等人，2017)，并且在复杂问题中也能足够好地管理 (Baker 等人，2020；Berner 等人，2019；Jaderberg 等人，2019)。

Third, an active line of research is exploring the development of communication skills. Due to the rise of deep learning methods, new computational approaches are available to investigate the emergence of language between interactive agents (Lazaridou and Baroni 2020). Despite the plethora of works that analyze emergent behaviors and semantics, many works propose methods that endow agents with communication skills. By expressing their intension, agents can align their coordination and find a consensus (Foerster et al. 2016). The non-stationarity from the perspective of a single learner can be eluded when agents disclose their history. Moreover, agents can share their local information with others to alleviate partial observability (Foerster et al. 2018b; Omidshafiei et al. 2017).

第三，一个活跃的研究方向是探索交流技能的发展。由于深度学习方法的兴起，新的计算方法可以用来研究交互式代理之间语言的出现 (Lazaridou 和 Baroni 2020)。尽管有许多作品分析了出现的性行为和语义，但许多作品提出了赋予代理交流技能的方法。通过表达它们的意图，代理可以协调一致并找到共识 (Foerster 等人 2016 年)。当代理披露它们的历史时，单个学习者的非平稳性可以被避免。此外，代理可以与其他代理共享他们的局部信息，以减轻部分可观测性 (Foerster 等人 2018b；Omidshafiei 等人 2017 年)。

Fourth and last, we note a clear trend towards the CTDE paradigm that enables the sha-ing of information during the training. Local information such as the observation-action history, function values, or policies can be made available to all agents during the training, which renders the environment stationary from the viewpoint of an individual agent and may diminish partial observability (Lowe et al. 2017). Further, the credit assignment problem can be addressed when information is available about all agents, and a centralized mechanism can attribute the individual contribution to the respective agent (Foerster et al. 2018b). Further challenges that can be loosened are coordination and scalability when the lack of information of an individual agent is compensated, and the learning process is accelerated (Gupta et al. 2017).

第四，也是最后一点，我们注意到一个明显的趋势是朝着 CTDE 范式发展，该范式使得在训练期间共享信息成为可能。诸如观察-动作历史、函数值或策略等局部信息可以在训练期间提供给所有代理，这使得从单个代理的角度来看环境是平稳的，并可能减少部分可观测性 (Lowe 等人 2017 年)。此外，当所有代理的信息可用时，可以解决信用分配问题，并且集中机制可以将个体贡献归因于相应的代理 (Foerster 等人 2018b)。当单个代理的信息不足得到补偿，学习过程加速时，可以缓解的进一步挑战是协调和可扩展性 (Gupta 等人 2017 年)。

6.2 Future work

6.2 未来工作

Next to our identified trends, which are already under active research, we recognize areas that have not been sufficiently explored yet. One such area is multi-goal learning where each agent has an individually associated goal that needs to be optimized. However, global optimality can only be accomplished if agents also allow others to be successful in their task (Yang et al. 2020). Typical scenarios are cooperative tasks such as public good dilemmas, where agents are obliged to the sustainable use of limited resources, or autonomous driving, where agents have individual destinations and are supposed to coordinate the path-finding to avoid crashes. A similar direction is multi-task learning where agents are expected to perform well not only on one single but also on related other tasks (Omid-shafiei et al. 2017; Taylor and Stone 2009). Besides multi-goal and multi-task learning, another avenue for future work is present in safe MADRL. Safety is a highly desired property because autonomously acting agents are expected to ensure system performance while holding to safety guarantees during learning and employment (García et al. 2015). Several works in single-agent RL are concerned with safety concepts, but its applicability to multiple agents is limited and still in its infancy (Zhang and Bastani 2019; Zhu et al. 2020). Akin to the growing interest in learning to communicate, a similar effect may happen in the multi-agent domain, where deep learning methods open new paths. For an application perspective on safe autonomous driving, one can consider the article by Shalev-Shwartz et al. (2016). Another possible direction for future research offers the intersection between MADRL and evolutionary methodologies. Evolutionary algorithms have been used in versatile contexts of multi-agent RL, e.g. for building intrinsic motivation (Wang et al. 2019), shaping rewards (Jaderberg et al. 2019), generating curricula (Long et al. 2020) and analyzing dynamics (Bloembergen et al. 2015). Since evolution requires many entities to adapt, multi-agent RL is a natural playground for such algorithms.

在我们已经识别并正在积极研究的发展趋势之外，我们还注意到一些尚未得到充分探索的领域。其中一个领域是多目标学习，其中每个代理都有与其个体关联的目标需要优化。然而，只有当代理也允许其他代理在其任务中取得成功时 (Yang et al. 2020)，才能实现全局最优。典型的场景是合作任务，如公共物品困境，其中代理有义务可持续地使用有限的资源，或者自动驾驶，其中代理有各自的目的地，并需要协调路径查找以避免碰撞。一个类似的方向是多任务学习，其中代理不仅在一个单一任务上，而且在相关的其他任务上也被期望表现良好 (Omid-shafiei et al. 2017; Taylor 和 Stone 2009)。除了多目标和多任务学习之外，未来工作的另一个途径体现在安全的 MADRL 中。安全性是一个高度期望的属性，因为自主行动的代理在学习和应用过程中被期望确保系统性能，同时遵守安全保证 (García et al. 2015)。单代理 RL 中的一些工作关注于安全概念，但其适用于多代理的范围有限，仍处于初级阶段 (Zhang 和 Bastani 2019; Zhu et al. 2020)。类似于对学习通信日益增长的兴趣，多代理领域中可能发生类似的效果，其中深度学习方法开辟了新的路径。对于自动驾驶应用角度的安全问题，可以考虑 Shalev-Shwartz 等人 (2016) 的文章。未来研究的另一个可能方向是 MADRL 和进化方法之间的交叉。进化算法已经在多代理 RL 的各种背景下得到应用，例如用于构建内在动机 (Wang et al. 2019)、塑造奖励 (Jaderberg et al. 2019)、生成课程 (Long et al. 2020) 和分析动态 (Bloembergen et al. 2015)。由于进化需要许多实体来适应，多代理 RL 自然成为这类算法的游乐场。

Beyond the current challenges and reviewed literature of Sect. 5, we identify two problems that we

regard as the most challenging problems to overcome by future work. We primarily choose these two problems since they are the ones that matter the most when it comes to the applicability of algorithms to real-world scenarios. Most research focuses on learning within homogeneous settings where agents share common interests and optimize a mutual goal. For instance, the learning of communication is mainly studied in dec-POMDPs, where agents are expected to optimize upon a joint reward signal. When agents share common interests, the CTDE paradigm is usually a beneficial choice to exchange information between agents, and problems like non-stationarity, partial observability, and coordination can be diminished. However, heterogeneity implies that agents may have their own interests and goals, individual experience and knowledge, or different skills and capabilities. Limited research has been conducted in heterogeneous scenarios, although many real-world problems naturally comprise a mixture of different entities. Under real-world conditions, agents have only access to local and heterogeneous information on which decisions must be taken. The fundamental problem in the multi-agent domain is and ever has been the curse of dimensionality (Busoniu et al. 2008; Hernandez-Leal et al. 2019). The state-action space and the combinatorial possibilities of agent interactions increase exponentially by the number of agents, which renders sufficient exploration itself a difficult problem. This is intensified when agents have only access to partial observations of the environment or when the environment is of continuous nature. Although powerful function approximators like neural networks can cope with continuous spaces and generalize well over large spaces, open questions remain like how to explore large and complex spaces sufficiently well and how to solve large combinatorial optimization problems.

