A REVIEW OF COOPERATION IN MULTI-AGENT LEARNING

A PREPRINT.

Yali Du
King's College London
yali.du@kcl.ac.uk

Joel Z. Leibo
Google DeepMind
jzl@deepmind.com

Usman Islam
King's College London
usman.islam@kcl.ac.uk

Richard Willis
King's College London
richard.willis@kcl.ac.uk

Peter Sunehag
Google DeepMind
sunehag@deepmind.com

ABSTRACT

Cooperation in multi-agent learning (MAL) is a topic at the intersection of numerous disciplines, including game theory, economics, social sciences, and evolutionary biology. Research in this area aims to understand both how agents can coordinate effectively when goals are aligned and how they may cooperate in settings where gains from working together are possible but possibilities for conflict abound. In this paper we provide an overview of the fundamental concepts, problem settings and algorithms of multi-agent learning. This encompasses reinforcement learning, multi-agent sequential decision-making, challenges associated with multi-agent cooperation, and a comprehensive review of recent progress, along with an evaluation of relevant metrics. Finally we discuss open challenges in the field with the aim of inspiring new avenues for research.

Keywords Cooperative AI · Reinforcement learning · Multi-agent systems · Multi-agent learning

1 Introduction

Cooperative multi-agent learning (MAL) delves into algorithms and strategies that allow multiple agents to learn how to collaborate, adapt, and make decisions in shared environments. As multi-agent systems become increasingly prevalent in our technologically driven world, the importance of ensuring effective and seamless cooperation between agents grows too.

Cooperative MAL naturally intersects with various other fields including economics [Zheng et al., 2021a, Johanson et al., 2022] and evolutionary biology [Jaderberg et al., 2019, Duéñez-Guzmán et al., 2023]. Other concepts from the social sciences also play a large role such as communication, norms, and trust [Hertz et al., 2023]. Game theory provides a robust foundation for understanding strategic interactions between agents including collaborative and non-collaborative decision-making [Shapley, 1953, Littman, 1994]. Its mathematical formalism aligns with economic principles, and is especially useful when agents need to maximise shared utilities or when mechanisms are required to encourage cooperation in settings rife with potential conflicts.

While the broader field of MAL encompasses a wide range of topics, we aim to focus on its collaborative dimension. As the momentum around Cooperative AI grows (e.g. [Dafoe et al., 2020]), it becomes imperative to offer readers a synthesised understanding of the area. The field has two main branches: team-based MAL (covered in Section 4) and mixed-motive MAL (covered in Section 5).

In team-based MAL, it is difficult to learn effectively coordinated joint policies because a single scalar reward signal provides the only feedback available on the activities of all agents on the team. Consider what happens when one agent takes a rewarding action while another agent act unhelpfully. The shared scalar reward cannot distinguish which agent's action was the one responsible for the reward. This makes credit-assignment difficult in this setting [Claus and Boutilier, 1998, Foerster et al., 2018a, Sunehag et al., 2018].

In the mixed-motive setting there are individual rewards, which are easier to learn from. However, such games contain many sub-optimal equilibria, a fact which gives rise to social dilemmas—-i.e. situations where there is tension between individual and collective rationality [Rapoport, 1974]. In MAL, the game-theoretic notion of a social dilemma has been generalised to the spatially/temporally-extended complex behaviour learning setting [Leibo et al., 2017]. This area has seen the development of a vast array of techniques for enabling cooperation that more resembles what is seen in the human world and, therefore, intersects more with social science and evolutionary biology where the emergence of cooperation is an important topic of study [Duéñez-Guzmán et al., 2023]. For convenience, we use the term 'co-player' to describe other agents in team-based and mixed-motive settings, as opposed to 'opponent' in zero-sum settings.

The structure of this paper is outlined as follows. Section 2 presents self-contained fundamental knowledge of multi-agent learning, including single agent and multi-agent RL, game theoretic formulations. Section 4 considers cooperative systems with pure motivation. Section 5 discuss the case where agents have mixed motivation. Section 6 reviews the benchmarks and evaluation metrics. Section 7 concludes with a discussion on challenges and open questions in the field.

2 Background

In this section, we provide the necessary background on reinforcement learning, in both single- and multi-agent settings. Specifically, we provide a brief overview of stochastic or Markov games, which are the most commonly used frameworks for describing the multi-agent learning setting.

2.1 Single-agent Reinforcement learning

Reinforcement Learning (RL) is a standard problem setting in ML consisting of an agent sequentially performing actions in an environment, leading to a change of observed state and a reward at each timestep. We first describe the single-agent setting using the Markov Decision Process (MDP) formalism and then extend this to the partially observable and multi-agent settings.

Definition 1 *A Markov Decision Process is defined as a five-tuple (S, A, P, r, γ) where: S is a set of states, A is a set of actions, $P(s'|s, a)$, the transition function, gives the probability that the environment transitions from state s to state s' given that the agent plays action a, and $r(s, a)$, the reward function, gives the expected reward obtained by the agent for performing action a in state s. $\gamma \in [0, 1]$ is the discount factor.*

Note that the transition and reward functions only depend on the current state and action, with no need for access to the history of the process. This is known as the Markov property. The Markov Decision Process (MDP) is a commonly used framework for modeling the decision-making process of an agent that has complete knowledge of the system state, denoted as s. In this framework, at each time step t, the agent selects an action a_t based on the current state s_t. This action leads to a transition to a new state, s_{t+1}, which follows a probability distribution denoted as $\mathcal{P}\left(\cdot \mid s_t, a_t\right)$. Additionally, the agent receives an immediate reward, denoted as $R\left(s_t, a_t, s_{t+1}\right)$. The primary objective in solving an MDP is to find a policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, which is a mapping from the state space $\mathcal{S}$ to a distribution over the action space $\mathcal{A}$. This policy, denoted as $a_t \sim \pi\left(\cdot \mid s_t\right)$, aims to maximise the expected sum of discounted rewards:

$$\mathbb{E}[\sum_{t\geq 0} \gamma^t R\left(s_t, a_t, s_{t+1}\right) \mid a_t \sim \pi(\cdot \mid s_t), s_0]. \tag{1}$$

In this context, two essential functions are defined under policy π: the state-action function (Q-function) and the value function, expressed as:

$$Q_\pi(s, a) = \mathbb{E}[\sum_{t\geq 0} \gamma^t r(s_t, a_t, s_{t+1}) \mid a_t \sim \pi(\cdot \mid s_t), a_0 = a, s_0 = s], \tag{2}$$

$$V_\pi(s) = \mathbb{E}[\sum_{t>0} \gamma^t r(s_t, a_t, s_{t+1}) \mid a_t \sim \pi(\cdot \mid s_t), s_0 = s]. \tag{3}$$

These functions quantify the expected cumulative rewards when starting from a specific state-action pair (s, a) and state s, respectively. When referring to the functions associated with the optimal policy π^*, they are commonly known as the optimal Q-function and the optimal value function.

In many situations, it is too strong to assume the agent can see all the state all the time. Therefore an adjusted formalism can be employed known as a partially observable MDP (POMDP), defined as a 7-tuple $(S, A, P, R, \Omega, O, \gamma)$ where: S is a set of states, A is a set of actions, and $P(s'|s, a)$ is the transition function giving the probability that the environment transitions from state s to state s' given that the agent plays action a. $r(s, a)$ is the reward function that gives the expected reward obtained by the agent for performing action a in state s. Ω is a set of observations. $O(o|s', a)$, the observation function, gives the probability of observing $o \in \Omega$ given the reached state s' and the action a. $\gamma \in [0, 1]$ is the discount factor. Different from the fully observable setting, at time t, the agent only observes o from the environment with probability $O(o|s', a)$. Partial observations result in reduced sample efficiency since the agent needs more experience to perceive the full variety of possible state. Moreover, observations may be modelled in a noisy way (e.g. taking sensor noise into account), reducing stability and requiring even more experience to gain a reliable idea of the effect of actions on the environment (See Cassandra [1998] for more details on POMDP Model). As we will see, these difficulties carry forward into the multi-agent partially observable case.

RL algorithms generally follow two paradigms: value-iteration and policy-iteration. The former seeks to optimise a parameterised value or action-value function using an objective derived by dynamic programming and setting the policy to maximise the learned function, while the latter uses directly explore the policy space.

Value-based Methods Value-based reinforcement learning techniques aim to calculate an accurate approximation of the state-action value function, specifically the optimal Q-function denoted as Q_{π^*}. The approximate optimal policy can then be determined by selecting the action with the highest Q-function estimate. One well-known value-based algorithm is Q-learning [Watkins and Dayan, 1992]. In Q-learning, the agent maintains an estimate of the Q-value function denoted as $\hat{Q}(s, a)$. When the agent transitions from a state-action pair (s, a) to the next state s', it receives a reward denoted as r and updates the Q-function as follows:

$$\hat{Q}(s, a) \leftarrow (1 - \alpha)\hat{Q}(s, a) + \alpha[r + \gamma \max_{a'} \hat{Q}(s', a')]. \tag{4}$$

Here, $\alpha > 0$ represents the learning rate or step size. Under certain conditions on α, Q-learning can be mathematically proven to converge to the optimal Q-value function almost surely [Watkins and Dayan, 1992, Szepesvári and Littman, 1999], with discrete and finite state and action spaces. Furthermore, when combined with neural networks for function approximation, deep Q-learning has shown remarkable empirical success in achieving human-level control in various applications, as demonstrated by [Mnih et al., 2015, 2013]. Another notable on-policy value-based method is SARSA. Its convergence properties were established in a study by [Singh et al., 2000], particularly in settings with for finite-space settings.

Policy-Based Methods Another category of reinforcement learning (RL) algorithms involves a direct exploration of the policy space, which typically utilise parameterised function approximators, namely approximating the policy by $\pi_\theta(\cdot \mid s)$ where θ denotes unknown parameters. Consequently, the straightforward approach of updating the parameter based on the gradient of long-term rewards has been implemented through the policy gradient (PG) method. The fundamental premise behind this concept is expressed as [Sutton et al., 1999]:

$$\nabla J(\theta) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s), s \sim \eta_{\pi_\theta}(\cdot)} \left[Q_{\pi_\theta}(s, a) \nabla \log \pi_\theta(a \mid s) \right]. \tag{5}$$

Here, $J(\theta)$ represents the expected return, Q_{π_θ} is the Q-function under policy π_θ, $\nabla \log \pi_\theta(a \mid s)$ denotes the policy's score function, and η_{π_θ} signifies the state occupancy measure, which can be either discounted or ergodic, under policy π_θ. Various policy gradient methods, including REINFORCE [Williams, 1992], and actor-critic algorithms [Konda and Tsitsiklis, 1999], have been introduced by estimating the gradient in various ways. This idea also applies to deterministic policies in continuous-action settings, and Silver et al. [2014] derived the policy gradient for such cases. In addition to gradient-based methods, several other policy optimisation techniques have demonstrated excellent performance in numerous applications. These include PPO [Schulman et al., 2017], TRPO [Schulman et al., 2015], and soft actor-critic [Haarnoja et al., 2018].

2.2 Multi-agent learning

Multi-agent learning considers the more general situation with multiple interacting agents. The single-agent RL algorithm may not be applicable as this induces non-stationarity, i.e. the environment itself is changing from the perspective of each agent as the other agents' policies evolve. Accordingly, the problem is much harder and specific algorithms must be employed. Figure 1 provides an taxnomy of multi-agent systems based on its utility structure. Multi-agent systems fall into the following categories: Cooperative, where the agents share a common reward, Competitive, where rewards are individual and sum to zero (i.e. one agent's gain is another's loss), mixed-motive, where rewards are individual and both cooperative and competitive motivations coexist.

Definition 2 *A Markov Game is defined by a set of N agents (or players) and the following: S is a set of states, A_i is a set of actions for agent $i \in \mathbb{N} = \{1, 2, ..., N\}$, $P(s'|s, a)$, the transition function, gives the probability that the*

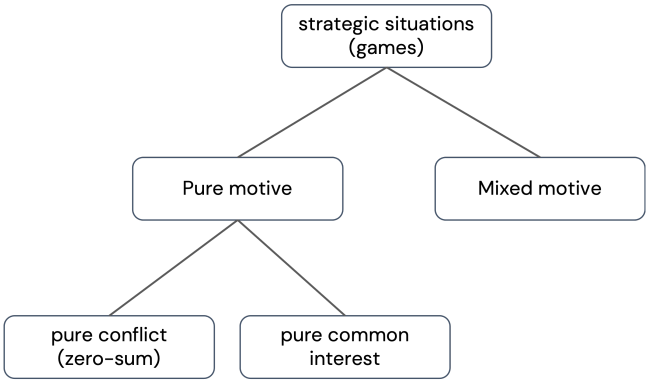

Figure 1: A taxonomy of multi-agent systems [Schelling, 1960].

environment transitions from state s to state s' given that the agents play the joint action $\boldsymbol{a} := (a_1, ..., a_k)$, $R_i(s, \boldsymbol{a})$, the reward function for agent i, gives the expected reward obtained by i after the agents perform the joint action a in state s, $\gamma \in [0, 1]$ is the discount factor.

