Deception in Social Learning: A Multi-Agent Reinforcement Learning Perspective

Paul C. Chelărescu
School of Informatics
University of Edinburgh
Edinburgh, UK, EH8 9AB
P.C.Chelarescu@sms.ed.ac.uk

Abstract

Within the framework of Multi-Agent Reinforcement Learning, Social Learning is a new class of algorithms that enables agents to reshape the reward function of other agents with the goal of promoting cooperation and achieving higher global rewards in mixed-motive games. However, this new modification allows agents unprecedented access to each other’s learning process, which can drastically increase the risk of manipulation when an agent does not realize it is being deceived into adopting policies which are not actually in its own best interest. This research review introduces the problem statement, defines key concepts, critically evaluates existing evidence and addresses open problems that should be addressed in future research.

1 Introduction

Recent successes in Artificial Intelligence have brought Reinforcement Learning (RL) to the forefront attention of the research community, through examples such as learning how to play Go, Chess and Atari Games with the same algorithm [Schrittwieser et al., 2019], solving a physical Rubik’s cube [OpenAI et al., 2019], controlling power grids [Glavic et al., 2017], routing vehicles [Nazari et al., 2018], improving tax policy [Zheng et al., 2020] or improving cyber security [Nguyen and Reddi, 2019]. Essentially, Reinforcement Learning is an area of Machine Learning which studies how individual agents should behave optimally in an environment. Agents learn by following signals, called rewards that indicate desirable actions, and have the sole responsibility of creating their own decisions. Agents can represent a variety of entities, from software programs to physical robots, but more recently, agents can represent experts of whom an artificial agent could learn better behaviours from [Zhang et al., 2019].

While the above examples are usually about single-agent problems, a more broad decision-making problem would capture the behaviour of other learning agents that are in the environment and recognize that they could be learning how to react to the learning of others. Multi-Agent Reinforcement Learning (MARL) is a learning approach for such problems, that extends the capabilities of Reinforcement Learning to model settings with multiple, collectively interacting agents that are engaged in a continual process of learning. MARL is a flexible approach and has enabled successes such as coordinating fleets of unmanned aerial vehicles [Alon and Zhou, 2020].

Whereas in classical MARL agents learn how to interact with each other, the reward function is a property of the environment and can only be altered by external factors. Agents are therefore focused on learning only insofar as it helps them collect their individual rewards while having no means to shape the reward function of others. Hence, agents can only influence others indirectly, through the ways in which they optimize their assigned reward function of the environment. While
this can be natural for Single-Agent Reinforcement Learning, it greatly restricts inter-agent interactions, limiting behaviours such as simulating trade deals and making negotiations [Lupu and Precup, 2020].

MARL can be extended with mechanisms that allow agents to directly, rather than indirectly, influence the learning process of each other. Such examples can be found in the works of Yang et al. [2020a] and Lupu and Precup [2020], in which agents can reward each other for behaviours that they themselves see as desirable, rather than relying solely on the environment’s judgement of desirability.

On one hand, the advances in mechanisms to directly affect others have allowed agents to collaborate unprecedentedly well in environments which reward selfishness in the short term at the cost of reduced future reward. So-called Social Dilemmas, these environments are specifically designed to lead collectively selfish behaviour to sub-optimal rewards, with the only sustainable and optimal policy being collective cooperation. One might draw parallels between Social Dilemmas and many challenges that humanity faces, such as Climate Change, where an individual country might be tempted to reduce its efforts towards sustainability due to priorities focused on the short-term while ignoring that if every country behaved the same, long-term goals would be negatively affected. Evidently, such stakes require large-scale, sustained cooperation, and the ever-increasing digitalization of the economy necessitates the study of the impact of Artificial Intelligence that learns how to cooperate with others.

On the other hand, allowing the judgment of other agents to influence an agent’s reward necessitates trusting the judgement of others. Most of the current research into incentivizing others, such as Lupu and Precup [2020] and Yang et al. [2020a], consider the population of agents homogeneous, in the sense that agent policies are either identical or very similar. Achieving cooperation in heterogeneous groups, in which agents are of different types and might have different judgements, is still largely an open problem [McKee et al., 2020]. One stark question to ask is what would happen if an agent were in competition with the rest of the group, while still having the power to influence others according to its own judgement? Could such an agent deceive the others with regards to its intentions and make them trust it? What mechanisms should be created to ensure that a collaborative group can safely send and receive incentives without worrying about misaligned goals? Can a group of agents form a coalition against an identified enemy and avoid getting exploited or deceived by adversaries?

This review investigates the delicacy of cooperation enhanced by social incentives, all within the framework of MARL. It asks why is it desirable, and then, which are the ways in which mechanisms that shape the reward functions of others create opportunities for deception. This review focuses on mixed cooperative-competitive settings, rather than purely cooperative or purely competitive settings since the first scenario is more amenable to deception and is less researched [McKee et al., 2020]. Methods that focus on deception in purely competitive settings, such as cyber-security games, are mostly omitted from this review.