在第 5 节中讨论的当前挑战和回顾的文献之外，我们确定了两项我们认为是未来工作需要克服的最具挑战性的问题。我们主要选择这两个问题，因为当涉及到算法在实际场景中的应用性时，它们是最重要的。大多数研究集中在同质环境中的学习，其中代理共享共同兴趣并优化共同目标。例如，通信的学习主要在 dec-POMDPs 中研究，其中代理被期望在联合奖励信号上进行优化。当代理共享共同兴趣时，CTDE 范式通常是代理之间交换信息的一个有益选择，并且可以减少非平稳性、部分可观测性和协调等问题。然而，异质性意味着代理可能有自己的兴趣和目标、个人经验和知识，或者不同的技能和能力。尽管许多现实世界问题自然包含不同实体的混合，但在异质场景中进行的研究却很有限。在现实条件下，代理只能访问局部和异质信息，必须在此基础上做出决策。多代理领域的基本问题始终是维度诅咒 (Busoniu 等人，2008 年；Hernandez-Leal 等人，2019 年)。状态-动作空间和代理交互的组合可能性随着代理数量的增加而指数增长，这使得足够探索本身就是一个难题。当代理只能访问环境的部分观察，或者环境是连续性质时，这个问题变得更加复杂。尽管强大的函数逼近器如神经网络能够处理连续空间并在大空间上泛化良好，但仍存在一些开放性问题，比如如何足够好地探索大而复杂的空间，以及如何解决大型组合优化问题。

7 Conclusion

7 结论

Even though multi-agent reinforcement learning enjoys a long record, historical approaches hardly exceeded the complexity of discretized environments with a limited amount of states and actions (Busoniu et al. 2008; Tuyls and Weiss 2012). Since the breakthrough of deep learning methods, the field is undergoing a rapid transformation, and many previously unsolved problems have become step by step tractable. Latest advances showed that tasks with real-world complexity could be mastered (Baker et al. 2020; Berner et al. 2019; Jader-berg et al. 2019; Vinyals et al. 2019). Still, MADRL is a young field which attracts growing interest, and the amount of published literature rises swiftly. In this article, we surveyed recent works that combine deep learning methods with multi-agent reinforcement learning. We analyzed training schemes that are used to learn policies, and we reviewed patterns of agent behavior that emerge when multiple entities interact simultaneously. In addition, we systematically investigated challenges that are present in the multi-agent context and studied recent approaches that are under active research. Finally, we outlined trends which we have identified in state-of-the-art literature and proposed possible avenues for future work. With this contribution, we want to equip interested readers with the necessary tools to understand the contemporary challenges in MADRL by providing a more holistic overview of the recent approaches. We want to emphasize its potential and reveal opportunities as well as its limitations. In the foreseeable future, we expect an abundance of new literature to emanate and, hence, we want to encourage the community for further developments in this interesting and young field of research.

尽管多智能体强化学习拥有悠久的历史，历史上的方法几乎无法超越离散化环境中的复杂性，这些环境具有有限的状态和动作数量 (Busoniu 等人，2008 年；Tuyls 和 Weiss，2012 年)。自从深度学习方法取得突破以来，该领域正在经历快速的变革，许多之前无法解决的问题已经逐步变得可处理。最新的进展

表明，具有现实世界复杂性的任务可以被掌握 (Baker 等人，2020 年；Berner 等人，2019 年；Jader-berg 等人，2019 年；Vinyals 等人，2019 年)。尽管如此，多智能体强化学习是一个年轻的领域，吸引了越来越多的关注，发表的文献数量也在迅速增长。在本文中，我们调研了结合深度学习方法和多智能体强化学习的近期工作。我们分析了用于学习策略的训练方案，并回顾了多个实体同时交互时出现的智能体行为模式。此外，我们系统地研究了多智能体背景下存在的挑战，并研究了正在积极研究的近期方法。最后，我们概述了我们在最新文献中识别的趋势，并提出了未来工作的可能途径。通过这项贡献，我们希望为感兴趣的读者提供必要的工具，以便通过提供对近期方法的更全面概述，理解 MADRL 中的当代挑战。我们希望强调其潜力，揭示机会以及局限性。在可预见的未来，我们预计将涌现大量新的文献，因此，我们希望鼓励社区在这个有趣且年轻的研究领域中进一步发展。

Acknowledgements We would like to thank the editor and the three anonymous reviewers for providing their comprehensive feedback. Without their suggestions, this manuscript would not look as it does in this final version. We want to thank our colleagues and friends who read through earlier versions of this manuscript. In particular, we appreciate the help of Matthias Kissel, Patrick Krämer, Anke Müller and Martin Gottwald.

致谢我们感谢编辑和三位匿名审稿人提供的全面反馈。如果没有他们的建议，这份手稿就不会有现在的最终版本。我们还想感谢阅读过这份手稿早期版本的同仁和朋友。特别是，我们感激 Matthias Kissel、Patrick Krämer、Anke Müller 和 Martin Gottwald 的帮助。

Funding Open Access funding enabled and organized by Projekt DEAL.

资助开放获取资金由 Projekt DEAL 资助和组织的。

References

参考文献

Ahilan S, Dayan P (2019) Feudal multi-agent hierarchies for cooperative reinforcement learning. CoRR arxiv: abs/1901.08492

Al-Shedivat M, Bansal T, Burda Y, Sutskever I, Mordatch I, Abbeel P (2018) Continuous adaptation via meta-learning in nonstationary and competitive environments. In: International conference on learning representations. https://openreview.net/forum?id=Sk2u1g-0-

Albrecht SV, Stone P (2018) Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif Intell 258:66-95. https://doi.org/10.1016/j.artint.2018.01.002.http://www.sciencedirect.com/science/article/pii/S0004370218300249

Amato C, Konidaris G, Cruz G, Maynor CA, How JP, Kaelbling LP (2015) Planning for decentralized control of multiple robots under uncertainty. In: 2015 IEEE international conference on robotics and automation (ICRA), pp 1241-1248. https://doi.org/10.1109/ICRA.2015.7139350

Amodei D, Olah C, Steinhardt J, Christiano PF, Schulman J, Mané D (2016) Concrete problems in AI safety. CoRR. arxiv: abs/1606.06565,

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D (2015) Vqa: Visual question answering. In: The IEEE international conference on computer vision (ICCV)

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) Deep reinforcement learning: a brief survey. IEEE Signal Process Mag 34(6):26-38. https://doi.org/10.1109/MSP.2017.2743240

Aubret A, Matignon L, Hassas S (2019) A survey on intrinsic motivation in reinforcement learning. arXiv e-prints arXiv:1908.06976,

Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I (2020) Emergent tool use from multi-agent autocurricula. In: International conference on learning representations. https://openreview.net/forum?id=SkxpxJBKwS

Bansal T, Pachocki J, Sidor S, Sutskever I, Mordatch I (2018) Emergent complexity via multi-agent competition. In: International conference on learning representations. https://openreview.net/forum?id=Sy0GnUxCb

Barde P, Roy J, Harvey FG, Nowrouzezahrai D, Pal C (2019) Promoting coordination through policy regularization in multi-agent reinforcement learning. arXiv e-prints arXiv:1908.02269,

Barrett S, Rosenfeld A, Kraus S, Stone P (2017) Making friends on the fly: cooperating with new teammates. Artif Intell 242:132-171

Beattie C, Leibo JZ, Teplyashin D, Ward T, Wainwright M, Küttler H, Lefrancq A, Green S, Valdés V, Sadik A, Schrittwieser J, Anderson K, York S, Cant M, Cain A, Bolton A, Gaffney S, King H, Has-sabis D, Legg S, Petersen S (2016) Deepmind lab. CoRR. arxiv: abs/1612.03801

Becker R, Zilberstein S, Lesser V, Goldman CV (2004) Solving transition independent decentralized Markov decision processes. J Artif Intell Res 22:423-455

Bellemare M, Srinivasan S, Ostrovski G, Schaul T, Saxton D, Munos R (2016) Unifying count-based exploration and intrinsic motivation. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 1471-1479. http://papers.nips.cc/paper/6383-unifying-count-based-exploration-and-intrinsic-motivation.pdf

Bellman R (1957) A Markovian decision process. J Math Mechanics 6(5):679-684. http://www.jstor.org/stable/24900506

Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning, ACM, New York, NY, USA, ICML '09, pp 41-48. https://doi.org/10.1145/1553374.1553380,

Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, Farhi D, Fischer Q, Hashme S, Hesse C, Józefowicz R, Gray S, Olsson C, Pachocki JW, Petrov M, de Oliveira Pinto HP, Raiman J, Salimans T, Schlatter J, Schneider J, Sidor S, Sutskever I, Tang J, Wolski F, Zhang S (2019) Dota 2 with large scale deep reinforcement learning. ArXiv arxiv: abs/1912.06680

Bernstein DS, Givan R, Immerman N, Zilberstein S (2002) The complexity of decentralized control of Markov decision processes. Math Oper Res 27(4):819-840. https://doi.org/10.1287/moor.27.4.819.297

Bertsekas DP (2012) Dynamic programming and optimal control, vol 2, 4th edn. Athena Scientific, Belmont

Bertsekas DP (2017) Dynamic programming and optimal control, vol 1, 4th edn. Athena Scientific, Belmont

Bloembergen D, Tuyls K, Hennes D, Kaisers M (2015) Evolutionary dynamics of multi-agent learning: a survey. J Artif Intell Res 53:659-697

Bono G, Dibangoye JS, Matignon L, Pereyron F, Simonin O (2019) Cooperative multi-agent policy gradient. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G (eds) Machine learning and knowledge discovery in databases. Springer International Publishing, Cham, pp 459-476

Boutsioukis G, Partalas I, Vlahavas I (2012) Transfer learning in multi-agent reinforcement learning domains. In: Sanner S, Hutter M (eds) Recent advances in reinforcement learning. Springer, Berlin, pp 249-260

Bowling M, Veloso M (2002) Multiagent learning using a variable learning rate. Artif Intell 136(2):215-250 3rockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W (2016) Openai gym. arXiv:1606.01540

Busoniu L, Babuska R, De Schutter B (2008) A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man Cybern Part C (Appl Rev) 38(2):156-172. https://doi.org/10.1109/TSMCC.2007.913919

Cai Y, Yang SX, Xu X (2013) A combined hierarchical reinforcement learning based approach for multi-robot cooperative target searching in complex unknown environments. In: 2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL), pp 52-59. https://doi.org/10.1109/ADPRL.2013

Cao K, Lazaridou A, Lanctot M, Leibo JZ, Tuyls K, Clark S (2018) Emergent communication through negotiation. In: International conference on learning representations. https://openreview.net/forum?id=Hk6WhagRW

Cao Y, Yu W, Ren W, Chen G (2013) An overview of recent progress in the study of distributed multi-agent coordination. IEEE Trans Industr Inf 9(1):427-438. https://doi.org/10.1109/TII.2012.2219061

Castellini J, Oliehoek FA, Savani R, Whiteson S (2019) The representational capacity of action-value networks for multi-agent reinforcement learning. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, international foundation for autonomous agents and multiagent systems, Richland, SC, AAMAS '19, pp 1862-1864. http://dl.acm.org/citat ion.cfm?id=3306127.3331944

Celikyilmaz A, Bosselut A, He X, Choi Y (2018) Deep communicating agents for abstractive summarization. CoRR arxiv: abs/1803.10357,

Chang Y, Ho T, Kaelbling LP (2004) All learning is local: Multi-agent learning in global reward games. In: Thrun S, Saul LK, Schölkopf B (eds) Advances in neural information processing systems 16, MIT Press, pp 807-814. http://papers.nips.cc/paper/2476-all-learning-is-local-multi-agent-learn ing-in-global-reward-games.pdf

Chen Y, Zhou M, Wen Y, Yang Y, Su Y, Zhang W, Zhang D, Wang J, Liu H (2018) Factorized q-learning for large-scale multi-agent systems. CoRR arxiv: abs/1809.03738

Chen YF, Liu M, Everett M, How JP (2016) Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. CoRR. arxiv: abs/1609.07845,

Chentanez N, Barto AG, Singh SP (2005) Intrinsically motivated reinforcement learning. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems 17, MIT Press, pp 1281-1288. http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf

Choi E, Lazaridou A, de Freitas N (2018) Multi-agent compositional communication learning from raw visual input. In: International conference on learning representations. https://openreview.net/forum?id=rknt2Be0-

Chu T, Chinchali S, Katti S (2020) Multi-agent reinforcement learning for networked system control. In: International conference on learning representations. https://openreview.net/forum?id=Syx7A3NFvH

Chu T, Wang J, Codecà L, Li Z (2020) Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans Intell Transp Syst 21(3):1086-1095

Chu X, Ye H (2017) Parameter sharing deep deterministic policy gradient for cooperative multi-agent reinforcement learning. CoRR arxiv: abs/1710.00336

Claus C, Boutilier C (1998) The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the fifteenth national conference on artificial intelligence and tenth innovative applications of artificial intelligence conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA, pp 746-752. http://www.aaai.org/Library/AAAI/1998/aaai98-106.php

Crandall JW, Goodrich MA (2011) Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. Mach Learn 82(3):281-314. https://doi.org/10.1007/s10994-010-5192-9

Da Silva FL, Costa AHR (2017) Accelerating multiagent reinforcement learning through transfer learning. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, AAAI Press, AAAI Press, AAAI'17, pp 5034-5035. http://dl.acm.org/citation.cfm?id=3297863.3297988

Da Silva FL, Costa AHR (2019) A survey on transfer learning for multiagent reinforcement learning systems. J Artif Int Res 64(1):645-703. https://doi.org/10.1613/jair.1.11396

Da Silva FL, Glatt R, Costa AHR (2017) Simultaneously learning and advising in multiagent reinforcement learning. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, international foundation for autonomous agents and multiagent systems, Richland, SC, AAMAS '17, pp 1100-1108. http://dl.acm.org/citation.cfm?id=3091210.3091280

Da Silva FL, Warnell G, Costa AHR, Stone P (2019) Agents teaching agents: a survey on inter-agent transfer learning. Auton Agent Multi-Agent Syst 34(1):9. https://doi.org/10.1007/s10458-019-09430-0

Das A, Kottur S, Moura JMF, Lee S, Batra D (2017) Learning cooperative visual dialog agents with deep reinforcement learning. In: The IEEE international conference on computer vision (ICCV)

Das A, Gervet T, Romoff J, Batra D, Parikh D, Rabbat M, Pineau J (2019) TarMAC: Targeted multi-agent communication. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 1538-1546. http://proceedings.mlr.press/v97/das19a.html

Dayan P, Hinton GE (1993) Feudal reinforcement learning. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems 5, Morgan-Kaufmann, pp 271-278. http://papers.nips.cc/paper/714-feudal-reinforcement-learning.pdf

De Cote EM, Lazaric A, Restelli M (2006) Learning to cooperate in multi-agent social dilemmas. In: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems, ACM, New York, NY, USA, AAMAS '06, pp 783-785. https://doi.org/10.1145/1160633.1160770

Diallo EAO, Sugiyama A, Sugawara T (2017) Learning to coordinate with deep reinforcement learning in doubles pong game. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA), pp 14-19. https://doi.org/10.1109/ICMLA.2017.0-184

Dibangoye J, Buffet O (2018) Learning to act in decentralized partially observable MDPs. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholms-mässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 1233-1242. http://proceedings.mlr.press/v80/dibangoye18a.html

Dobbe R, Fridovich-Keil D, Tomlin C (2017) Fully decentralized policies for multi-agent systems: an information theoretic approach. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 2941-2950. http://papers.nips.cc/paper/6887-fully-decentralized-policies-for-multi-agent-systems-an-information-theoretic-approach.pdf

Duan Y, Schulman J, Chen X, Bartlett PL, Sutskever I, Abbeel P (2016) Rl: fast reinforcement learning via slow reinforcement learning. CoRR arxiv: abs/1611.02779,

Eccles T, Bachrach Y, Lever G, Lazaridou A, Graepel T (2019) Biases for emergent communication in multi-agent reinforcement learning. In: Wallach H, Larochelle H, Beygelzimer A, Alche-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems 32, Curran Associates, Inc., pp 13111-13121. http://papers.nips.cc/paper/9470-biases-for-emergent-communication-in-multi-agent-reinforcement-learning.pdf

Everett R, Roberts S (2018) Learning against non-stationary agents with opponent modelling and deep reinforcement learning. In: 2018 AAAI Spring symposium series