Along the same spirit with POMDP, in Markov games, it is not always true to assume that the agent can see all the state all the time. Therefore an adjusted formalism can be employed known as a Partially Observable Markov Games (POMG). Players are unable to directly observe each state; rather, they obtain a partial observation of the state, which is defined by the observation function $o_i \in \mathbb{R}^d$, which is determined by an observation function $\mathcal{O} : \mathcal{S} \times \mathbb{N} \to \mathbb{R}^d$. The partially-observed case has been predominant in the research of mixed-motive settings and is sometimes omitted to reduce the clutter in terminology. The framework of Markov games is a general umbrella of various multi-agent learning settings, namely cooperative, competitive and mixed-motive settings.

Cooperative Setting In completely cooperative scenarios, all agents typically work under a unified reward function, symbolised as $R_1 = R_2 = ... = R_N = R$. This model is recognised as multi-agent MDPs [Boutilier, 1996, Lauer and Riedmiller, 2000] or team Markov games [Wang and Sandholm, 2002].

Given this framework, both the value function and Q-function are consistent across all agents. This uniformity allows the use of single-agent RL techniques, like the Q-learning update, when all agents function as a single decision entity. The apex of cooperation in this context aligns with a Nash equilibrium in the game's landscape.

In a variant setting, agents can possess distinct reward functions, possibly confidential to each agent [Zhang et al., 2018]. The cooperative objective is to enhance the enduring reward based on the average reward $\bar{R} = \frac{1}{N} \sum R^i(s, \boldsymbol{a}, s')$ for any given $(s, \boldsymbol{a}, s')$ set in $S \times A \times S$. This model, emphasising agent variation, counts the previously discussed model as a subset. It not only maintains agent privacy but also propels the creation of independent multi-agent RL algorithms [Zhang et al., 2018, Qu et al., 2020]. This heterogenity further demands integrating communication structures into multi-agent reinforcement learning and examining communication-efficient strategies.

Competitive Setting For competitive environments are typically represented as zero-sum Markov games. In essence, the collective rewards of all participants equal zero for every combination of (s, a, s'). The bulk of studies has predominantly centred on dual-agent competing with each other [Littman, 1994] since, in this scenario, one agent's gain directly corresponds to the other's loss. Apart from direct applications in gameplay [Littman, 1994, Silver et al., 2016, OpenAI, 2018], such zero-sum scenarios are also employed to advance robust learning, as the uncertainty that prohibits learning could be modelled as a fictitious opponent that has opposite interest with the robust learning agent. The Nash equilibrium gives rise to a policy that's tailored for optimising rewards in the harshest scenarios.

Mixed Setting The mixed setting, often referred to as the general-sum game framework, doesn't place limitations on the objectives or interactions of agents [Hu and Wellman, 2003, Littman et al., 2001, Littman, 1994]. These agents act based on their individual interests, and their rewards can sometimes be at odds with those of other agents. Foundational game theory concepts like Nash equilibrium [Başar and Olsder, 1998] greatly shape the algorithms tailored for such a setting. Multi-agent problems can also be considered under scenarios combining both fully cooperative and competitive agents within this setting. An example would be two teams in a zero-sum competition, where agents within each team collaborate [OpenAI, 2018, Jaderberg et al., 2019]. In this paper, we focus on the cooperation in both cooperative settings and mixed settings.

3 Challenges in Multi-agent Cooperation

In contrast to single-agent reinforcement learning, where the agent's objective is to efficiently maximise long-term returns, the learning objectives in multi-agent reinforcement learning are sometimes less clearly-defined. The field debates whether MAL is to be understood as a question or as an answer [Shoham et al., 2007]. In fact both motivations exist simultaneously. There are two main lines of research: one focuses on maximising joint values, described as Team Markov Games or Multi-agent MDPs, and the other seeks to find common ground among agents and promote their social welfare or avoid social dilemmas. These scenarios are often modelled through mixed-motive (general sum) games. In Section 3.2, we summarise below the key challenges for multi-agent cooperation in Team Markov Games. While mixed motivation games share a similar set of challenges with Team Markov Games, they also involve new challenges arising from the self-motivated nature of the agents, which are discussed in Section 3.3.

3.1 Shared challenges appearing in both team and mixed-motive settings

Non-stationarity and scalability in the number of agents Multi-agent systems are highly non-stationary, since any agent's policy improvement is experienced by other agents in the system as a change in the distribution of their experience [Foerster et al., 2017, Lowe et al., 2017a, Leibo et al., 2019a]. From each agent's perspective, the environment involves all the other agents and their own interactions.

When you increase the number of agents you also increase the size of the joint action space exponentially. This may lead to exploding training times for algorithms that attempt to model the joint action space directly. A naive way to sidestep this problem is to use independent and decentralized learning algorithms for each agent [Claus and Boutilier, 1998]. However, performance degrades in this case because non-stationarity is not taken into account [Foerster et al., 2017]. Exploration, a key part of any RL scheme, also ostensibly becomes more difficult since the choice of actions to explore now needs to be coordinated in joint action space, rather than each agent naively adding independent noise to its action as one might do in single-agent reinforcement learning [Bard et al., 2020]. On the other hand, Leibo et al. [2019b] argued that the non-stationarities induced by increasing population sizes can actually be a blessing in disguise since under certain conditions a rising number of players may induce a kind of consistently-directed exploration called "exploration by exploitation" which would be hard to motivate otherwise due to the presence of training plateaus and attractive local optima (see also Sunehag et al. [2023]).

Cooperating with Novel Partners (Generalisable Social behaviour) This challenge concerns developing agents that are capable of cooperating with novel individuals who they never encountered during prior training [Rahman et al., 2021, Hu et al., 2020, Leibo et al., 2021, Jaderberg et al., 2019, Carroll et al., 2019]. It is also called the problem of "ad hoc teamwork" [Stone et al., 2010], but that name is somewhat misleading since it implies a team sports metaphor which is too limited in practice to cover much of the work in this area which includes a wide range of efforts to study and improve generalization to novel social situations. Some work in this area is concerned with pure common-interest situations where agents are aligned with strangers on the goal but still must effectively coordinate to achieve it. Other work in this area is concerned with mixed-motive situations where agents meet strangers with whom they may not even be aligned on the overall goal [Leibo et al., 2021, Agapiou et al., 2022]. To be robust to novel agents, an agent needs to be able to engage in reciprocal cooperation with like-minded partners, while remaining robust to exploitation attempts from others. In order to do this, the agents need a broad skill set, including the social skills to determine who they can trust to stick to their agreements or reciprocate, as well as the ability to flexibly adapt when they encounter conventions different to those previously experienced [Dafoe et al., 2020]. They also need flexibility to adapt when they encounter different conventions to those previously experienced. A core challenge when developing these capabilities in an agent is to expose them to sufficiently diverse behaviours during training [Strouse et al., 2021, Madhushani et al., 2023] (see also Balduzzi et al. [2019] for discussion of the analogous need for training diversity in situations of pure conflict).

In this setting, coordination can be difficult to achieve because it is not possible to rely on prior knowledge to anticipate how a novel player will act. One example approach is to develop policies that are effective against a predetermined set of teammate types. An agent then classifies each novel teammate based upon their action history and selects a suitable behavioural policy for their type [Strouse et al., 2021, Lou et al., 2023, Li et al., 2023]. A zero-sum version of this idea appears in the OPRE algorithm of Vezhnevets et al. [2020], which was extended to the mixed-motive setting in Leibo et al. [2021] and Agapiou et al. [2022].

3.2 Team Markov Games

Credit Assignment In single-agent RL, credit assignment is the problem of distinguishing which past actions contributed significantly to the current reward, or equivalently how much the current action will contribute to future rewards, which is sometimes termed temporal credit assignment or reward redistribution [Sutton and Barto, 2018].

Credit assignment is difficult in multi-agent team scenarios where agents strive to optimise a joint reward. The problem stems from the inability of decentralized agents to precisely determine their own contribution to the reward. It may lead to complacency in some agents over time (an issue sometimes called the lazy agent problem [Sunehag et al., 2018]) and ineffective coordination. Moreover, reinforcement learning agents generally have a hard time differentiating between their teammates' exploration behaviour and random environmental factors [Claus and Boutilier, 1998]. Partial observability also exacerbates credit assignment issues since it makes it difficult for an agent to get reliable information about co-player behaviour and how it may impact their shared reward.

While a rich body of research has focused on distributing rewards among team players in dense reward settings [Jeon et al., 2022, Wang et al., 2020a], the area of sparse and delayed rewards in the context of multi-agent reinforcement learning (MARL) remains relatively unexplored. Recent efforts have utilized transformers for global reward decomposition into local rewards [Chen et al., 2023, She et al., 2022]. However, success has been limited to small-scale scenarios and often lacks interpretability. It is particularly noteworthy that understanding the episodic reward setting, where rewards are only observed at the end of an episode, is challenging and not well-understood.

3.3 Mixed-motive games

Heterogeneous incentives In mixed-motive situations there is the additional complication that the agents desire different outcomes from each other. This means that an agent cannot rely on their partners to act in their best interests. For example, a game of football is a team game, each player wants their team to win, and the challenge is to determine and enact the most successful strategy taking into account the capabilities of their teammates and their opponents. If the most successful strategy were known, each teammate would be content to play their corresponding role. Now, suppose there is an additional financial incentive given to the players for each goal that they score. This makes the game a mixed-motive game, as each player now has a preference to win the game, as before, but also to personally score goals [Köster et al., 2020]. Whereas without the financial incentive, a teammate can be trusted to pass the ball if this is more likely to result in a goal, now they cannot be fully relied upon to do so, because they might prefer to be the player who takes the shot.

Here, at least the personal incentives still broadly align with the collective objective, as scoring goals will increase the chance that the team will win. In some mixed-motive games, called social dilemmas, individual incentives actually conflict with the group's objective [Rapoport, 1974]. For example, suppose the unscrupulous manager of the opposing team were to offer a far larger financial incentive to the first player to score an own-goal. Now, what is best for the players (scoring an own-goal before anyone else) conflicts with what is best for their team (winning). Even if the strategy that maximises the probability of the team winning were known, the players would have incentives to deviate from it. In these situations, if each player attempts to do what is best for themselves, this can lead to poor outcomes for the group. Notable examples are free-riders who benefit from public goods without contributing their share [Hughes et al., 2018], and the tragedy of the commons where a common pool resource is degraded by unrestrained consumption [Perolat et al., 2017].

Collective good Consider the case where a user is training a group of agents and is deciding what to use as a performance metric. Suppose that the user wishes to maximise the collective good or social welfare. Due to differences in individual preferences, it can be difficult to understand what is best for the collective. First, there are many possible notions of the collective good, such as the utilitarian, which measures the sum of all individual agent rewards, or Rawls [1971]'s metric, which measures the reward of the worst-off agent. Second, even with a clear understanding of the chosen concept of social welfare, there is also the issue of fairness to take into account. Even though a given outcome may maximise the sum of player rewards, if it results in an unfair distribution of rewards, this may not be desirable. Therefore another metric, such as the Gini coefficient [Ceriani and Verme, 2012] that measure inequality could be used in tandem [Perolat et al., 2017]. Once desirable performance has been specified, it is still non-trivial to achieve it. Due to the misalignment between individual incentives, simply training each agent to optimise their own rewards is unlikely to result in good outcomes [Leibo et al., 2017, McKee et al., 2020]. Additional mechanisms are typically required to achieve cooperation in these settings.

4 Cooperation in Team Games

In the fully cooperative situation, all agents typically work under a unified reward function, symbolised as $R_1 = R_2 = ... = R_N = R$. Due to its relevance to single-agent tasks, it is arguably the most studied domain. In this section, we first provide a high-level explanation of the taxonomy, which is constructed based on the learning paradigm and type of policies, and then of the representative algorithms for team game settings. Table 1 summarises recent popular algorithms for solving common-payoff games.

Table 1: Summary of representative algortihms for MARL with common payoffs. VB and PG represent learning types, with VB standing for Value-based and PG for Policy Gradient.

Algorithm	Centralised/ Decentralised Critic	Centralised/ Decentralised Actors	Learning	Communication
MADDPG [Lowe et al., 2017b]	C	D	PG	✗
COMA [Foerster et al., 2018a]	C	D	PG	✗
LIIR[Du et al., 2019]	C	D	PG	✗
GridNet [Han et al., 2019]	C	C	PG	✓
MA2C [Iqbal and Sha, 2019]	C	D	PG	✗
HAPPO/HATRPO [Kuba et al., 2021]	C	D	PG	✗
MAPPO [Yu et al., 2022]	C	D	PG	✗
VDN [Sunehag et al., 2018]	C	D	VB	✗
QMIX [Rashid et al., 2018]	C	D	VB	✗
QTRAN [Son et al., 2019]	C	D	VB	✗
Qatten [Yang et al., 2020a]	C	D	VB	✗
QPLEX [Wang et al., 2020b]	C	D	VB	✗
SHAQ [Wang et al., 2020a]	C	D	VB	✗
FMA-FQI [Wang et al., 2021a]	C	D	VB	✗
RIAL/DIAL [Foerster et al., 2016]	C	D	VB	✓
CommNet [Sukhbaatar and Fergus, 2016]	C	D	PG	✓
ATOC [Jiang and Lu, 2018]	C	D	PG	✓
TarMAC [Das et al., 2019]	C	D	PG	✓
IC3Net [Singh et al., 2019]	C	D	PG	✓
SchedNet [Kim et al., 2019]	C	D	PG	✓
DCG [Böhmer et al., 2020]	C	D	VB	✓
DGN [Jiang et al., 2020]	C	D	VB	✓
Networked AC [Zhang et al., 2018]	D	D	PG	✓
NeurComm [Chu et al., 2020]	D	D	PG	✓

4.1 Learning Paradigm

In comparison to the single-agent scenario, multi-agent reinforcement learning (MARL) introduces a more complex information structure, which determines who has knowledge during training and execution. For instance, in the context of Markov games, it is sufficient for each agent to observe the current state s to make decisions, as the local policy π maps from states to actions.