Evtimova K, Drozdov A, Kiela D, Cho K (2018) Emergent communication in a multi-modal, multi-step referential game. In: International conference on learning representations. https://openreview.net/forum?id=rJGZq6g0-

Finn C, Levine S (2018) Meta-learning and universality: deep representations and gradient descent can approximate any learning algorithm. In: International conference on learning representations. https://openreview.net/forum?id=HyjC5yWCW

Foerster J, Assael IA, de Freitas N, Whiteson S (2016) Learning to communicate with deep multi-agent reinforcement learning. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 2137-2145. http://papers.nips.cc/paper/6042-learning-to-communicate-with-deep-multi-agent-reinforcement-learning.pdf

Foerster J, Nardelli N, Farquhar G, Afouras T, Torr PHS, Kohli P, Whiteson S (2017) Stabilising experience replay for deep multi-agent reinforcement learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 1146-1155. http://proceedings.mlr.press/v70/foerster17b

Foerster J, Chen RY, Al-Shedivat M, Whiteson S, Abbeel P, Mordatch I (2018a) Learning with opponent-learning awareness. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 122-130. http://dl.acm.org/citation.cfm?id=3237383.3237408

Foerster J, Farquhar G, Afouras T, Nardelli N, Whiteson S (2018b) Counterfactual multi-agent policy gradients. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17193

Foerster J, Song F, Hughes E, Burch N, Dunning I, Whiteson S, Botvinick M, Bowling M (2019) Bayesian action decoder for deep multi-agent reinforcement learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 1942-1951. http://proceedings.mlr.press/v97/foerster19a.html

Fulda N, Ventura D (2007) Predicting and preventing coordination problems in cooperative q-learning systems. In: Proceedings of the 20th international joint conference on artifical intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'07, pp 780-785

García J, Fern, o Fernández (2015) A comprehensive survey on safe reinforcement learning. J Mach Learn Res 16(42):1437-1480. http://jmlr.org/papers/v16/garcia15a.html

Ghavamzadeh M, Mahadevan S, Makar R (2006) Hierarchical multi-agent reinforcement learning. Auton Agent Multi-Agent Syst. https://doi.org/10.1007/s10458-006-7035-4

Gleave A, Dennis M, Wild C, Kant N, Levine S, Russell S (2020) Adversarial policies: Attacking deep reinforcement learning. In: International conference on learning representations. https://openreview.net/forum?id=HJgEMpVFwB

Goldman CV, Zilberstein S (2004) Decentralized control of cooperative systems: categorization and complexity analysis. J Artif Int Res 22(1):143-174. http://dl.acm.org/citation.cfm?id=1622487.1622493

Grover A, Al-Shedivat M, Gupta J, Burda Y, Edwards H (2018) Learning policy representations in multiagent systems. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol

80, pp 1802-1811. http://proceedings.mlr.press/v80/grover18a.html

Guestrin C, Koller D, Parr R (2002) Multiagent planning with factored mdps. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems 14, MIT Press, pp 1523-1530. http://papers.nips.cc/paper/1941-multiagent-planning-with-factored-mdps.pdf

Gupta JK, Egorov M, Kochenderfer M (2017) Cooperative multi-agent control using deep reinforcement learning. In: Sukthankar G, Rodriguez-Aguilar JA (eds) autonomous agents and multiagent systems. Springer, Cham, pp 66-83

Hadfield-Menell D, Milli S, Abbeel P, Russell SJ, Dragan A (2017) Inverse reward design. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 6765-6774. http://papers.nips.cc/paper/7253-inverse-reward-design.pdf

Han D, Boehmer W, Wooldridge M, Rogers A (2019) Multi-agent hierarchical reinforcement learning with dynamic termination. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, pp 2006-2008. http://dl.acm.org/citation.cfm?id=3306127.3331992

Hansen EA, Bernstein D, Zilberstein S (2004) Dynamic programming for partially observable stochastic games. In: AAAI

Hardin G (1968) The tragedy of the commons. Science 162(3859):1243-1248

Hausknecht M, Stone P (2015) Deep recurrent q-learning for partially observable mdps. https://www.aaai.org/ocs/inde

Havrylov S, Titov I (2017) Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 2149-2159. http://papers.nips.cc/paper/6810-emergence-of-language-with-multi-agent-games-learning-to-communicate-with-sequences-of-symbols.pdf

He H, Boyd-Graber J, Kwok K, III HD (2016) Opponent modeling in deep reinforcement learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd international conference on machine learning, PMLR, New York, New York, USA, Proceedings of Machine Learning Research, vol 48, pp 1804-1813. http://proceedings.mlr.press/v48/he16.html

He H, Chen D, Balakrishnan A, Liang P (2018) Decoupling strategy and generation in negotiation dialogues. CoRR arxiv: abs/1808.09637,

Heinrich J, Silver D (2016) Deep reinforcement learning from self-play in imperfect-information games. CoRR arxiv: abs/1603.01121,

Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D (2018) Deep reinforcement learning that matters. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669

Hernandez-Leal P, Kaisers M, Baarslag T, de Cote EM (2017) A survey of learning in multiagent environments: dealing with non-stationarity. CoRR arxiv: abs/1707.09183,

Hernandez-Leal P, Kartal B, Taylor ME (2019) Agent modeling as auxiliary task for deep reinforcement learning. CoRR arxiv: abs/1907.09597,

Hernandez-Leal P, Kartal B, Taylor ME (2019) A survey and critique of multiagent deep reinforcement learning. Auton Agent Multi-Agent Syst 33(6):750-797. https://doi.org/10.1007/s10458-019-09421-1

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hong Z, Su S, Shann T, Chang Y, Lee C (2017) A deep policy inference q-network for multi-agent systems. CoRR arxiv: abs/1712.07893,

Hoshen Y (2017) Vain: Attentional multi-agent predictive modeling. In: Proceedings of the 31st international conference on neural information processing systems, Curran Associates Inc., USA, NIPS'17, pp 2698-2708. http://dl.acm.org/citation.cfm?id=3294996.3295030

Houthooft R, Chen X, Chen X, Duan Y, Schulman J, De Turck F, Abbeel P (2016) Vime: variational information maximizing exploration. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems 29, Curran Associates, Inc., pp 1109-1117. http://papers.nips.cc/paper/6591-vime-variational-information-maximizing-exploration.pdf

Hu J, Wellman MP (1998) Multiagent reinforcement learning: theoretical framework and an algorithm. In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, ICML '98, pp 242-250. http://dl.acm.org/citation.cfm?id=645527.657296

Hu J, Wellman MP (2003) Nash q-learning for general-sum stochastic games. J Mach Learn Res 4:1039-1069

Hughes E, Leibo JZ, Phillips M, Tuyls K, Dueñez Guzman E, García Castañeda A, Dunning I, Zhu T, McKee K, Koster R, Roff H, Graepel T (2018) Inequity aversion improves cooperation in intertempo-

ral social dilemmas. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 3326-3336. http://papers.nips.cc/paper/7593-inequity-aversion-improves-cooperation-in-intertemporal-social-dilemmas.pdf

Iqbal S, Sha F (2019) Actor-attention-critic for multi-agent reinforcement learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 2961-2970. http://proceedings.mlr.press/v97/iqbal19a.html

Islam R, Henderson P, Gomrokchi M, Precup D (2017) Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. CoRR arxiv: abs/1708.04133,

Jaderberg M, Czarnecki WM, Dunning I, Marris L, Lever G, Castañeda AG, Beattie C, Rabinowitz NC, Morcos AS, Ruderman A, Sonnerat N, Green T, Deason L, Leibo JZ, Silver D, Hassabis D, Kavukcuoglu K, Graepel T (2019) Human-level performance in 3d multiplayer games with population-based reinforcement learning. Science 364(6443):859-865

Jain U, Weihs L, Kolve E, Rastegari M, Lazebnik S, Farhadi A, Schwing AG, Kembhavi A (2019) Two body problem: Collaborative visual task completion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Jaques N, Lazaridou A, Hughes E, Gülgehre Ç, Ortega PA, Strouse D, Leibo JZ, de Freitas N (2018) Intrinsic social motivation via causal influence in multi-agent RL. CoRR arxiv: abs/1810.08647,