Various learning approaches stemming from different information structures result in numerous algorithmic variations. An extreme case is the independent learning scheme, where agents only have visibility of local actions and rewards. This scheme generally encounters convergence issues [Tan, 1993]. Independent learning is a straightforward method that adapts single-agent RL algorithms for multi-agent environments, where each agent operates as a separate learner. In such a framework, the actions of other agents are seen as environmental factors. This concept was initially conceptualised in Tan [1993], where the Q-learning algorithm was adapted for this context, leading to what is known as Independent Q-Learning (IQL). The primary hurdle with IQL lies in its non-stationarity, given that actions from individual agents aiming for local objectives influence environmental transitions.

In order to address the challenge of handling partial information as described earlier, a substantial body of research has operated under the assumption of a central controller. This central entity is responsible for gathering data such as collective actions, shared rewards, and joint observations, and is even design policies for all participating agents. However, it is important to note that in most practical applications, a centralised controller is not readily available, except in cases where access to a simulator is easily attainable, such as in video games and robotics [Peng et al., 2017, Han et al., 2019]

To strike a balance between decentralised control and non-stationarity, the popular learning scheme of centralised-learning-decentralised-execution (CTDE) has emerged. This concept originated from research in planning for the partially observed setting, specifically, Dec-POMDP [Oliehoek et al., 2016]. In CTDE framework, the critic (i.e. value or action-value function neural network) takes all agents' observations and actions into account but each agent's policy only takes its own observation into account. Here the actors are decentralised, leading to fast exploration, but training of the critic incorporates all agents in order to capture the non-stationarity and coordination of behaviours. When the same team rewards are distributed among all agents, a single critic model suffices. However, when rewards are localised and private, each agent should have its critic model for training.

4.2 Solution approaches

Policy-based Methods The first class of algorithms employs the policy-iteration paradigm discussed earlier. The difference is in the structure of the policy/policies across agents.

According to the Policy Gradient Theorem, the objective for policy updates in the single-agent case is

$$g = \mathbb{E}_{\pi}[Q(s,a)\nabla_{\theta} \ln \pi_{\theta}(a|\tau)], \tag{6}$$

where $Q(s, a)$ is the critic. While $\pi_{\theta}(a|\tau)$ can be modelled as the joint policies of individual $\pi_i(a|\tau)$, this approach necessitates a centralised controller, and the research in this pathway usually aim to achieve high scalability to number of actors [Han et al., 2019, Peng et al., 2017]. For instance, GridNet [Han et al., 2019] represents the state information as a grid feature map and employs convolutional neural networks as the policy network to achieve scalable control of an arbitrary number of agents. On the other hand, BicNet [Peng et al., 2017] models the interdependencies of agents through the use of bi-directional RNN. Although it operates in a centralised manner, it offers flexibility in controlling varying numbers of agents. This method is said to employ a centralised actor since there is only one policy and this takes all agents' states into account in order to produce all actions.

For decentralised actors, it can be extended to:

$$g_i = \mathbb{E}_{\pi}[Q(s,\boldsymbol{a})\nabla_{\theta} \ln \pi_i(a_i|\tau_i)], \tag{7}$$

for agent i. To reduce variance, we can subtract from the critic a baseline function which is independent of the action of that agent (this will not change the optimal parameters). This is similar to the single-agent case but here the baseline can incorporate the other agents' actions since the policy is trained in a decentralised way:

$$g_i = \mathbb{E}_{\pi}[(Q(s,\boldsymbol{a}) - b(s,\boldsymbol{a}_{-i}))\nabla_{\theta} \ln \pi_i(a_i|\tau_i)]. \tag{8}$$

Note that we have estimated the action-value function (A.K.A. Q-function) using the reward obtained using the current action and the value function of the next state. Representative algorithms include MAPPO [Yu et al., 2022], HAPPO [Kuba et al., 2021], COMA [Foerster et al., 2018a] and LIIR [Du et al., 2019].

In cases where agents' actions are continuous, deterministic policies are employed for agents. According to single-agent DDPG [Lillicrap et al., 2015], the deterministic policy gradient update is

$$\nabla_{\theta} J(\theta) = \mathbb{E}_s[\nabla_{\theta}\mu_{\theta}(a|s)\nabla_{\theta} Q^{\mu}(s,a)]. \tag{9}$$

Here a deterministic policy μ_{θ} is used instead of π_{θ}. The multi-agent extension for agent i is given as follows

$$\nabla_{\theta_i} J(\mu_i) = \mathbb{E}_{\boldsymbol{x},a}[\nabla_{\theta_i}\mu_i(a_i|s_i)\nabla_{a_i} Q_i^{\mu}(\boldsymbol{x}, a_1, ..., a_N)], \tag{10}$$

where the critic is trained by standard TD learning. This is the basis for Multi-agent DDPG [Lowe et al., 2017b]. Note that each agent has its own critic, meaning that this can model cooperative, competitive or mixed behaviours. We assume the agents know each others actions, but if this is not the case the authors present a method to infer these, by maximising the log probability of other agents' actions with an entropy regulariser. MADDPG also solves the Markov game problem, but the limitation for such algorithm is difficulties in proving convergence.

MAPPO [Yu et al., 2022] is a straightforward extension of PPO to the CTDE multi-agent case, with the caveat that agents must share policy parameters. It solves the Dec-POMDP problem so we assume shared rewards and value functions which take in all the agents' actions as opposed to the policies which are decentralised. Whereas MAPPO requires agents to share parameters, HAPPO and HATRPO [Kuba et al., 2021] overcome this limitation. They are multi-agent extensions of PPO and TRPO respectively, justified theoretically by a multi-agent advantage decomposition result. Namely, the joint advantage function can be decomposed into the sum of agents' local advantages. Since parameter sharing is not necessary, these methods solve the Markov game problem (where competitive and mixed behaviour is possible).

The standard baseline function is the value function $V(s^t)$, leading to the advantage function $Q(s^t, \boldsymbol{a}) - V(s^t) = r + \gamma V(s^{t+1}) - V(s^t)$ being substituted into the policy gradient objective. However, this choice only takes global rewards into account so it does not do a good job of assigning credit for rewards to specific agents. Counterfactual Multi-agent Policy Gradients (COMA) [Foerster et al., 2018a] instead uses a counterfactual baseline, i.e. the Q-function with the given player marginalised out:

$$b(s,\boldsymbol{a}_{-i})) := \sum_{a_i'} \pi_i(a_i'|s_i) Q(s,(a_i',\boldsymbol{a}_{-i})). \tag{11}$$

The adjusted advantage function now compares the Q-value using all agents' actions to one using only the other agents' actions. The marginalisation is done over the current estimate of the agent's policy. Each agent's policy objective now encodes more specific information about the agent's contribution.

In the similar vein, Learning Individual Intrinsic Reward (LIIR) [Du et al., 2019] adds a learned intrinsic reward function $r_i^{in}(s_i, a_i)$ per agent to the standard extrinsic reward r^{ex} from the environment. $r_i^{in}(s_i, a_i)$ is parameterised by η_i and takes in the given agent's state and action. It is trained to maximise the extrinsic (i.e. standard) discounted return J_ex so that it is in line with the overall MARL problem, while using the experience of the specific agent. The proxy reward is defined as $r^{proxy} := r^{ex} + r^{in}$ and the proxy discounted return J^{proxy} is then used to train the policy parameters θ_i. Note that since r^{in} is trained to maximise the extrinsic return, the proxy return preserves the MARL objective for policy updates. However, each agent's return now reflects its own specific experience as well, leading to increased diversity between agents. This is a CTDE method for fully cooperative multi-agent systems and the overall update is done as a bilevel optimisation using meta-gradient.

Value-based Methods In the following we review some popular methods that focus on the critic and learn using CTDE. A centralised critic is one which is shared among the agents and takes in all state and actions. This lacks scalability as it must be trained on the joint action space. On the other hand, a decentralised critic for a given agent only takes in the observations and actions of that agent. Using decentralised critics does not account for the non-stationarity inherent in multi-agent systems. Therefore, a better alternative to these two approaches is value factorisation, where we start with decentralised critics and pass the outputs into a mixing network whose output represents a combined (centralised) critic network. This allows fast training based on only localised experience while also taking agent interactions into account. We would like that the optimiser of the centralised critic Q_{tot} also individually optimises each decentralised critic Q_i. This is known as the Individual-global Maximisation (IGM) constraint:

$$\text{argmax}_{\boldsymbol{a}} Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = (\text{argmax}_{a_1} Q_1(\tau_1, a_1), ..., \text{argmax}_{a_n} Q_n(\tau_n, a_n)). \tag{12}$$

The methods differ in how exactly they factorise the value function. We will consider **linear** and **nonlinear factorisation approaches**. Value Decomposition Network (VDN) [Sunehag et al., 2018] simply sums the decentralised critics to produce the full critic so that:

$$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_i Q_i(\tau_i, a_i). \tag{13}$$

TD learning is then performed on the full critic so that the agents are not trained using their own specific rewards. This method is scalable due to the simple summation and it satisfies the IGM constraint. However, linear factorisations are a limited representation and there is no global convergence guarantee for value-based learning.

Although the IGM constraint is satisfied by VDN, the summation factorisation is a rather strong requirement. QMIX [Rashid et al., 2018] notes that it is sufficient to have a global argmax on Q_{tot} that yields the same result as a set of individual argmax operations on the decentralised critics. Therefore it makes a weaker requirement, namely that Q_{tot} is monotonically increasing with each individual Q_i. In other words, the partial derivative of Q_{tot} with respect to Q_i is non-negative:

$$\frac{\partial Q_{tot}}{\partial Q_i} \geq 0, \forall i. \tag{14}$$

The mixing network is not a summation but a full neural network with the constraint that the weights must be non-negative, thus ensuring the above criterion. This allows for a much more complex representation of the centralised critic while allowing decentralised policies to be extracted using linear-time individual argmax operations (i.e. over individual actions rather than the joint action).

QTRAN [Son et al., 2019] builds on the above two methods but presents an even weaker assumption on the mixing network that is nevertheless sufficient to satisfy the IGM constraint. The assumption requires the non-negativity of an expression involving Q_{tot}, $\sum_i Q_i$ as well as the (state-)value function. The authors note that this property is also necessary under an affine transformation. The end result is a more general mixing network that can model a wider class of MARL problems than VDN and QMIX. Qatten [Yang et al., 2020a] employs a multi-head attention based mixing network for Q_{tot}, allowing the critic to explicitly measure the importance of each individual Q_i to Q_{tot}. FMA-FQI [Wang et al., 2021a] formalise a multi-agent fitted Q-iteration framework to analyse factorised multi-agent Q-learning. Within this framework, they explore linear value factorisation and uncover that multi-agent Q-learning with this straightforward decomposition implicitly achieves a robust counterfactual credit assignment, though it might not converge in certain scenarios.

QPLEX [Wang et al., 2020b] uses a dueling mixing network based on dueling DQN [Wang et al., 2016]. This involves expressing the action-value functions (both joint and for individual agents) in terms of advantage functions and state-value functions. Specifically, we have:

$$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = V_{tot}(\boldsymbol{\tau}) + A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}), \tag{15}$$

$$Q_i(\tau_i, a_i) = V_i(\tau_i) + A_i(\tau_i, a_i)), \tag{16}$$

where $V_{tot}(\boldsymbol{\tau}) = \max_{\boldsymbol{a'}} Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a'})$ (and similarly for the individual value functions). The idea is to move the IGM constraint from the Q-functions to the advantage functions:

$$\text{argmax}_{\boldsymbol{a}} A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = (\text{argmax}_{a_1} A_1(\tau_1, a_1), ..., \text{argmax}_{a_n} A_n(\tau_n, a_n)). \tag{17}$$

The benefit of this is that the constraint can be directly realised by limiting the value of advantage functions. It is known as the advantage-based IGM constraint and this is an equivalent transformation since the state-value terms do not affect the action selection. To obtain the final factorisation, we need to substitute the individual state-values and advantages into the expression for Q_{tot}. Since the state-value function does not take in actions, we can simply set $V_{tot}(\boldsymbol{\tau}) = \sum_i V_i(\tau_i)$. For the advantages, we will need importance weights since it involves the actions:

$$A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_i \lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) A_i(\tau_i, a_i), \tag{18}$$

where the importance weights satisfy $\lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) > 0$. Overall, we then have the full factorisation as:

$$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_i V_i(\tau_i) + \sum_i \lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) A_i(\tau_i, a_i). \tag{19}$$

This is trained end-to-end with the importance weights being learned using multi-head attention.

Several limitations of CTDE are worth noting. First, $Q(s, \boldsymbol{a})$ represents only an estimation of the global return and is trained using value iteration. This potentially limits its scalability to large scale problems. Second, CTDE does not adapt to the specific characteristics or requirements of individual agents. In the following section, we discuss how to distinguish between individual agents' contributions in the global return.