Jaques N, Lazaridou A, Hughes E, Gulcehre C, Ortega P, Strouse D, Leibo JZ, De Freitas N (2019) Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: International conference on machine learning, pp 3040-3049

Jiang J, Lu Z (2018) Learning attentional communication for multi-agent cooperation. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 7254-7264. http://papers.nips.cc/paper/7956-learning-attentional-communication-for-multi-agent-cooperation.pdf

Johnson M, Hofmann K, Hutton T, Bignell D (2016) The malmo platform for artificial intelligence experimentation. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, AAAI Press, IJCAI'16, pp 4246-4247. http://dl.acm.org/citation.cfm?id=3061053.3061259

Jorge E, Kägebäck M, Gustavsson E (2016) Learning to play guess who? and inventing a grounded language as a consequence. CoRR arxiv: abs/1611.03218,

Juliani A, Berges V, Vckay E, Gao Y, Henry H, Mattar M, Lange D (2018) Unity: a general platform for intelligent agents. CoRR arxiv: abs/1809.02627,

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4(1):237-285. http://dl.acm.org/citation.cfm?id=1622737.1622748

Kasai T, Tenmoto H, Kamiya A (2008) Learning of communication codes in multi-agent reinforcement learning problem. In: 2008 IEEE conference on soft computing in industrial applications, pp 1-6

Kim W, Cho M, Sung Y (2019) Message-dropout: An efficient training method for multi-agent deep reinforcement learning. In: Proceedings of the AAAI conference on artificial intelligence 33(01):6079-6086

Kirby S (2002) Natural language from artificial life. Artif Life 8(2):185-215. https://doi.org/10.1162/106454602320184248

Kok JR, Vlassis N (2006) Collaborative multiagent reinforcement learning by payoff propagation. J Mach Learn Res 7:1789-1828. http://dl.acm.org/citation.cfm?id=1248547.1248612

Kollock P (1998) Social dilemmas: the anatomy of cooperation. Annu Rev Sociol 24(1):183-214. https://doi.org/10.1146/annurev.soc.24.1.183

Kong X, Xin B, Liu F, Wang Y (2017) Revisiting the master-slave architecture in multi-agent deep reinforcement learning. CoRR arxiv: abs/1712.07305,

Kraemer L, Banerjee B (2016) Multi-agent reinforcement learning as a rehearsal for decentralized planning. Neurocomputing 190:82-94

Kumar S, Shah P, Hakkani-Tür D, Heck LP (2017) Federated control with hierarchical multi-agent deep reinforcement learning. CoRR arxiv: abs/1712.08266,

Lanctot M, Zambaldi V, Gruslys A, Lazaridou A, Tuyls K, Perolat J, Silver D, Graepel T (2017) A unified game-theoretic approach to multiagent reinforcement learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 4190-4203. http://papers.nips.cc/paper/7007- a-unified-game-theoretic-approach-to-multiagent-reinforcement-learning.pdf

Lange PAV, Joireman J, Parks CD, Dijk EV (2013) The psychology of social dilemmas: a review. Organ Behav Hum Decis Process 120(2):125-141

Lauer M, Riedmiller M (2000) An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In: In Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, pp 535-542

Laurent GJ, Matignon L, Fort-Piat NL (2011) The world of independent learners is not markovian. Int J Knowl-Based Intell Eng Syst 15(1):55-64. http://dl.acm.org/citation.cfm?id=1971886.1971887

Lazaridou A, Baroni M (2020) Emergent multi-agent communication in the deep learning era. ArXiv arxiv: abs/2006.02419

Lazaridou A, Peysakhovich A, Baroni M (2017) Multi-agent cooperation and the emergence of (natural) language. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. https://openreview.net/forum?id= Hk8N3Sclg

Lazaridou A, Hermann KM, Tuyls K, Clark S (2018) Emergence of linguistic communication from referential games with symbolic and pixel input. In: International conference on learning representations. https://openreview.net/forum?id=HJGv1Z-AW

Le HM, Yue Y, Carr P, Lucey P (2017) Coordinated multi-agent imitation learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 1995-2003. http://proceedings.mlr.press/v70/le17a.html

Lee J, Cho K, Weston J, Kiela D (2017) Emergent translation in multi-agent communication. CoRR arxiv: abs/1710.06922,

Lee Y, Yang J, Lim JJ (2020) Learning to coordinate manipulation skills via skill behavior diversification. In: International conference on learning representations. https://openreview.net/forum?id=ryxB21BtvH

Leibo JZ, Zambaldi V, Lanctot M, Marecki J, Graepel T (2017) Multi-agent reinforcement learning in sequential social dilemmas. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '17, pp 464-473. http://dl.acm.org/citation.cfm?id=3091125.3091194

Leibo JZ, Hughes E, Lanctot M, Graepel T (2019) Autocurricula and the emergence of innovation from social interaction: a manifesto for multi-agent intelligence research. CoRR arxiv: abs/1903.00742,

Lerer A, Peysakhovich A (2017) Maintaining cooperation in complex social dilemmas using deep reinforcement learning. CoRR arxiv: abs/1707.01068,

Letcher A, Foerster J, Balduzzi D, Rocktäschel T, Whiteson S (2019) Stable opponent shaping in differentiable games. In: International conference on learning representations. https://openreview.net/forum?id=SyGjjsC5tQ

Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. Journal of Machine Learning Research 17(1):1334-1373. http://dl.acm.org/citation.cfm?id=2946645.2946684

Lewis M, Yarats D, Dauphin YN, Parikh D, Batra D (2017) Deal or no deal? end-to-end learning for negotiation dialogues. CoRR arxiv: abs/1706.05125,

Li F, Bowling M (2019) Ease-of-teaching and language structure from emergent communication. In: Wallach H, Larochelle H, Beygelzimer A, Alche-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems 32, Curran Associates, Inc., pp 15851-15861. http://papers.nips.cc/paper/9714-ease-of-teaching-and-language-structure-from-emergent-communication.pdf

Li S, Wu Y, Cui X, Dong H, Fang F, Russell S (2019a) Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. Proc AAAI Conf Artif Intell 33(01):4213-4220

Li X, Sun M, Li P (2019b) Multi-agent discussion mechanism for natural language generation. Proc AAAI Conf Artif Intell 33(01):6096-6103

Li Y (2018) Deep reinforcement learning. CoRR arxiv: abs/1810.06339,

Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2016) Continuous control with deep reinforcement learning. In: ICLR (Poster). http://arxiv.org/arxiv: abs/1509.02971

Lin K, Zhao R, Xu Z, Zhou J (2018) Efficient large-scale fleet management via multi-agent deep reinforcement learning. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, ACM, New York, NY, USA, KDD '18, pp 1774-1783. https://doi.org/10.1145/3219819.3219993,

Lin X, Beling PA, Cogill R (2018) Multiagent inverse reinforcement learning for two-person zero-sum games. IEEE Trans Games 10(1):56-68. https://doi.org/10.1109/TCIAIG.2017.2679115

Littman M (2001) Value-function reinforcement learning in markov games. Cogn Syst Res 2:55-66

Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the eleventh international conference on international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'94, pp 157-163. http://dl.acm.org/citation.cfm?id=3091574.30

Liu IJ, Yeh RA, Schwing AG (2020) Pic: Permutation invariant critic for multi-agent deep reinforcement learning. In: PMLR, proceedings of machine learning research, vol 100, pp 590-602. http://proceedings.mlr.press/v100/liu20a.html

Liu S, Lever G, Heess N, Merel J, Tunyasuvunakool S, Graepel T (2019) Emergent coordination through competition. In: International conference on learning representations. https://openreview.net/forum?id=BkG8sjR5Km

Long Q, Zhou Z, Gupta A, Fang F, Wu Y, Wang X (2020) Evolutionary population curriculum for scaling multi-agent reinforcement learning. In: International conference on learning representations. https://openreview.net/forum?id=SJxbHkrKDH