4.3 Communication

A usual premise in decentralised policies is that agents are allowed to communicate with neighbors to a given extent. Such communication has been instrumental in enhancing exploration, maximising reward, and diversifying solutions in intricate optimisation simulations [Barkoczi and Galesic, 2016], as well as in human studies [Lazer and Friedman, 2007]. Section 4.2 has reviewed many non-communicative techniques prioritising stabilised training through centralised value assessment.

One line of work employs fixed communication topology for both training and execution. Works like RIAL/DIAL [Foerster et al., 2016] leverage shared policy fingerprints among agents, while CommNet [Sukhbaatar and Fergus, 2016] utilises distinct communication channels combined with average pooling to assimilate information. Approaches such as DCG [Böhmer et al., 2020], ATOC [Jiang and Lu, 2018] and DGN [Jiang et al., 2020] identify neighboring agents through the K-nearest neighbor mechanism. In contrast, studies like GridNet [Han et al., 2019] employ convolution techniques in their policy networks to indirectly harness neighboring information. Research like CommNet [Sukhbaatar and Fergus, 2016] delves into communication structures among various learning agents and primarily focuses on static topologies. Similarly, Networked MARL methods like MA2C [Chu et al., 2020], Networked Actor-Critic [Zhang et al., 2018] and Gupta et al. [2020] consider scenarios where communication is confined to connected neighborhoods in interconnected systems.

Recently, attention-based communication methods form the third group, focusing on discerning communication recipients. Examples include VAIN [Hoshen, 2017], which adapts the approach from CommNet to introduce an attention vector for agent selection. Techniques like TarMAC [Das et al., 2019], IC3Net [Singh et al., 2019], MAGIC [Niu et al., 2021] and SchedNet [Kim et al., 2019] employ binary attention mechanisms to govern communication instances.

Another special line of work studied decentralised learning with networked agents where agents are allowed to communicate with their neighbors over a prescribed network, such as traffic grid and mobile sensor networks. Zhang et al. [2018] studied actor critic algorithm with guarantee to convergence to a consensus. In consensus algorithms or in the fully observable critic model, it's presupposed that agents can share their observations, actions, or rewards with their counterparts. The aspiration here is that, equipped with information from their peers, they can collectively discern the optimal policy. Zhang et al. [2019] provide a review of recent advances on decentralised training for networked agents.

Table 2: Social Dilemmas

	C	D
C	R,R	S,T
D	T,S	P,P

(a) Social Dilemmas

	C	D
C	$3,3$	$0,4$
D	$4,0$	$1,1$

(b) Prisoner's Dilemma

	C	D
C	$3,3$	$1,4$
D	$4,1$	$0,0$

(c) Chicken

	C	D
C	$4,4$	$0,3$
D	$3,0$	$1,1$

(d) Stag Hunt

4.4 Credit assignment

Credit assignment problem concerns how to distinguish agents contributions to the rewards, while only a shared team rewards available, and agents do not have a sound idea of other agent's contributions. The approaches to credit assignment can vary, whether they are indirect or direct methods. Indirect methods as seen in previous work [Wang et al., 2022, 2020a, Foerster et al., 2018a], where the collective state-action value is seen as a combination of each agent's state-action value, subsequently allocating the communal global rewards based on individual actions. For example, ShapleyQ [Wang et al., 2022, 2020a] present a cooperative game-theoretical framework, extending the Shapley value to Markov games, termed the Shapley Q-value. This Shapley Q-value provide individualised critic for based on each agent's individual contribution, offering a contrast to the shared critic approach. The indirect strategies often grapple with constraints in their representational power, especially in continuous action domains, while Direct methods grapple with discerning the individual impacts of separate agent actions on the collective rewards [Wolpert and Tumer, 2001]. LIIR [Du et al., 2019] firstly introduced intrinsic motivation in multi-agent team games to encourage the diverse behaviour. To regulate the diverse behaviours to be reward-seeking, Du et al. [2019] presented a bilevel programming approach to guide the training of intrinsic reward generator. Sparse and delayed rewards in the multi-agent reinforcement learning (MARL) context remain less explored. Recent attempts leverage transformer to perform global reward decomposition, into local rewards [Chen et al., 2023, She et al., 2022].

4.5 Team-Based Coordination with Novel Partners

The aim of team-based coordination with novel partners, also known as zero-shot coordination, is to construct AI agents that can coordinate in pure common-interest scenarios with novel partners they have not seen before, such as humans or other AI players. While autonomous agents are trained with a given set of partners on completing some tasks, one challenge is to reason about the best way to collaborate with other agents and people without prior coordination, with applications seen in service robots, team sports and autonomous driving. Recent advances were mainly seen in resolving games such as Hanabi [Hu et al., 2020, Bard et al., 2020] and Overcooked [Carroll et al., 2019, Lou et al., 2023], as well as in models of non-verbal communication [Bullard et al., 2020].

Learning algorithms aiming to facilitate zero-shot coordination at test time typically either depend on prior knowledge of environmental symmetries [Hu et al., 2020] or rely on training with a maximally diverse set of co-players [Carroll et al., 2019, Lupu et al., 2021, Zhao et al., 2021]. For instance, in an example of the latter approach, Strouse et al. [2021] operate as an online evolutionary algorithm, continuously adjusting policy parameters and executing policy substitutions within the population to which its learning agents train to best respond. Likewise Lupu et al. [2021] and Zhao et al. [2021] introduce additional auxiliary objective to improve the diversity of the population of partner policies it trains with.

5 Cooperation with Mixed Motivation

This section will focus on the topic of mixed motivation, as shown in Figure 1. Unlike pure common interest, where the focus is on solving practical problems such as sensor networks or a fleet of warehouse robots that jointly achieve higher common payoffs, the emphasis in mixed motivation also includes individual performance.

5.1 Social dilemmas

Central to the study of cooperation arising from mixed motivations is the concept of social dilemma: a situation where there is tension between individual and collective rationality [Rapoport, 1974]. In a social dilemma agents may be conflicted between playing strategies for the good of all players (cooperating) and playing selfish but often individually rational strategies (defecting). More specifically, consider a two-player matrix game with two actions per player, interpreted as a cooperate action, C and a defect action, D, respectively. Table 2a shows three matrix games which are

Table 3: Matrix game social dilemma inequalities

	Inequality	Comment
1	$R > P$	Players prefer mutual cooperation over mutual defection.
2	$R > S$	Players prefer mutual cooperation over unilateral cooperation.
3	$2R > T + S$	Players prefer mutual cooperation over an equal probability of unilateral cooperation and defection.
	At least one of:	
4a	$T > R$	Players prefer unilateral defection to mutual cooperation, which is known as greed.
4b	$P > S$	Players prefer mutual defection to unilateral cooperation, which is known as fear.

commonly considered to be canonical social dilemmas in the literature [Macy and Flache, 2002, Leibo et al., 2017]. There are four relevant payoffs here:

- *Reward* (R) of mutual cooperation.
- *Punishment* (P) arising from mutual defection.
- *Temptation* (T) the outcome for the player who defects while their co-player cooperates.
- S *Sucker* (S) the outcome for the player who cooperates while their co-player defects.

This game is a social dilemma when the payoffs satisfy the matrix game social dilemma inequalities given in Table 3:

While matrix game social dilemmas have been widely applied in social science, economics and biology, they have several shortcomings as models of social dilemmas in real life [Leibo et al., 2017]. First, real-life dilemmas are stateful and temporally extended, instead of being stateless and one-shot. Furthermore, cooperativeness may not be binary here, and an agent could display behaviour on a spectrum of cooperativeness, which may vary over time. This means that players will react to an ongoing pattern of play by other players in a more complicated environment. In addition, the initiation and effects of cooperate or defect behaviours may not occur simultaneously, and their effects as when player 1 starts executing a temporally extended strategy, player 2 may observe this and react accordingly. Players may also only have partial information of the state of the environment. Finally, and most importantly, in complex environments one must learn not just which strategic choice to take as an "atomic" unit but rather must instead learn a whole policy to implement whatever choice they make. The dynamics of learning to implement one's choices also affect the outcomes [Hertz et al., 2023]. Although some of these additions can be modelled with repeated play, continuous action spaces and extended-form games, a natural choice is to use Markov games.

Sequential Social Dilemmas The above issues motivate the idea of a sequential social dilemma (SSD). We now consider Markov games instead of matrix games and policy spaces instead of action spaces. Specifically, we policy sets Π^C and Π^D that implement cooperative and defecting policies respectively. The players can choose policies $\pi^C \in \Pi^C$ or $\pi^D \in \Pi^D$ from these. As for payoffs, notice that this new situation calls for considering the long-term sum of rewards of a particular policy given the other player's policy. We therefore use the expected long-term rewards, defined analogously to the payoffs in the previous section. First, we define player i's value function:

$$V_i^{\pi}(s_0) = \mathbb{E}_{a_t \sim \pi(O(s_t)), s_{t+1} \sim T(s_t, a_t)} \sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t), \tag{20}$$

where $\pi = (\pi_1, \pi_2)$ and a_t is the joint action at time t. Now we can define R, P, T and S as functions of the state $s \in S$:

$$R(s) := V_1^{\pi^C, \pi^C}(s) = V_2^{\pi^C, \pi^C}(s), \tag{21a}$$

$$P(s) := V_1^{\pi^D, \pi^D}(s) = V_2^{\pi^D, \pi^D}(s), \tag{21b}$$

$$T(s) := V_1^{\pi^D, \pi^C}(s) = V_2^{\pi^C, \pi^D}(s), \tag{21c}$$

$$S(s) := V_1^{\pi^C, \pi^D}(s) = V_2^{\pi^D, \pi^C}(s). \tag{21d}$$

The game is now a sequential social dilemma if the social dilemma inequalities (Table 3) hold for $R(s)$, $P(s)$, $T(s)$ and $S(s)$. In practice, the expected long-term payoffs are estimated by fixing the policies of the agents and averaging over multiple simulations, a technique known as empirical game-theoretic analysis [Walsh et al., 2002, Wellman, 2006, Tuyls et al., 2020, Viqueira et al., 2020]. We now present a formal definition.

Definition 3 (Sequential Social Dilemma [Leibo et al., 2017]) *A sequential social dilemma is a tuple $(\mathcal{M}, \Pi^C, \Pi^D)$. $\mathcal{M}$ = a Markov game. Π^C = set of cooperative policies. Π^D = set of defecting policies. Consider the empirical payoff matrix $(R(s), P(s), S(s), T(s))$, induced by policies $(\pi^C \in \Pi^C, \pi^D \in \Pi^D)$ via the payoffs defined above. $(\mathcal{M}, \Pi^C, \Pi^D)$ is an SSD if its empirical payoff matrix $(R(s), P(s), S(s), T(s))$ satisfies the social dilemma inequalities in Table 3.*

The generalisation to more than two players is as follows. An N-player sequential social dilemma is a tuple $(\mathcal{M}, \Pi = \Pi_c \sqcup \Pi_d)$ of a Markov game and two disjoint sets of policies, said to implement cooperation and defection respectively, satisfying the following properties. Consider the strategy profile $(\pi_c^1, \dots, \pi_c^\ell, \pi_d^1, \dots, \pi_d^m) \in \Pi_c^\ell \times \Pi_d^m$ with $\ell + m = N$. We shall denote the average payoff for the cooperating policies by $R_c(\ell)$ and for the defecting policies by $R_d(\ell)$.

We say that $(\mathcal{M}, \Pi)$ is a sequential social dilemma iff the following hold:

1. Mutual cooperation is preferred over mutual defection: $R_c(N) > R_d(0)$.
2. Mutual cooperation is preferred to being exploited by defectors: $R_c(N) > R_c(0)$.
3. Either the fear property, the greed property, or both:
   - Fear: mutual defection is preferred to being exploited. $R_d(i) > R_c(i)$ for sufficiently small i.
   - Greed: exploiting a cooperator is preferred to mutual cooperation. $R_d(i) > R_c(i)$ for sufficiently large i.

A sequential social dilemma is intertemporal if the choice to defect is optimal in the short-term. More precisely, consider an individual i and an arbitrary set of policies for the rest of the group. Given a starting state, for all k sufficiently small, the policy $\pi_k^i \in \Pi$ with maximum return in the next k steps is a defecting policy. There is thus a tension between short-term individual incentives and the long-term collective interest.

Schelling diagrams A Schelling diagram is a game representation which highlights interdependencies between agents, showing how the choices of others shape one's own incentives [Schelling, 1973]. It represents games with any number of players $N \geq 2$, and assumes that each individual faces a binary choice. Thus it is said to be a model of binary choice with externalities. We refer to one choice as cooperation (C), and the other as defection (D). In a game governed by a Schelling diagram, the payoffs from choosing either option are driven only by one's own choice and the number of other individuals that chose the C option. This means the effect of one's choice is influenced by the cumulative effect of externalities arising from the choices of others.

A Schelling diagram plots the curves $R_c(\ell + 1)$ and $R_d(\ell)$, as shown in Figure 2. Intuitively, the diagram displays the two possible payoffs to the N^{th} player given that ℓ of the remaining players elect to cooperate and the rest defect. The minimum number of players who must cooperate in order for each player cooperating to do better than a defector does when all players defect is called the minimum viable coalition size. A social dilemma with a larger minimum viable coalition size generally requires more coordination to resolve than a social dilemma with a smaller minimum viable coalition size.