Lowe R, WU Y, Tamar A, Harb J, Pieter Abbeel O, Mordatch I (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 6379-6390. http://papers.nips.cc/paper/7217-multi-agent-actor-critic-for-mixed-cooperative-competitive-environments.pdf

Lowe R, Foerster JN, Boureau Y, Pineau J, Dauphin YN (2019) On the pitfalls of measuring emergent communication. CoRR arxiv: abs/1903.05168,

Luketina J, Nardelli N, Farquhar G, Foerster JN, Andreas J, Grefenstette E, Whiteson S, Rocktäschel T (2019) A survey of reinforcement learning informed by natural language. CoRR arxiv: abs/1906.03926,

Luong NC, Hoang DT, Gong S, Niyato D, Wang P, Liang Y, Kim DI (2019) Applications of deep reinforcement learning in communications and networking: a survey. IEEE Communications Surveys Tutorials pp 1-1. https://doi.org/10.1109/COMST.2019.2916583

Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. Nature 397(6719):498-500. https://doi.org/10.1038/17290

Lyu X, Amato C (2020) Likelihood quantile networks for coordinating multi-agent reinforcement learning. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems, pp 798-806

Ma J, Wu F (2020) Feudal multi-agent deep reinforcement learning for traffic signal control. In: Segh-rouchni AEF, Sukthankar G, An B, Yorke-Smith N (eds) Proceedings of the 19th international conference on autonomous agents and multiagent systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020, International Foundation for Autonomous Agents and Multiagent Systems, pp 816-824. https://dl.acm.org/doi/arxiv: abs/10.5555/3398761.3398858

Makar R, Mahadevan S, Ghavamzadeh M (2001) Hierarchical multi-agent reinforcement learning. In: Proceedings of the fifth international conference on autonomous agents, ACM, New York, NY, USA, AGENTS ' 01, pp 246–253. https://doi.org/10.1145/375735.376302,

Matignon L, Laurent GJ, Le Fort-Piat N (2007) Hysteretic q-learning : an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In: 2007 IEEE/RSJ international conference on intelligent robots and systems, pp 64-69

Matignon L, Jeanpierre L, Mouaddib AI (2012a) Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes. https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5038

Matignon L, Gj Laurent, Le fort piat N, (2012b) Review: independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. Knowl Eng Rev 27(1):1-31. https://doi.org/10.1017/S0269888912000057

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wier-stra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518:529 EP -. https://doi.org/10.1038/nature14236

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd international conference on machine learning, PMLR, New York, New York, USA, Proceedings of machine learning research, vol 48, pp 1928-1937. http://proceedings.mlr.press/v48/mniha 16.html

Moerland TM, Broekens J, Jonker CM (2018) Emotion in reinforcement learning agents and robots: a survey. Mach Learn 107(2):443-480. https://doi.org/10.1007/s10994-017-5666-0

Mordatch I, Abbeel P (2018) Emergence of grounded compositional language in multi-agent populations. https://aaai.org/ocs/index.php/AAAI/AAAI/AAAI18/paper/view/17007

Nair R, Tambe M, Yokoo M, Pynadath D, Marsella S (2003) Taming decentralized pomdps: towards efficient policy computation for multiagent settings. In: Proceedings of the 18th international joint conference on artificial intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'03,

pp 705-711. http://dl.acm.org/citation.cfm?id=1630659.1630762

Narvekar S, Sinapov J, Leonetti M, Stone P (2016) Source task creation for curriculum learning. In: Proceedings of the 2016 international conference on autonomous agents & multiagent systems, international foundation for autonomous agents and multiagent systems, Richland, SC, AAMAS '16, pp 566-574. http://dl.acm.org/citation.cfm?id=2936924.2937007

Nedic A, Ozdaglar A (2009) Distributed subgradient methods for multi-agent optimization. IEEE Trans Autom Control 54(1):48-61

Ng AY, Russell SJ (2000) Algorithms for inverse reinforcement learning. In: Proceedings of the seventeenth international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '00, pp 663-670. http://dl.acm.org/citation.cfm?id=645529.657801

Ng AY, Harada D, Russell S (1999) Policy invariance under reward transformations: theory and application to reward shaping. In: In Proceedings of the sixteenth international conference on machine learning, Morgan Kaufmann, pp 278-287

Nguyen DT, Kumar A, Lau HC (2017a) Collective multiagent sequential decision making under uncertainty. https://aaai.org/ocs/index.php/AAAI/AAAI/7/paper/view/14891

Nguyen DT, Kumar A, Lau HC (2017b) Policy gradient with value function approximation for collective multiagent planning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 4319-4329. http://papers.nips.cc/paper/7019-policy-gradient-with-value-function-approximation-for-collective-multiagent-planning.pdf

Nguyen DT, Kumar A, Lau HC (2018) Credit assignment for collective multiagent rl with global rewards. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 8102-8113. http://papers.nips.cc/paper/8033-credit-assignment-for-collective-multiagent-rl-with-global-rewards.pdf

Nguyen TT, Nguyen ND, Nahavandi S (2020) Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Trans Cybern 50(9):3826-3839

Oliehoek FA, Amato C (2016) A Concise Introduction to Decentralized POMDPs, 1st edn. Springer Publishing Company, Berlin

Oliehoek FA, Spaan MTJ, Vlassis N (2008) Optimal and approximate q-value functions for decentralized pomdps. J Artif Int Res 32(1):289-353. http://dl.acm.org/citation.cfm?id=1622673.1622680

Omidshafiei S, Pazis J, Amato C, How JP, Vian J (2017) Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of machine learning research, vol 70, pp 2681-2690. http://proceedings.mlr.press/v70/omidshafiei17a.html

Omidshafiei S, Kim DK, Liu M, Tesauro G, Riemer M, Amato C, Campbell M, How JP (2019) Learning to teach in cooperative multiagent reinforcement learning. Proc AAAI Conf Artif Intelli 33(01):6128-6136

Oroojlooyjadid A, Hajinezhad D (2019) A review of cooperative multi-agent deep reinforcement learning. ArXiv arxiv: abs/1908.03963

Oudeyer PY, Kaplan F (2007) What is intrinsic motivation? A typology of computational approaches. Front Neurorobotics 1:6-6

Palmer G, Tuyls K, Bloembergen D, Savani R (2018) Lenient multi-agent deep reinforcement learning. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 443-451. http://dl.acm.org/citation.cfm?id=3237383.3237451

Palmer G, Savani R, Tuyls K (2019) Negative update intervals in deep multi-agent reinforcement learning. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, pp 43-51

Panait L, Luke S (2005) Cooperative multi-agent learning: the state of the art. Auton Agent Multi-Agent Syst 11(3):387-434. https://doi.org/10.1007/s10458-005-2631-2

Panait L, Sullivan K, Luke S (2006) Lenient learners in cooperative multiagent systems. In: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems, association for computing machinery, New York, NY, USA, AAMAS '06, pp 801-803. https://doi.org/10.1145/1160633.1160776,

Papoudakis G, Christianos F, Rahman A, Albrecht SV (2019) Dealing with non-stationarity in multi-agent deep reinforcement learning. CoRR arxiv: abs/1906.04737,

Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-driven exploration by self-supervised prediction. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 2778-2787. http://proceedings.mlr.press/v70/pathak17a.html

Peng P, Yuan Q, Wen Y, Yang Y, Tang Z, Long H, Wang J (2017) Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. CoRR arxiv: abs/1703.10069,

Pérolat J, Leibo JZ, Zambaldi V, Beattie C, Tuyls K, Graepel T (2017) A multi-agent reinforcement learning model of common-pool resource appropriation. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 3643-3652. http://papers.nips.cc/paper/6955-a-multi-agent-reinforcem ent-learning-model-of-common-pool-resource-appropriation.pdf

Peysakhovich A, Lerer A (2018) Prosocial learning agents solve generalized stag hunts better than selfish ones. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, international Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 2043-2044. http://dl.acm.org/citation.cfm?id=3237383.3238065

Pinto L, Davidson J, Sukthankar R, Gupta A (2017) Robust adversarial reinforcement learning. In: Pre-cup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of machine learning research, vol 70, pp 2817-2826. http://proceedings.mlr.press/v70/pinto17a.html