Asymmetrical dilemmas In reality, agents typically have individual differences in their reward structures, affordances (actions), and sensors (observations), which lead to asymmetric interactions. For example, in ecology [Sunehag et al., 2019] and economics [Zheng et al., 2021b, Johanson et al., 2022], agents are seen as belonging to different species or as playing different social roles. Such role-dependent differences in capabilities and tastes create opportunities for cooperation that differ fundamentally from symmetric scenarios. In particular, they make it possible to realize overall welfare gains from trade. The formal definitions above apply to symmetrical games, where the rewards to a player depend only on their own policy, and the combination of policies chosen by their co-players. An attempt to generalise the definition of a social dilemma has been made by Willis et al. [2023].

5.2 Cooperation in sequential social dilemmas

Numerous methods have been proposed to develop agents who exhibit good cooperation in sequential social dilemmas. These methods have been introduced to achieve better group performance, often measured as the sum of reward for all agents. In this section, we refer to the social welfare, SW, as the mean episode reward for every agent:

$$SW(\boldsymbol{R}) = \frac{1}{n} \sum_i R_i. \tag{22}$$

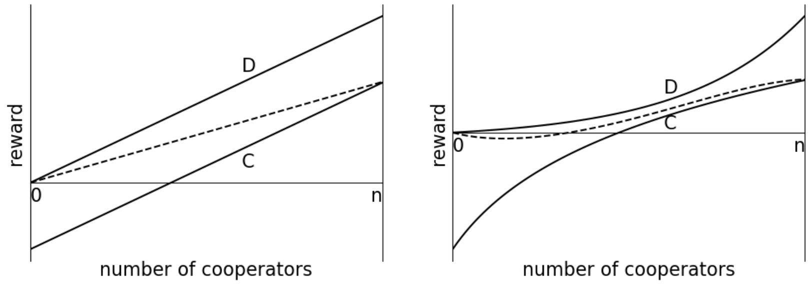

Figure 2: For a population of size n, the Schelling diagram shows the payoff for an $n+1$ agent choosing to either defect or cooperate. The dotted line shows the average reward of the population.

We discuss alternative notions of collective good in Section 6.2.

A common approach to induce cooperation is to use reward shaping, whereby an agent is provided with intrinsic reward in addition to its extrinsic game reward. These additional rewards serve to motivate particular behaviours, for example by including preferences for increasing social welfare. In this paper, we will refer to the specific case of reward shaping involving only the extrinsic rewards of other agents as intrinsic motivation, because it represents the agents having social motivations beyond their own selfish reward. Specifically, let $r^{\text{ex},i}$ denote extrinsic environment reward for agent i and r^{in} denote the intrinsic motivation, agent i's immediate reward is modified as follow,

$$r^i := r^{\text{ex},i} + r^{\text{in},i}. \tag{23}$$

Altruism A particularly common form of intrinsic motivation is called Altruism. While in colloquial use, altruism begets notions of self-sacrifice, in this field it refers to an agent caring about the wellbeing of others, but not necessarily at their own expense. For example, the model of altruism proposed by Chen and Kempe [2008] is as follows. An agent receives a reward that is composed of a proportion of their extrinsic game reward, $p_i(s)$, and a proportion of the mean group social welfare, $SW(\boldsymbol{r})$. A parameter $\beta \in [0,1]$ controls the tradeoff between the two.

$$\forall i \in N, \quad r^i(\beta) = (1-\beta) r^{\text{ex},i} + \frac{\beta}{n} SW(\boldsymbol{r}^{\text{ex}}). \tag{24}$$

If agents care about the collective return, this directly addresses the core problem of a social dilemma - that there are opportunities to profit at the expense of the group. Now, though an agent maintains their own interests, if they take actions that harm the collective, this too will lower their own reward. Additional models of altruism are discussed by Apt and Schäfer [2014].

Game extensions Another popular method is to extend the game to include additional mechanisms, which agents can utilise with an expanded action space. An example of this is contracting [Hughes et al., 2020, Christoffersen et al., 2023]. In one version of this idea, at each timestep, a pair of agents can propose a joint-action. If both agents propose the same joint-action, then they must take their corresponding action, otherwise both agents choose their actions as normal. Other extensions include methods where agents explicitly affect the rewards of other agents through dedicated means, such as by gifting [Lupu and Precup, 2020] or exchanging reward [Willis and Luck, 2023].

Models of the emergence of cooperation Much of the literature in this area has been concerned with the question of how cooperation of self-interested agents may come about and remain stable, despite the ever-present opportunities for conflict, overconsumption, free-riding, and defection that threaten it. This line of work is concerned with modeling humans, and as such is intended to connect with long-established theories in economics and other social sciences [Hertz et al., 2023] which have been interested in self-interested conceptions of agent motivation (in part as a result of adherence to the methodological individualism used in these fields [Heath, 2020]). Moreover, it is preferable to exhibit a simpler account of cooperation over a more a complex one on grounds of parsimony. A common refrain in this area is to say that one is "searching for the minimal set of maximally general priors".

Table 4: Summary of representative algorithms for resolving SSDs

Types	Description	Algorithms
Other-regarding preferences	Agents either intrinsically care about the welfare of others, or they modify the extrinsic game rewards of other agents	Prosocial [Peysakhovich and Lerer, 2018a] Inequity aversion [Hughes et al., 2018] Evolving motivations [Wang et al., 2019a] PED-DQN [Hostallero et al., 2020] SVO [McKee et al., 2020] Gifting [Lupu and Precup, 2020] Gifting for prosociality [Wang et al., 2021b] Relational networks [Haeri et al., 2022] D3C [Gemp et al., 2022] Auto-aligning incentives [Kwon et al., 2023]
Other-influence	Agents consider how their actions will impact the future behaviour of their fixed co-players	Hierarchical social agency [Kleiman-Weiner et al., 2016] CCC [Peysakhovich and Lerer, 2018b] LOLA [Foerster et al., 2018b] Imitation [Eccles et al., 2019] Cooperation degree [Wang et al., 2019b] SOS [Letcher et al., 2019] Social influence [Jaques et al., 2019] LIO [Yang et al., 2020b] Cooperative learning [Jacq et al., 2020] M-FOS [Lu et al., 2022]
Reputation and Norms	Agents assess whether their co-players comply with social rules and modify their behaviour accordingly	Competitive altruism [McKee et al., 2021] Reputation dynamics [Anastassacos et al., 2021] CNM [Vinitsky et al., 2023]
Contracts	Agents are able to commit to taking joint-actions or promises of future rewards	Contracts [Hughes et al., 2020] RUSP [Baker, 2020] Contracts with payments [Christoffersen et al., 2023] Reward exchange [Willis and Luck, 2023] Reward shares [Schmid et al., 2023]

5.3 Solution Approaches

We now review some popular classes of approaches for tackling SSD problems. Table 4 summarises the related algorithms, where we have classified them based upon the mechanism that they use.

Other-regarding preferences When an intrinsic reward is proportional to the mean collective reward [Peysakhovich and Lerer, 2018a], it is regarded as a prosocial reward, and it is equivalent to altruism. This same collective reward was used by Hostallero et al. [2020], though they constrained the intrinsic reward to use only those of other agents within an agent's observation. The authors argue that by limiting the intrinsic motivation to include the reward of local agents, this can help with the credit assignment problem (introduced in Section 4.4).

Hughes et al. [2018] split the typical altruism reward into two parts: advantageous inequity or guilt when an agent outperforms the mean, and disadvantageous inequity or envy when an agent underperforms the means. The intrinsic motivation for agent i's is given as

$$r^{\text{in},i} := -\frac{\alpha}{N-1}\sum_{j\neq i}\max(e_t^j(s_t^j,a_t^j)-e_t^i(s_t^i,a_t^i),0)-\frac{\beta}{N-1}\sum_{j\neq i}\max(e_t^i(s_t^i,a_t^i)-e_t^j(s_t^j,a_t^j),0). \tag{25}$$

Therefore, the immediate reward for agent i is given as $r^i := r^{\text{ex},i} + r^{\text{in},i}$, where $r^{\text{ex},i}$ represents the extrinsic reward. $e_t^j(s_t^j,a_t^j)$ for agents $j = 1, ..., N$ are their extrinsic rewards smoothed in a manner analogous to eligibility traces [Sutton and Barto, 2018]:

$$e_t^j(s_t^j,a_t^j) := \gamma\lambda e_{t-1}^j(s_{t-1}^j,a_{t-1}^j) + r^{\text{ex},j}(s_t^j,a_t^j), \tag{26}$$

for discount factor γ and hyperparameter λ. α and β are hyperparameters which respectively control the importance of envy and guilt. The authors experimentally found that in different social dilemmas, different proportions of envy and guilt performed best. While Hughes et al. [2018] determined these hyperparameters with a grid search, Wang et al. [2019a] determined them using a technique called Population Based Training [Jaderberg et al., 2017] which they interpreted as a model of evolution operating on the reward function. In the variant where cooperation emerged it was specifically a model of multi-level (group) selection [Duéñez-Guzmán et al., 2023], this was an expected result since it has been known since Darwin that individual fitness maximization does not on-its-own lead to the evolution of altruistic traits [Nowak, 2006].

A general method for constructing different attitudes or preferences an agent may have for the relationship between their own reward and the mean group reward is is Social Value Orientation (SVO) [McKee et al., 2020], which takes inspiration from interdependence theory. The social values of an agent can be specified using a point on a circle centred at the origin where the x-axis is the agent's reward and the $y-axis$ is the arithmetic mean $\bar{r}_{-i}$ of all other agents' rewards. Given reward vector $\mathbf{r}$, the reward angle $\theta^i(\mathbf{r})$ for agent i satisfies

$$\tan\theta^i(\mathbf{r}) = \frac{\bar{r}^{-i}}{r^i} \tag{27}$$

and an agent's preferred reward angle θ^i_{SVO} is its SVO. The reward is then augmented with an intrinsic reward penalising deviation of the reward angle (given observed rewards) from the SVO:

$$R^i(s^i_t, a^i_t) := r^i(s^i_t, a^i_t) - w|\theta^i_{SVO} - \theta^i(\mathbf{r})|. \tag{28}$$

This is depicted in Figure 3. The authors demonstrate the positive effects of heterogeneity on achieving complex behavioural variation in a manner consistent with interdependence theory. The heterogeneous agents ultimately outperform homogeneous populations on the SSDs considered. Somewhat analogously, the final method [Haeri et al., 2022] uses pre-defined social structures, represented with a graph, and qualitatively assesses the performance of these relationships in the presence of agent imbalance.

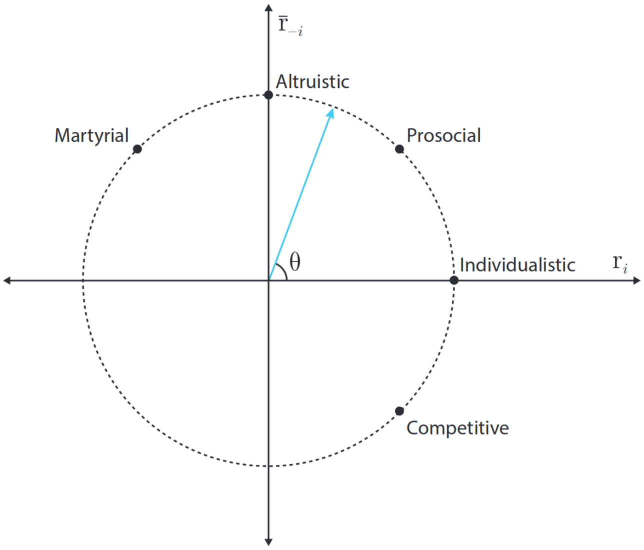

Figure 3: An agent's Social Value Orientation can be expressed by the parameter θ, which specifies the ideal relationship between an agent's own reward and that of the other agents

Rather than having intrinsic motivations, which relies upon agents adopting preferences beyond their own self-interest, by transferring reward an agent is able to modify the extrinsic game reward of their co-players. If this modification leads to a more cooperative approach in the recipients, the agent transferring rewards can experience a net benefit. The most simple method is gifting [Lupu and Precup, 2020, Wang et al., 2021a], where agents are able as part of their actions to give rewards directly to other agents, in a zero-sum manner. This mechanism allows agents to reward their peers for cooperation. Gemp et al. [2022] introduce D3C which is also used by Kwon et al. [2023]. This algorithm takes a collective approach, and finds an optimal mixture of reward transfer or loss sharing to improve the worst-case performance of the group.

Other-influence An agent can strategically encourage cooperative behaviour by mirroring the prosocial behaviour of others. In this way, an agent remains robust to exploitation while offering to engage in reciprocal cooperation with an co-player, as inspired by the game theory strategy for iterated normal-form social dilemmas called Tit-For-Tat [Axelrod, 1980]. Whereas altruism incentivises an agent to take prosocial actions regardless of the behaviour of other agents, here prosocial actions are dependent on those of other agents. A common approach is to train two policies, a cooperative, prosocial policy, and a self-interested policy. If the co-player is deemed to be behaving pro-socially, typically as assessed by a classifier [Wang et al., 2019b, Kleiman-Weiner et al., 2016] or by analysis of one's own rewards [Peysakhovich and Lerer, 2018b], then the agent uses the prosocial policy, otherwise it uses the self-interested policy. In the case of Wang et al. [2019b], a full range of cooperative behaviours is achieved by synthesising the two policies to match the degree of cooperativeness exhibited by the co-player.

In Eccles et al. [2019], the authors use two types of agents; innovators and imitators. Innovators are standard learners whose rewards are purely extrinsic (selfish). Imitators will include an intrinsic reward term which seeks to minimise the difference in niceness between themselves and an innovator, thus mimicking the cooperativeness of the innovator. Specifically, suppose we have an imitator (im) and an innovator (in) and associated trajectories T^{im} and T^{in}. The imitator's intrinsic reward is defined as

$$r^{\text{im}}(t) := -(N^{\text{im}}(T^{\text{im}}) - N^{\text{in}}(T^{\text{in}}))^2, \tag{29}$$

where the niceness function $N^i(T^i)$ for agent $i \in \{\text{im}, \text{in}\}$ measures the effect on an agent's reward of the other agent's actions through a trajectory. Formally, define first the niceness of an innovator action for the imitator as

$$n^{\text{im}}_{\pi_{\text{in}}}(s^{\text{in}}_t, a^{\text{in}}_t) := Q^{\text{im}}_{\pi_{\text{in}}}(s^{\text{in}}_t, a^{\text{in}}_t) - V^{\text{im}}_{\pi_{\text{in}}}(s^{\text{in}}_t), \tag{30}$$

where $V^{\text{im}}_{\pi_{\text{in}}}$ and $Q^{\text{im}}_{\pi_{\text{in}}}$ are neural network models for the value and the action-value functions respectively. The niceness function of an imitator action is defined equivalently.

These functions quantify the effect of the other agent's action on one's own reward. Note that the action-value functions are estimating the discounted return to the agent from time t given the other agent's states and actions. The niceness function of a trajectory for agent $i \in \{\text{im}, \text{in}\}$ can then be defined w.r.t. agent $j \in \{\text{im}, \text{in}\}$
i as the discounted sum of action niceness values

$$N^i(T^i) := \sum_{k=1}^{t} \gamma^{t-k} n^j_{\pi_i}(s^i_k, a^i_k), \tag{31}$$

where t is the length of the trajectory T^i. This niceness function can be used to calculate the intrinsic reward, which is normalised before being added to the extrinsic reward.

An independent algorithm that treats other agents as part of the environment is called a naive learner. Because the environment is assumed to be static, naive learners do not appreciate that their co-players may change their policies, and so not do take into account how changes in their own policy can cause responses in their co-players. The following methods take into account the learning of their co-players, and are called co-player shaping. Co-player shaping is a rare departure from the usual intrinsic reward mechanism, because the other agents' policy parameters are used directly in the value function optimisation stage. Unlike reward transfer methods, where an agent will directly modify the extrinsic rewards of their co-player and leave them to learn how to optimise it, co-player shaping directly considers the policy updates their co-player will make. The during training, an agent using this method attempts to ensure that their co-player will learn to take actions that benefit the agent using co-player shaping.

The first method to introduce this technique was LOLA [Foerster et al., 2018b], which assumed it was facing a naive learner. Instead of optimising the standard value function $V(\theta_1, \theta_2)$, where θ_1 and θ_2 are the policy parameters of the current agent 1 and those of the other agent respectively, this method optimises $V(\theta_1, \theta_2 + \Delta\theta_2)$, where $\Delta\theta_2$ is one naive update of agent 2's parameters. The other agent's parameters are inferred from its state-action trajectories through a form of co-player modelling similar to behaviour cloning. This method is able to achieve cooperation via the following approach: when an co-player takes the action that you want them to, reciprocate so that they receive higher reward and are more likely to take this action next time. Otherwise, punish them to dissuade other actions.

Letcher et al. [2019] and Lu et al. [2022] improved upon LOLA, increasing its robustness and relaxing requirements to have a differentiable model of their co-player. Jacq et al. [2020] extended these concepts to include arbitrary learning algorithms for n-player games, and developed an approach whereby a group of agents can retaliate against any selfish behaviour.

A similar method, Learning to Incentivise Others (LIO) [Yang et al., 2020b] extends co-player shaping to include reward transfers to the co-player, which allows an agent to shape the behaviour of their co-player directly, without needing to influence the co-player's 's extrinsic rewards through the consequences of their own actions alone. Here, a reward giver agent i learns an incentive function $r^{\eta^i} : O \times A^{-i} \to \mathbb{R}^{N-1}$ where O is its own observation space, A^{-i} is the joint action space of all other agents and η^i is a parameter vector. The incentive function determines the rewards given by i to the other agents based on its observation of the state and their actions. Consequently, a recipient agent j has a new reward r^j. Let $r^j_{\eta^i}$ be the reward given by i to j. Then j's new reward is

$$r^j(s_t, \boldsymbol{a_t}, \eta^{-j}) := r^{\text{ex},j}(s_t, \boldsymbol{a_t}) + \sum_{i \neq j} r^j_{\eta^i}(o^i_t, \boldsymbol{a_t}^{-i}). \tag{32}$$

While i optimises its incentive function r_{η^i} to maximise a separate objective, r^j is then optimised as usual with policy gradients. The latter consists of a positive term representing i's expected extrinsic reward due to the reward recipients' new actions as well as a negative term representing the discounted sum of rewards given (i.e. a cost for reward giving). This approach facilitates gifting with gifted rewards determined by a dedicated selfish objective, thereby encouraging cooperation to emerge from self-interest. Notice that the policy updates require knowledge of other agents' parameters. The authors relax this requirement with co-player modelling using other agents' observed states and actions.

Another method uses the idea that agents may seek to maximize their social influence [Jaques et al., 2019], i.e. the extent to which their action changes the decisions of others. Mathematically, the full reward is $r^i := \alpha r^{\text{ex},i} + \beta r^{\text{in},i}$

where $r^{\text{in},i}$ is calculated as the sum over other agents $j \neq i$ of KL-divergences between j's policy conditioned on i's action and j's marginal policy (independent of i's action):

$$r^i := \sum_{j \neq i} D_{KL}[p(a_t^j | a_t^i, s_t^j) || p(a_t^j | s_t^j)]. \tag{33}$$

The quantity $p(a_t^j | s_t^j)$ is estimated as $\sum_{\tilde{a}_t^i} p(a_t^j | \tilde{a}_t^i, s_t^j) p(\tilde{a}_t^i | s_t^j)$, where $\tilde{a}_t^i$ are counterfactual samples of k's action. Agent i's intrinsic reward can be seen as a measure of i's social influence and adding this to the reward encourages choosing more socially influential actions, thus leading to increased cooperation. In a second level of their method, the authors introduce an explicit communication channel. Namely, each agent outputs an extra vector m_t^i representing a message. All the agents' messages are concatenated and then provided as input to each agent in the next step. The intrinsic reward in this case is similar to the above but instead of conditioning agent j's distribution on i's action a_t^i, it is conditioned on i's message m_{t-1}^i from the previous step. This has a similar effect of representing social influence but this time assuming that influence travels through the dedicated communication channel. Empirical results show improved coordination, communication and collective return due to social influence.

Reputation and norms A norm is a collective pattern of behaviour supported by a shared pattern of sanctioning. Some work has taken inspiration from how humans use their own reputation within the group as motivation to cooperate. One such method is McKee et al. [2021], where the reward is as before a sum of extrinsic reward and an intrinsic reward, calculated as

$$r^{\text{in},i} := -\alpha \max(\bar{c} - c^{\text{self}}, 0) - \beta \max(c^{\text{self}} - \bar{c}, 0), \tag{34}$$

where c^{self} is a measure of one's own contribution level and $\bar{c}$ is the observed or estimated average group contribution level. This equation is analogous to inequity aversion described in Eq. (25) (but with contribution levels instead of smoothed rewards) and accordingly α and β are scalar parameters. The terms in this intrinsic reward penalise large discrepancies between one's own contribution level and that of the group as a whole. The authors build a computational model of human behaviour and show that humans can effectively cooperate on the Cleanup task when other players are identifiable and their reputations can be tracked. However, they fail to cooperate under conditions of anonymity. The MARL agents also demonstrate these behaviours with regard to identifiability and anonymity. Anastassacos et al. [2021] assess which reputation norms lead to the emergence of cooperation in a population of reinforcement learning agents playing matrix game social dilemmas.

In the Classifier Norm Model (CNM) [Vinitsky et al., 2023], agents learn social norms through public sanctioning. This is inspired by human societies, where people adjust their behaviour based on their internal understanding of the behaviours to which society as a whole gives its approval or disapproval. In this model, agents have opportunities to sanction one another using a grounded in-game mechanism, whereby they may zap each other with a punishment beam that causes negative reward in any player hit by it. Agent A has an opportunity to sanction agent B whenever agent B is in range of their zapper. Whenever zapping events occur it is assumed that everyone gossips about them, so there is public knowledge of sanctioning opportunities and whether or not they led to a zap, which is interpreted as disapproval of the agent who was zapped's recent behaviour, or did not lead to a zap, in which case it would be interpreted as approval of the not-zapped agent's recent behaviour. Each agent learns a personal representation of what behaviour is acceptable to its group. This takes the form of a classifier of whether or not a given behaviour is likely to provoke approval or disapproval. The classifier is trained with all the group's public sanctioning data. Finally, agents are assumed to have an intrinsic motivation to align their own sanctioning behaviour (zapping) with the predictions of the classifier they learned (i.e. with their personal representation of social norm). They are motivated to zap the in the same context that others in their group would also zap. Therefore, this system exhibits a runaway bandwagon effect where most agents then end up supporting a particular norm, i.e. sanctioning the same behaviour as one another and complying with the norm by refraining from emitting the sanctioned behaviour. This is a model where the content of a social norm proscribing a behaviour emerges from the dynamics of multi-agent learning.

The CNM model is in line with a decentralised vision for mixed-motive agents. For example, the agents could be seen like models of autonomous vehicles learning the local norms of a new town by observing sanctioning events such as other cars honking their horns at one another. Humans may also participate in this using the same interface for sanctioning as the agents (e.g. humans already understand the social meaning of honking a car horn).

Contracts A separate paradigm for encouraging cooperation involves allowing agents to more directly affect other agents' rewards through dedicated structures rather than indirectly through the environment. For example, contracts [Hughes et al., 2020] can be used for this purpose. A contract is a joint state-action dependent vector of rewards which can be proposed and accepted by agents. The authors formulate an augmented game where agents' actions are augmented with the ability to propose contracts, and if both players propose the same joint-action, it becomes binding for the next timestep.

Christoffersen et al. [2023] extended these contracts to include a possible side payment. This approach gives agents the ability to share rewards with other accepting agents, given that the accepting agents follow action trajectories satisfying given specifications. This can be thought of as committing to reward transfer conditional on behaviour. For example, a contract could be interpreted as "I'll pay you to clean the river (so that I can pick apples)".

Rather than specifying contracts on an action-level, which can be burdensome, agents could alternatively enter an agreement that covers the whole episode. Schmid et al. [2023] allow agents to buy and sell stakes in the future rewards of their co-players, in an analogous manner to buying shares in a company, while Willis and Luck [2023] find the minimum proportion of reward that agents must commit to exchanging between themselves over an episode to learn cooperative policies in a sequential social dilemma. In Willis et al. [2023], the authors generalise their method to allow unequal transfers between agents and find transfer arrangements that resolve matrix game dilemmas while retaining the greatest possible proportion of their own extrinisic reward or self-interest. Baker [2020] uses a sampling approach and randomises the transfers (which can be thought of as relationships) between the agents. Furthermore, the agents are uncertain about these relationships. Their experiments suggest that both of these features help the learning algorithms to develop reciprocal behaviours.

6 Evaluation

The evaluation of MAL algorithms can be challenging, it has been noticed that there is a substantial variability in reported results even for the same algorithm on the same task [Gorsane et al., 2022]. The community calls for standardised evaluation protocols, focusing on default parameters such as training time, standardised uncertainty quantification and more complete reporting on failure cases. Below, we provide a review of representative tasks and evaluation metrics.

6.1 Environments

Aligning with the methodology described in Section 4 and 5, we describe the benchmarks for fully cooperative team-based tasks and mixed-motive tasks respectively. Below are open-sourced simulators and benchmarks for fully cooperative tasks.

- **StarCraft Multi-agent Challenge (SMAC):** StarCraft II [Vinyals et al., 2017] is a popular testbed for MARL algorithms, presenting various battle scenarios where agents correspond to units (i.e. combatants) who must cooperate to defeat computer controlled opponents. SMAC has become a standard evaluation suite for MARL, much like Atari in the single-agent case.
- **Overcooked:** Overcooked [Carroll et al., 2020] is a video game environment where agents must cooperate to prepare and serve onion soup. They must split cooking/serving responsibilities among each other and learn to work with varied teammates.
- **MuJoCo Soccer Environment:** The MuJoCo Soccer environment [Liu et al., 2019] is a simple 2v2 team game where agents can move in 3D space to kick a ball into the opponent team's goal. This type of environment allows investigation of emergent cooperative behaviours such as ball chasing.
- **Google Research Football (GRF):** A much more realistic football game is Google Research Football [Kurach et al., 2020], which resembles popular football games available on games consoles. Here, all the standard rules of the game apply such as corner kicks, fouls, cards, kick-off, offside, etc. In addition, the physical representation of the players is highly realistic. This environment allows for a much more complex range of learning behaviours to be studied, alongside customising the difficulty level.