Pinyol I, Sabater-Mir J (2013) Computational trust and reputation models for open multi-agent systems: a review. Artif Intell Rev 40(1):1-25. https://doi.org/10.1007/s10462-011-9277-z

Potter MA, De Jong KA (1994) A cooperative coevolutionary approach to function optimization. In: Davi-dor Y, Schwefel HP, Männer R (eds) Parallel problem solving from nature - PPSN III. Springer, Berlin, pp 249-257

Qu G, Wierman A, Li N (2020) Scalable reinforcement learning of localized policies for multi-agent networked systems. PMLR, The Cloud, Proceedings of machine learning research, vol 120, pp 256-266. http://proceedings.mlr.press/v120/qu20a.html

Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SMA, Botvinick M (2018) Machine theory of mind. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4218- 4227. http://proceedings.mlr.press/v80/rabinowitz18a.html

Raghu M, Irpan A, Andreas J, Kleinberg B, Le Q, Kleinberg J (2018) Can deep reinforcement learning solve Erdos-Selfridge-Spencer games? In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4238-4246. http://proceedings.mlr.press/v80/raghu18a.html

Raileanu R, Denton E, Szlam A, Fergus R (2018) Modeling others using oneself in multi-agent reinforcement learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4257-4266. http://proceedings.mlr.press/v80/raileanu18a.html

Ramchurn SD, Huynh D, Jennings NR (2004) Trust in multi-agent systems. Knowl Eng Rev 19(1):1–25. https://doi.org/10.1017/S0269888904000116

Rashid T, Samvelyan M, Schroeder C, Farquhar G, Foerster J, Whiteson S (2018) QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4295-4304. http://proceedings.mlr.press/v80/rashid18a.html

Russell S, Zimdars AL (2003) Q-decomposition for reinforcement learning agents. In: Proceedings of the twentieth international conference on international conference on machine learning, AAAI Press, ICML'03, pp 656-663. http://dl.acm.org/citation.cfm?id=3041838.3041921

Schaul T, Horgan D, Gregor K, Silver D (2015) Universal value function approximators. In: Proceedings of the 32nd international conference on international conference on machine learning - volume 37, JMLR.org, ICML'15, pp 1312-1320

Schmidhuber J (2010) Formal theory of creativity, fun, and intrinsic motivation (1990-2010). IEEE Trans Auton Ment Dev 2(3):230-247. https://doi.org/10.1109/TAMD.2010.2056368

Schmidhuber J, Zhao J, Wiering M (1996) Simple principles of metalearning. Tech. rep

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. CoRR arxiv: abs/1707.06347,

Sen S, Weiss G (1999) Multiagent systems. MIT Press, Cambridge, MA, USA. http://dl.acm.org/citation.cfm?id=3056

Sequeira P, Melo FS, Prada R, Paiva A (2011) Emerging social awareness: exploring intrinsic motivation in multiagent learning. In: 2011 IEEE international conference on development and learning (ICDL), vol 2, pp 1-6. https://doi.org/10.1109/DEVLRN.2011.6037325

Shalev-Shwartz S, Shammah S, Shashua A (2016) Safe, multi-agent, reinforcement learning for autonomous driving. CoRR arxiv: abs/1610.03295,

Shapley LS (1953) Stochastic games. Proc Nat Acad Sci 39(10):1095-1100

Shoham Y, Leyton-Brown K (2008) Multiagent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press, USA

Shoham Y, Powers R, Grenager T (2003) Multi-agent reinforcement learning: a critical survey. Tech. rep

Silva FLD, Taylor ME, Costa AHR (2018) Autonomously reusing knowledge in multiagent reinforcement learning. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18, International joint conferences on artificial intelligence organization, pp 5487-5493. https://doi.org/10.24963/ijcai.2018/774,

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of go with deep neural networks and tree search. Nature 529:484 EP -. https://doi.org/10.1038/nature16961

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D (2018) A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science 362(6419):1140-1144

Singh A, Jain T, Sukhbaatar S (2019) Learning when to communicate at scale in multiagent cooperative and competitive tasks. In: International conference on learning representations. https://openreview.net/forum?id=ry

Son K, Kim D, Kang WJ, Hostallero DE, Yi Y (2019) Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In: International conference on machine learning, pp 5887-5896

Song J, Ren H, Sadigh D, Ermon S (2018) Multi-agent generative adversarial imitation learning. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, Curran Associates, Inc., vol 31, pp 7461-7472. https://proceedings.neurips.cc/paper/2018/file/240c945bb72980130446fc2b40fbb8e0-Paper.pdf

Song Y, Wang J, Lukasiewicz T, Xu Z, Xu M, Ding Z, Wu L (2019) Arena: A general evaluation platform and building toolkit for multi-agent intelligence. CoRR arxiv: abs/1905.08085,

Spooner T, Savani R (2020) Robust market making via adversarial reinforcement learning. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems, pp 2014-2016

Srinivasan S, Lanctot M, Zambaldi V, Perolat J, Tuyls K, Munos R, Bowling M (2018) Actor-critic policy optimization in partially observable multiagent environments. In: Bengio S, Wallach H, Larochelle H , Grauman K , Cesa-Bianchi N , Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 3422-3435. http://papers.nips.cc/paper/7602-actor-critic-policy-optimization-in-partially-observable-multiagent-environments.pdf

Stone P, Veloso M (2000) Multiagent systems: a survey from a machine learning perspective. Auton Robots 8(3):345–383. https://doi.org/10.1023/A:1008942012299

Strouse D, Kleiman-Weiner M, Tenenbaum J, Botvinick M, Schwab DJ (2018) Learning to share and hide intentions using information regularization. In: Bengio S, Wallach H, Larochelle H, Grauman K , Cesa-Bianchi N , Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 10249-10259. http://papers.nips.cc/paper/8227-learning-to-share-and-hide-intentions-using-information-regularization.pdf

Sukhbaatar S, szlam a, Fergus R (2016) Learning multiagent communication with backpropagation. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 2244-2252. http://papers.nips.cc/paper/6398-learning-multiagent-communication-with-backpropagation.pdf

Sukhbaatar S, Kostrikov I, Szlam A, Fergus R (2017) Intrinsic motivation and automatic curricula via asymmetric self-play. CoRR arxiv: abs/1703.05407,

Sunehag P, Lever G, Gruslys A, Czarnecki WM, Zambaldi V, Jaderberg M, Lanctot M, Sonnerat N, Leibo JZ, Tuyls K, Graepel T (2018) Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 2085-2087. http://dl.acm.org/citation.cfm?id=3237383.3238080

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Adaptive computation and machine learning, MIT Press. http://www.worldcat.org/oclc/37293240

Sutton RS, Precup D, Singh S (1999) Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. Artif Intell 112(1):181-211

Svetlik M, Leonetti M, Sinapov J, Shah R, Walker N, Stone P (2017) Automatic curriculum graph generation for reinforcement learning agents. https://aaai.org/ocs/index.php/AAAI/AAAI/7/paper/view/14961

Tacchetti A, Song HF, Mediano PAM, Zambaldi V, Kramár J, Rabinowitz NC, Graepel T, Botvinick M, Battaglia PW (2019) Relational forward models for multi-agent learning. In: International conference on learning representations. https://openreview.net/forum?id=rJIEojAqFm

Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, Aru J, Vicente R (2017) Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE 12(4):1-15. https://doi.org/10.1371/journal.po

Tan M (1993) Multi-agent reinforcement learning: Independent vs. cooperative agents. In: In Proceedings of the tenth international conference on machine learning, Morgan Kaufmann, pp 330-337

Tang H, Hao J, Lv T, Chen Y, Zhang Z, Jia H, Ren C, Zheng Y, Fan C, Wang L (2018) Hierarchical deep multiagent reinforcement learning. CoRR arxiv: abs/1809.09332,

Taylor A, Dusparic I, Cahill V (2013) Transfer learning in multi-agent systems through parallel transfer. In: in Workshop on theoretically grounded transfer learning at the 30th international conference on machine learning (Poster

Taylor ME, Stone P (2009) Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res 10:1633-1685. http://dl.acm.org/citation.cfm?id=1577069.1755839

Tesauro G (2004) Extending q-learning to general adaptive multi-agent systems. In: Thrun S, Saul LK, Schölkopf B (eds) Advances in neural information processing systems 16, MIT Press, pp 871-878. http://papers.nips.cc/paper/2503-extending-q-learning-to-general-adaptive-multi-agent-systems.pdf

Tumer K, Wolpert DH (2004) Collectives and the design of complex systems. Springer, Berlin

Tuyls K, Weiss G (2012) Multiagent learning: basics, challenges, and prospects. AI Mag 33(3):41

Vezhnevets AS, Osindero S, Schaul T, Heess N, Jaderberg M, Silver D, Kavukcuoglu K (2017) FeUdal networks for hierarchical reinforcement learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 3540-3549. http://proceedings.mlr.press/v70/vezhnevets17a.html

Vezhnevets AS, Wu Y, Leblond R, Leibo JZ (2019) Options as responses: grounding behavioural hierarchies in multi-agent RL. CoRR arxiv: abs/1906.01470,

Vinyals O, Ewalds T, Bartunov S, Georgiev P, Vezhnevets AS, Yeo M, Makhzani A, Küttler H, Agapiou J, Schrittwieser J, Quan J, Gaffney S, Petersen S, Simonyan K, Schaul T, van Hasselt H, Silver D, Lill-icrap TP, Calderone K, Keet P, Brunasso A, Lawrence D, Ekermo A, Repp J, Tsing R (2017) Starcraft II: a new challenge for reinforcement learning. CoRR arxiv: abs/1708.04782,

Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P, Oh J, Horgan D, Kroiss M, Danihelka I, Huang A, Sifre L, Cai T, Agapiou JP, Jad-erberg M, Vezhnevets AS, Leblond R, Pohlen T, Dalibard V, Budden D, Sulsky Y, Molloy J, Paine TL, Gulcehre C, Wang Z, Pfaff T, Wu Y, Ring R, Yogatama D, Wünsch D, McKinney K, Smith O, Schaul T, Lillicrap T, Kavukcuoglu K, Hassabis D, Apps C, Silver D (2019) Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature 575(7782):350-354. https://doi.org/ 10.1038/s41586-019-1724-z

Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, Botvin-ick M (2016a) Learning to reinforcement learn. CoRR arxiv: abs/1611.05763,

Wang JX, Hughes E, Fernando C, Czarnecki WM, Duéñez Guzmán EA, Leibo JZ (2019) Evolving intrinsic motivations for altruistic behavior. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, pp 683-692. http://dl.acm.org/citation.cfm? id=3306127.3331756

Wang S, Wan J, Zhang D, Li D, Zhang C (2016b) Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. Comput Netw 101:158- 168. https://doi.org/10.1016/j.comnet.2015.12.017.http://www.sciencedirect.com/science/article/ pii/S1389128615005046, industrial Technologies and Applications for the Internet of Things

Wang T, Dong H, Lesser VR, Zhang C (2020a) ROMA: multi-agent reinforcement learning with emergent roles. CoRR arxiv: abs/2003.08039

Wang T, Wang J, Wu Y, Zhang C (2020b) Influence-based multi-agent exploration. In: International conference on learning representations. https://openreview.net/forum?id=BJgy96EYvr

Wang T, Wang J, Zheng C, Zhang C (2020c) Learning nearly decomposable value functions via communication minimization. In: International conference on learning representations. https://openr

eview.net/forum?id=HJx-3grYDB

Wei E, Luke S (2016) Lenient learning in independent-learner stochastic cooperative games. J Mach Learn Res 17(84):1-42. http://jmlr.org/papers/v17/15-417.html

Wei E, Wicke D, Freelan D, Luke S (2018) Multiagent soft q-learning. https://www.aaai.org/ocs/index.php/SSS/SSS18

Wei Ren, Beard RW, Atkins EM (2005) A survey of consensus problems in multi-agent coordination. In: Proceedings of the 2005, American control conference, 2005., pp 1859-1864 vol. 3. https://doi.org/10.1109/ACC.2005.1

Weiß G (1995) Distributed reinforcement learning. In: Steels L (ed) The biology and technology of intelligent autonomous agents. Springer, Berlin, pp 415-428

Weiss G (ed) (1999) Multiagent systems: a modern approach to distributed artificial intelligence. MIT Press, Cambridge

Wiegand RP (2004) An analysis of cooperative coevolutionary algorithms. PhD thesis, USA, aAI3108645

Wolpert DH, Tumer K (1999) An introduction to collective intelligence. CoRR cs.LG/9908014. http:// arxiv.org/arxiv: abs/cs.LG/9908014

Wu C, Rajeswaran A, Duan Y, Kumar V, Bayen AM, Kakade S, Mordatch I, Abbeel P (2018) Variance reduction for policy gradient with action-dependent factorized baselines. In: International conference on learning representations. https://openreview.net/forum?id=H1tSsb-AW

Yang E, Gu D (2004) Multiagent reinforcement learning for multi-robot systems: a survey. Tech. rep

Yang J, Nakhaei A, Isele D, Fujimura K, Zha H (2020) Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. In: International conference on learning representations. https://openreview.net/forum?id=S11EX04tPr

Yang T, Meng Z, Hao J, Zhang C, Zheng Y (2018a) Bayes-tomop: a fast detection and best response algorithm towards sophisticated opponents. CoRR arxiv: abs/1809.04240,

Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J (2018b) Mean field multi-agent reinforcement learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 5571-5580. http://proceedings.mlr.press/v80/yang18d.html

Yu C, Zhang M, Ren F (2013) Emotional multiagent reinforcement learning in social dilemmas. In: Boella G, Elkind E, Savarimuthu BTR, Dignum F, Purvis MK (eds) PRIMA 2013: principles and practice of multi-agent systems. Springer, Berlin, pp 372-387

Yu H, Shen Z, Leung C, Miao C, Lesser VR (2013) A survey of multi-agent trust management systems. IEEE Access 1:35-50. https://doi.org/10.1109/ACCESS.2013.2259892

Yu L, Song J, Ermon S (2019) Multi-agent adversarial inverse reinforcement learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning,

PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 7194-7201. http://proceedings.mlr.press/v97/yu19e.html

Zhang K, Yang Z, Basar T (2018) Networked multi-agent reinforcement learning in continuous spaces. In: 2018 IEEE conference on decision and control (CDC), pp 2771-2776

Zhang K, Yang Z, Liu H, Zhang T, Basar T (2018) Fully decentralized multi-agent reinforcement learning with networked agents. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 5872-5881. http://proceedings.mlr.press/v80/zhang18n.html

Zhang K, Yang Z, Başar T (2019) Multi-agent reinforcement learning: a selective overview of theories and algorithms. ArXiv arxiv: abs/1911.10635

Zhang W, Bastani O (2019) Mamps: Safe multi-agent reinforcement learning via model predictive shielding. ArXiv arxiv: abs/1910.12639

Zheng Y, Meng Z, Hao J, Zhang Z (2018a) Weighted double deep multiagent reinforcement learning in stochastic cooperative environments. In: Geng X, Kang BH (eds) PRICAI 2018: trends in artificial intelligence. Springer International Publishing, Cham, pp 421-429

Zheng Y, Meng Z, Hao J, Zhang Z, Yang T, Fan C (2018b) A deep bayesian policy reuse approach against non-stationary agents. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 954-964. http://papers.nips.cc/paper/7374-a-deep-bayesian-policy-reuse-approach-against-non-stationary-agents.pdf

Zhu H, Kirley M (2019) Deep multi-agent reinforcement learning in a common-pool resource system. In: 2019 IEEE congress on evolutionary computation (CEC), pp 142-149. https://doi.org/10.1109/CEC.2019.8790001

Zhu Z, Biyik E, Sadigh D (2020) Multi-agent safe planning with gaussian processes. ArXiv arxiv: abs/2008.04452

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and

出版商注记 Springer Nature 对已发表地图中的领土主张保持中立，