Recently, a surging interest is seen in enhancing cooperation in mixed-motive tasks. Some representative tasks are described below.

- **Clean-up:** Clean-up [Hughes et al., 2018] was inspired by social dilemmas of public good provision. Agents gain reward by harvesting apples available in an orchard. There is a river which feeds the apple orchard. However, pollution is being dumped into the river from outside the environment so it fills up with pollution with a constant probability over time. As the proportion of the river filled with pollution increases, the growth rate of apples monotonically decreases. No apples grow at all once pollution levels exceed a threshold. Individuals can spend time cleaning the river to remove its pollution, an extended course of action that is analogous to making a contribution to the public good of size proportional to the amount of time they clean and their skill in doing so. This is a sequential social dilemma due to the temptation of neglecting the river cleaning to instead free-ride by collecting apples in the orchard, a course of action which leads to ruin if all elect it simultaneously.
- **Commons Harvest:** Commons Harvest was inspired by common-pool resource appropriation scenarios (see Janssen et al. [2010]). The MAL environment was introduced in Perolat et al. [2017] under the name Commons. Later it

was renamed to Harvest [Hughes et al., 2018]. Subsequent work used both names interchangeably. More recently it has usually been called Commons Harvest to try reduce the naming confusion [Leibo et al., 2021, Agapiou et al., 2022]. In Commons Harvest agents must navigate a 2D world to collect apples. The apple spawn rate in each location depends positively on the number of nearby apples, so that they grow in groups. If there are no apples in a local area then the probability of new growth in that area is zero. This is a sequential social dilemma because there is conflict between the short-term reward of collecting all apples in a particular area vs the long-term cost of apples never growing back in the area. Most of the mixed-motive methods we have discussed were evaluated on both Clean-up and Commons Harvest.

- **Coin Game:** Another environment that is commonly used by the discussed methods in Coin Game [Lerer and Peysakhovich, 2017], where agents navigate a gridworld to collect randomly placed coins, each of a randomly assigned colour. Agents obtain a reward of +1 for collecting any coin but if it is of the other agent's colour, the latter receives a reward of -2. If both agents succomb to this temptation, their expected reward is 0. Agents must therefore learn to sacrifice and coordinate based on colours in order to maintain long-term rewards.
- **Level-Based Foraging (LBF):** In the LBF tasks [Christianos et al., 2020, Papoudakis et al., 2021], agents navigate a grid-like domain where their goal is to gather items. These items and agents have distinct levels assigned to them. To successfully gather an item, the combined levels of cooperating agents must meet or surpass the item's level. Successfully acquiring an item grants agents a reward corresponding to the item's level. Full visibility of the environment is the default setting for agents, but a variation limits their vision to a nearby 5x5 square.
- **Melting Pot:** Melting Pot [Leibo et al., 2021, Agapiou et al., 2022] is an evaluation suite tailored to assess the capability of MARL algorithms to interact and adapt when faced with unfamiliar agents within familiar games that are here called substrates. A substrate together with a population of agents to face is called a scenario. The suite encompasses a breadth of social dynamics, such as cooperation, competition, deception, and trust in a range of over 50 substrates (for the expanded Meltingpot 2.0) and over 250 scenarios. After training agents on specific multi-agent games, the real test comes from evaluating their adaptability with new agents in the same game contexts. The unique aspect of Melting Pot is its scalability: by simply introducing a new set of opponent agents, an entirely new evaluation challenge is created, without altering the inherent game. Melting Pot brings together many pre-existing environments, including several of those mentioned above: Overcooked (which it calls Collaborative Cooking), Clean-up, Commons Harvest, and Coin Game.

6.2 Metrics

We will now outline the evaluation metrics used in the literature for the methods we have discussed. The choice of metric depends on the principles being evaluated. In the team game case, there is a common reward for all agents at each step and all incentives are therefore the same. This makes metric choice easy as there is only one option, namely the value of the common reward achieved during an episode (or equivalently the common win/success rate). Accordingly, all team game methods discussed use the common reward or win rate as their evaluation metric. One slight variation here is that agents have distinct rewards but still the collective reward is considered for the team performance.

In the mixed-motive setting, the situation is much more complex as here there is an inherent conflict between agents' individual incentives and those of the group. Which incentives do we prefer the system to optimise, those of the individual or the group? Indeed how do we even measure group incentives? The methods we have discussed for this setting all agree on the importance of collective return or utilitarian metric, which is the total episodic reward received over all agents:

$$R_C = \sum_{i=1}^{N} R^i. \tag{35}$$

Note that in some cases this is substituted by the average return over agents but the concept here is identical. Most works additionally employ some social metrics aimed at quantifying the level of cooperativeness achieved. We find that, although a selected few metrics are used by multiple works, there is in general little consensus on how to measure cooperativeness and the metrics tend to be heavily task-specific or subjective. In the following we will go through the social metrics used in the mixed-motive literature.

Sustainability Capturing the notion of sacrificing selfish immediate rewards in order to maintain long-term rewards for the group, sustainability measures the average time at which rewards are achieved:

$$S = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} t^i], \tag{36}$$

where $t^i = \mathbb{E}[t|r_t^i > 0]$. This sustainability metric was introduced in Perolat et al. [2017] and later used by Inequity Aversion [Hughes et al., 2018], Reciprocity via motivation to imitate [Eccles et al., 2019] and Gifting [Lupu and Precup, 2020].

Equality The Gini coefficient is commonly used to measure the inequality of achieved rewards within the system:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |U_i - U_j|}{2n^2 \bar{U}}. \tag{37}$$

The metric $E := 1 - G$ of equality is used by the three works mentioned above alongside sustainability.

Mechanism-based metrics Given a novel methodological mechanism that has been introduced, a common evaluative technique is to measure how well that mechanistic component has done its job. For example, Gifting introduces the mechanism of allowing agents to gift rewards to others and accordingly, one evaluation metric used is the gifting action frequency. Another example is Learning to Incentivise Others (LIO) [Yang et al., 2020b] which introduces a similar mechanism to allow agents to provide incentives to other agents, which can also be used as metrics for evaluation. Classifier Norm Model (CNM) [Vinitsky et al., 2023] used the fact that agents can sanction or "zap" other agents and measured this by monitoring the evolution of the zap likelihood. In Social Influence [Jaques et al., 2019], there are influencer agents who emit symbols before performing actions in order to influence others. The authors introduce some custom metrics to evaluate this mechanism. Speaker consistency measures how consistently an influencer emits a particular symbol before performing a particular action and instantaneous coordination measures the mutual information between a) the influencer's symbol and the influencee's next action and b) the influencer's action and the influencee's next action.

Task-based metrics Many metrics measure the accomplishment of task-specific goals assumed to relate to enhancing cooperativeness within the particular task. Examples of this include the waste cleared metric used for Clean-up by Inequity Aversion [Hughes et al., 2018] and Reciprocity [Eccles et al., 2019] via motivation to imitate [Eccles et al., 2019], apple consumption and berry fraction evolution used respectively by Inequity Aversion and CNM [Vinitsky et al., 2023] for their fruit-gathering based tasks, and proportion of own coins collected, used by Reciprocity and LOLA [Foerster et al., 2018b] for their Coin Game experiments. Social Value Orientation (SVO) [McKee et al., 2020] analyses abstention from depleting resources (a task-specific measure of sustainability) and interagent distance, measuring how well agents divide up the map between themselves. Finally, in Reputation [McKee et al., 2021], there are metrics measuring group contribution, territoriality and turn-taking, each assessing cooperative behaviours relevant to Clean-up.

7 Conclusion and Future Directions

In this paper, we have presented a comprehensive exploration of cooperation within the realm of multi-agent learning. An exhaustive review has been conducted, scrutinising contemporary advancements in cooperation across diverse scenarios — be it unified payoffs or distinct individual payoffs, and spanning both centralised and decentralised training paradigms. Furthermore, we proffer an encompassing compilation of benchmarks tailored explicitly for the evaluation of cooperative multi-agent learning endeavors. Conclusively, we delineate prospective research trajectories and underscore extant gaps in the literature that warrant future scholarly attention. Below we suggest some areas we believe would be highly beneficial for future research based on the overview we have presented. Specifically, we highlight possible future advances in emergent cooperation, generalisation and evaluation techniques.

Foundation models as Agents Recently, large language models (LLMs) have demonstrated remarkable potential in achieving human-level intelligence through the acquisition of vast amounts of web knowledge. This has sparked an upsurge in studies investigating LLM-based autonomous agents. In the context of multi-agent cooperation, we are interested in promote cooperation harnessing LLMs. Early attempts [Hong et al., 2023, Zhang et al., 2023] show that these models often perform better when prompted with one specific objective and relevant context than with a larger range of goals and sub-goals at once. The techniques of retrieving relevant context and individual prompting have also been used to create social simulations where a larger number of players go about their daily lives in a simulated world [Park et al., 2023, Vezhnevets et al., 2023]. The potential for studying multi-agent cooperation in such settings is still not fully understood though it likely to be substantial.

Cooperation with Novel Agents (zero-shot social generalization) Zero-shot generalization is a setting where there is a relative dearth of dedicated RL-based methods. We refer here to environments where agents must learn to cooperate with heterogeneous other agents and achieve good zero-shot cooperation performance with strangers at test time. In this case the demand is for agents and populations of agents capable of functioning both as visitors to an unfamiliar culture not seen during training, as well as in the role of the dominant "resident" population—joined by visitors from outside who were never encountered during training. The two cases are quite different. When an agent visits a larger unfamiliar group it often must adapt to the local conventions, which may be unfamiliar. When a population of agents are resident (they are the majority), then it is up to them to provide public goods and resolve social dilemmas, and to do so while remaining robust to distraction from new joiners. Melting Pot [Leibo et al., 2021, Agapiou et al., 2022] is a large suite of environments specifically for studying zero-shot generalization, broadly construed. Most environments in Melting Pot are mixed motive though Melting Pot also includes some pure common-interest environments (Overcooked) and some team-based zero-sum environments (capture the flag and king of the hill).

In the case of test-time agents being humans or based on human data, there is a scarcity in adequate evaluation benchmarks, given that it can be rather cumbersome to obtain human or human data-based heterogeneous agents. PECAN [Lou et al., 2023] is a good example of this very approach evaluated on Overcooked, but more such environment suites and algorithms are needed. One possible idea is a tournament-style evaluation of algorithms amidst participating human agents, such as was done in Axelrod [1980] for Iterated Prisoner's Dilemma.

Evaluation Most evaluation benchmarks in cooperative MARL, especially in mixed-motive settings, tend to be toy environments, e.g. gridworld-based. Some team game environments like SMAC and Google Research Football are more complex and realistic but more work is needed to model this realism when individual incentives may not be aligned to the group incentive. Future work could focus on domains like autonomous driving and robotics, where realistic simulators are already available. In these cases, significant strategic complexity could be added through e.g. road map design, social vehicle heterogeneity or large systems of cooperating warehouse robots. We also identified a gap in the field concerning the development of standardised evaluation metrics specifically designed to measure cooperativeness. As discussed, sustainability and equality are two existing examples, but more are needed. This is evident in the fact that most works introduce highly specific task- and mechanism-based metrics. A new set of generally applicable metrics covering all the core concepts of the currently used specific metrics would be of great value to the community and would allow standardised benchmarking of newly introduced techniques.

References

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. The ai economist: Optimal economic policy design via two-level deep reinforcement learning. *arXiv preprint arXiv:2108.02755*, 2021a.

Michael Bradley Johanson, Edward Hughes, Finbarr Timbers, and Joel Z Leibo. Emergent bartering behaviour in multi-agent reinforcement learning. *arXiv preprint arXiv:2205.06760*, 2022.

Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

Edgar A Duéñez-Guzmán, Suzanne Sadedin, Jane X Wang, Kevin R McKee, and Joel Z Leibo. A social path to human-like artificial intelligence. *Nature Machine Intelligence*, pages 1–8, 2023.

Uri Hertz, Raphael Koster, Marco Janssen, and Joel Z Leibo. Beyond the matrix: Experimental approaches to studying social-ecological systems. 2023.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open Problems in Cooperative AI, December 2020.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752):2, 1998.

Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018a.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 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 (AAMAS)*, pages 2085–2087, 2018.

Anatol Rapoport. Prisoner's dilemma—recollections and observations. In *Game Theory as a Theory of a Conflict Resolution*, pages 17–34. Springer, 1974.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

Anthony Rocco Cassandra. *Exact and approximate algorithms for partially observable Markov decision processes*. Brown University, 1998.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

Csaba Szepesvári and Michael L Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060, 1999.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38:287–308, 2000.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063, 1999.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pages 1861–1870, 2018.

Thomas C Schelling. *The Strategy of Conflict: with a new Preface by the Author*. Harvard university press, 1960.

Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*, pages 195–210, 1996.

Martin Lauer and Martin A Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the seventeenth international conference on machine learning*, pages 535–542, 2000.

Xiaofeng Wang and Tuomas Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. *Advances in neural information processing systems*, 15, 2002.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning (ICML)*, pages 5872–5881, 2018.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33:2074–2086, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

OpenAI. Openai five. https://blog.openai.com/openai-five/, 2018.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Michael L Littman et al. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328, 2001.

Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial intelligence*, 171(7):365–377, 2007.

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1702.08887*, 2017.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017a.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019a.

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Joel Z Leibo, Julien Perolat, Edward Hughes, Steven Wheelwright, Adam H Marblestone, Edgar Duéñez-Guzmán, Peter Sunehag, Iain Dunning, and Thore Graepel. Malthusian Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pages 1099–1107, Montreal, QC, Canada, May 2019b.

Peter Sunehag, Alexander Sasha Vezhnevets, Edgar Duéñez-Guzmán, Igor Mordach, and Joel Z Leibo. Diversity through exclusion (dte): Niche identification for reinforcement learning through value-decomposition. *arXiv preprint arXiv:2302.01180*, 2023.

Muhammad A Rahman, Niklas Hopner, Filippos Christianos, and Stefano V Albrecht. Towards open ad hoc teamwork using graph-based policy learning. In *International Conference on Machine Learning*, pages 8776–8786. PMLR, 2021.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play"' for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*, pages 6187–6199. PMLR, 2021.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1504–1509, 2010.

John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.

DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.

Udari Madhushani, Kevin R McKee, John P Agapiou, Joel Z Leibo, Richard Everett, Thomas Anthony, Edward Hughes, Karl Tuyls, and Edgar A Duéñez-Guzmán. Heterogeneous social value orientation leads to meaningful diversity in sequential social dilemmas. *arXiv preprint arXiv:2305.00768*, 2023.

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *International Conference on Machine Learning*, pages 434–443. PMLR, 2019.

Xingzhou Lou, Jiaxian Guo, Junge Zhang, Jun Wang, Kaiqi Huang, and Yali Du. Pecan: Leveraging policy ensemble for context-aware zero-shot human-ai coordination. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 679–688, 2023.

Yang Li, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Cooperative open-ended learning framework for zero-shot coordination. *arXiv preprint arXiv:2302.04831*, 2023.

Alexander Vezhnevets, Yuhuai Wu, Maria Eckstein, Rémi Leblond, and Joel Z Leibo. Options as responses: Grounding behavioural hierarchies in multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 9733–9742. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. In *International Conference on Machine Learning*, pages 10041–10052. PMLR, 2022.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7285–7292, 2020a.

Sirui Chen, Zhaowei Zhang, Yali Du, and Yaodong Yang. Stas: Spatial-temporal return decomposition for multi-agent reinforcement learning. *arXiv preprint arXiv:2304.07520*, 2023.

Jennifer She, Jayesh K Gupta, and Mykel J Kochenderfer. Agent-time attention for sparse rewards multi-agent reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1723–1725, 2022.

Raphael Köster, Kevin R McKee, Richard Everett, Laura Weidinger, William S Isaac, Edward Hughes, Edgar A Duéñez-Guzmán, Thore Graepel, Matthew Botvinick, and Joel Z Leibo. Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. *arXiv preprint arXiv:2010.09054*, 2020.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, 30, 2017.

John Rawls. *A Theory of Justice*. The Belknap Press of Harvard University Press, 1971.

Lidia Ceriani and Paolo Verme. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10(3):421–443, September 2012. ISSN 1569-1721, 1573-8701. doi:10.1007/s10888-011-9188-x.

Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6379–6390, 2017b.

Yali Du, Lei Han, Meng Fang, Tianhong Dai, Ji Liu, and Dacheng Tao. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Lei Han, Peng Sun, Yali Du, Jiechao Xiong, Qing Wang, Xinghai Sun, Han Liu, and Tong Zhang. Grid-wise control for multi-agent reinforcement learning in video game ai. In *International Conference on Machine Learning (ICML)*, pages 2576–2585, 2019.

Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 2961–2970. PMLR, 2019.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624, 2022.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 4292–4301, 2018.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*. International Conference on Machine Learning Organizing Committee, 2019.

Yaodong Yang, Jianye Hao, Ben Liao, Kun Shao, Guangyong Chen, Wulong Liu, and Hongyao Tang. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*, 2020a.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations*, 2020b.

Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, and Chongjie Zhang. Towards understanding cooperative multi-agent q-learning with value factorization. *Advances in Neural Information Processing Systems*, 34:29142–29155, 2021a.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2137–2145, 2016.

Sainbayar Sukhbaatar and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2244–2252, 2016.

Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7254–7264, 2018.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning (ICML)*, pages 1538–1546, 2019.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations (ICLR)*, 2019.

Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. Learning to schedule communication in multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Wendelin Böhmer, Vitaly Kurin, and Shimon Whiteson. Deep coordination graphs. In *In International Conference on Machine Learning (ICML)*, pages 01–11, 2020.

Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020. URL `https://openreview.net/forum?id=HkxdQkSYDB`.

Tianshu Chu, Sandeep Chinchali, and Sachin Katti. Multi-agent reinforcement learning for networked system control. In *International Conference on Learning Representations (ICLR)*, 2020. URL `https://openreview.net/forum?id=Syx7A3NFvH`.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.

Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

Daniel Barkoczi and Mirta Galesic. Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, 7(1):1–8, 2016.

David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007.

Shubham Gupta, Rishi Hazra, and Ambedkar Dukkipati. Networked multi-agent reinforcement learning with emergent communication. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1858–1860, 2020.

Yedid Hoshen. Vain: Attentional multi-agent predictive modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2701–2711, 2017.

Yaru Niu, Rohan Paleja, and Matthew Gombolay. Multi-agent graph-attention communication and teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 964–973, 2021.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Decentralized multi-agent reinforcement learning with networked agents: Recent advances. *arXiv preprint arXiv:1912.03821*, 2019.

Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. Shaq: Incorporating shapley value theory into multi-agent q-learning. *Advances in Neural Information Processing Systems*, 35:5941–5954, 2022.

David H Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(02n03):265–279, 2001.

Kalesha Bullard, Franziska Meier, Douwe Kiela, Joelle Pineau, and Jakob Foerster. Exploring zero-shot emergent communication in embodied multi-agent populations. *arXiv preprint arXiv:2010.15896*, 2020.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pages 7204–7213. PMLR, 2021.

Rui Zhao, Jinming Song, Hu Haifeng, Yang Gao, Yi Wu, Zhongqian Sun, and Yang Wei. Maximum entropy population based training for zero-shot human-ai coordination. *arXiv preprint arXiv:2112.11701*, 2021.

Michael W Macy and Andreas Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(suppl_3):7229–7236, 2002.

William E Walsh, Rajarshi Das, Gerald Tesauro, and Jeffrey O Kephart. Analyzing Complex Strategic Interactions in Multi-Agent Systems. *AAAI Technical Report WS-02-06*, June 2002.

Michael P Wellman. Methods for Empirical Game-Theoretic Analysis. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pages 1552–1556. AAAI Press, 2006.

Karl Tuyls, Julien Perolat, Marc Lanctot, Edward Hughes, Richard Everett, Joel Z. Leibo, Csaba Szepesvári, and Thore Graepel. Bounds and dynamics for empirical game theoretic analysis. *Autonomous Agents and Multi-Agent Systems*, 34(1):7, April 2020. ISSN 1387-2532, 1573-7454. doi:10.1007/s10458-019-09432-y.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Improved Algorithms for Learning Equilibria in Simulation-Based Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pages 79–87, Auckland, New Zealand, May 2020. International Foundation for Autonomous Agents and Multiagent Systems. doi:10.5555/3398761.3398776.

Thomas C Schelling. Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict resolution*, 17(3):381–428, 1973.

Peter Sunehag, Guy Lever, Siqi Liu, Josh Merel, Nicolas Heess, Joel Z Leibo, Edward Hughes, Tom Eccles, and Thore Graepel. Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems. In *Artificial life conference proceedings*, pages 103–110. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 2019.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The ai economist: Optimal economic policy design via two-level deep reinforcement learning, 2021b.

Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. Resolving social dilemmas with minimal reward transfer. *arXiv preprint arXiv:2310.12928*, 2023.

Po-An Chen and David Kempe. Altruism, selfishness, and spite in traffic routing. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 140–149, 2008.

Krzysztof R Apt and Guido Schäfer. Selfishness level of strategic games. *Journal of Artificial Intelligence Research*, 49:207–240, 2014.

Edward Hughes, Thomas W Anthony, Tom Eccles, Joel Z Leibo, David Balduzzi, and Yoram Bachrach. Learning to Resolve Alliance Dilemmas in Many-Player Zero-Sum Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pages 538–547, Auckland, New Zealand, 2020. International Foundation for Autonomous Agents and Multiagent Systems. doi:10.5555/3398761.3398827.

Phillip JK Christoffersen, Andreas A Haupt, and Dylan Hadfield-Menell. Get it in writing: Formal contracts mitigate social dilemmas in multi-agent rl. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 448–456, 2023.

Andrei Lupu and Doina Precup. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on autonomous agents and multiagent systems*, pages 789–797, 2020.

Richard Willis and Michael Luck. Resolving social dilemmas through reward transfer commitments. In *Adaptive and Learning Agents Workshop*, 2023.

Joseph Heath. Methodological Individualism. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2020.

Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2043–2044, 2018a.

Jane X Wang, Edward Hughes, Chrisantha Fernando, Wojciech M Czarnecki, Edgar A Duéñez-Guzmán, and Joel Z Leibo. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 683–692, 2019a.

David Earl Hostallero, Daewoo Kim, Sangwoo Moon, Kyunghwan Son, Wan Ju Kang, and Yung Yi. Inducing cooperation through reward reshaping based on peer evaluations in deep multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 520–528, 2020.

Woodrow Z. Wang, Mark Beliaev, Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, and Dorsa Sadigh. Emergent Prosociality in Multi-Agent Games Through Gifting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 434–442, Montreal, Canada, 2021b. ijcai.org. doi:10.24963/ijcai.2021/61.

Hossein Haeri, Reza Ahmadzadeh, and Kshitij Jerath. Reward-sharing relational networks in multi-agent reinforcement learning as a framework for emergent behavior. *arXiv preprint arXiv:2207.05886*, 2022.

Ian Gemp, Kevin R McKee, Richard Everett, Edgar Duéñez-Guzmán, Yoram Bachrach, David Balduzzi, and Andrea Tacchetti. D3c: Reducing the price of anarchy in multi-agent learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 498–506, 2022.

Minae Kwon, John Agapiou, Edgar Duéñez-Guzmán, Georgios Piliouras, Kalesha Bullard, and Ian Gemp. Auto-Aligning Multiagent Incentives with Global Objectives. In *Proceedings of the Adaptive and Learning Agents Workshop*, Online, May 2023. doi:10.5555/3398761.3398825.

Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*, 2016.

Alexander Peysakhovich and Adam Lerer. Towards AI that can solve social dilemmas. In *2018 AAAI Spring Symposia*, page 7, Stanford University, Palo Alto, California, USA, March 2018b. AAAI Press.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130, 2018b.

Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*, 2019.

Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*, pages 11:1–11:7, Beijing China, October 2019b. ACM. ISBN 978-1-4503-7656-3. doi:10.1145/3356464.3357712.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable Opponent Shaping in Differentiable Games. In *7th International Conference on Learning Representations*, New Orleans, LA, USA, May 2019. OpenReview.net.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.

Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems*, 33:15208–15219, 2020b.

Alexis Jacq, Julien Perolat, Matthieu Geist, and Olivier Pietquin. Foolproof Cooperative Learning. In *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pages 401–416. PMLR, November 2020.

Chris Lu, Timon Willi, Christian Schroeder de Witt, and Jakob Foerster. Model-Free Opponent Shaping. In *Proceedings of the 39th International Conference on Machine Learning Research*, volume 162, pages 14398–14411. PMLR, July 2022.

Kevin R McKee, Edward Hughes, Tina O Zhu, Martin J Chadwick, Raphael Koster, Antonio Garcia Castaneda, Charlie Beattie, Thore Graepel, Matt Botvinick, and Joel Z Leibo. A multi-agent reinforcement learning model of reputation and cooperation in human groups. *arXiv preprint arXiv:2103.04982*, 2021.

Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. Cooperation and Reputation Dynamics with Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, pages 115–123, Virtual Event, United Kingdom, May 2021. ACM. doi:10.5555/3463952.3463972.

Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2):26339137231162025, 2023.

Bowen Baker. Emergent Reciprocity and Team Formation from Randomized Uncertain Social Preferences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, virtual, December 2020.

Kyrill Schmid, Michael Kölle, and Tim Matheis. Learning to Participate through Trading of Reward Shares. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, volume 1, pages 355–362, Lisbon, Portugal, February 2023. SCITEPRESS. doi:10.5220/0011781600003393.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006. doi:10.1126/science.1133755. URL `https://www.science.org/doi/abs/10.1126/science.1133755`.

Robert Axelrod. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution*, 24(1):3–25, March 1980. ISSN 0022-0027, 1552-8766. doi:10.1177/002200278002400101.

Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems*, 35:5510–5521, 2022.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, and Julian Schrittwieser. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination, 2020.

Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, and Thore Graepel. Emergent coordination through competition. *arXiv preprint arXiv:1902.07151*, 2019.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4501–4510, 2020.

Marco A Janssen, Robert Holahan, Allen Lee, and Elinor Ostrom. Lab experiments for the study of social-ecological systems. *Science*, 328(5978):613–617, 2010.

Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv e-prints*, pages arXiv–1707, 2017.

Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. Metagpt: Meta programming for multi-agent collaborative framework, 2023.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. Proagent: Building proactive cooperative ai with large language models. *arXiv preprint arXiv:2308.11339*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*, 2023.