1.1 Related Work

Mechanisms to directly influence other agents through changing their reward function are novel in the literature, but the context in which they appeared is grounded in a few key areas. First of all, these mechanisms pursue goals aligned with the field of Cooperative AI, which has been highlighted as a key research area in several recent surveys [Dafoe et al., 2020] [OroojlooyJadid and Hajinezhad, 2019]. Furthermore, these mechanisms are related to the concept of “Reward Shaping” [Hostallero et al., 2020] [Zheng et al., 2018] [Devlin et al., 2014] [Grzes, 2017]. Predicting the effects of influencing others is related to inferring the mental state of other agents, which has been studied before as the “Theory of Mind” [Shum et al., 2019] [Yuan et al., 2019]. Influencing others can allow one to deceive, and while not in the context of directly influencing others in MARL, deception has been studied before in cyber-security as a means to gain an advantage [Islam and Al-Shaer, 2020]. Moreover, controlling the learning experience of reinforcement learning agents to ensure the safety of their learned behaviour has been studied within the context of ‘Reward Tampering’ [Everitt and Hutter, 2019], which is related to ensuring that influencing others does not produce undesired effects in their behaviour.

The mechanisms that enable directly influencing others can produce undesired behaviours, though self-enforcing feedback loops between agents. Safely learning from human demonstrators has been
studied before as a way to ensure the alignment between the intentions of the teacher and the learner [Brown et al., 2020] and could potentially address the issue of alignment. Relying on the judgement of other agents that can freely enter and leave the environment has been studied before within the context of trust and reputation in multi-agent systems [Hazard, 2010].

Social incentives can be deployed in safety-critical applications, such as between self-driving cars. Ensuring cooperation in complex and dynamic environments might require studying how to satisfy all the involved parties through tactics of negotiation, which could be critical in self-driving [Shalev-Shwartz et al., 2016]. An agent that perceives itself as compensated unfairly for its contributions might be more willing to partake in deception tactics to promote its own fairness, something which could have unintended consequences. Mechanisms that promote fairness within a population of heterogeneous agents, where agents are of varying abilities, might prove helpful in this regard [Zhu and Oh, 2018]. Last but not least, the field of 'Behavioural Economics' could offer insights about decision-making in ill-structured environments, which can then be further integrated into computational models [Chen et al., 2017].

2 Background

There are numerous challenges in MARL that need to be addressed in order to ensure the safety of agents seeking to ensure cooperation in contexts of incentivized setups. What follows is a brief introduction into the field of Reinforcement Learning, its extension to Multi-Agent environments, an overview of current methods that enable social cooperation and key scenarios and behaviours that can threaten the safety of agents involved.

The key distinction that separates mechanisms for direct incentivizing through reward-giving and those which promote cooperation unilaterally [Yang et al., 2020b] is the degree of decentralization in the former in contrast with centralization in the latter. Centralization is a strong assumption when a single unit is in charge of taking the decisions for every agent, with the goal of achieving cooperation between units. Previous work has tried to decouple this strong assumption and maintain decentralization only at testing time, but limitations still include the necessity of a centralized unit at training time [Chert, 2019].

In a multi-agent environment, the experience and outcomes of agents are interdependent. The interactions created by this interdependence can be sorted into two categories, based on the alignments of their incentives: pure-motive and mixed-motive. In pure-motive interactions, motives are either entirely aligned or entirely opposed, corresponding to settings of pure collaboration and pure competition. In mixed-motive interactions, the environment allows for more complex behaviours such as coalitions and betrayals. A recent survey has found that situations of mixed-motives interactions present under-explored opportunities [Dafoe et al., 2020].

The methods explored in this review focus mainly on decentralized, mixed-motive interactions. Furthermore, Reinforcement Learning algorithms with tabular forms are usually not sufficient for large state/action spaces, therefore, most of the focus in this review will be on algorithms that use function approximators, usually neural networks.

2.1 Single-Agent Reinforcement Learning

An intelligent agent acting in an environment can be modelled as a Markov Decision Process (MDP). An MDP is a mathematical framework which can be used to model decision making, by modelling possible scenarios as states and possible actions as transitions between states. Finding an optimal policy is the problem of taking actions in states in order to maximize a scalar signal, called reward. Rewards are attributed to an agent, classically only by the environment (but they can be given by other agents as well [Lupu and Precup, 2020] [Yang et al., 2020a]) when an agent takes a desirable action in a particular state.

Reinforcement Learning approximates the optimal policy for and MDP in the following way: at timestep $t$, an agent finds itself in a state $s_t \in S$, in which $S$ is set representing the state space, and is required to take an action $a_t \in A(s_t)$, where $A(s_t)$ is the set of valid actions in state $s_t$. Upon executing action $a_t$, the agent finds itself in a new state $s_{t+1}$ and receives a reward $r_t \in \mathbb{R}$ for transitioning into that state. The goal of the agent is to learn a policy that maximizes its long-term collected reward. A policy $\pi$ represents a probability distribution over all actions $A(s)$ for every
state \( s \in S \). The Value function \( V^\pi(s) \) of a state \( s \) is the cumulative reward that can be collected by starting in that state and following \( \pi \). In equation form, this is given by

\[
V^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r_{t+1} | s_0 = s \right],
\]

where \( \gamma \in [0, 1) \) is the discounting factor used to balance between short-term and long-term rewards. For a given transition probability distribution \( p(s_{t+1} | s_t, a_t) \) and reward function \( r(s_t, a_t) \), the following equation, defined by Bellman [1957], holds for every state \( s_t \) at any timestep \( t \):

\[
V^\pi(s_t) = \sum_{a \in A(s_t)} \pi(a | s_t) \sum_{s_{t+1} \in S} p(s_{t+1} | s_t, a) \left[ r(s_t, a) + \gamma V^\pi(s_{t+1}) \right],
\]

The optimal state-value and optimal policy \( \pi^* \) can be obtained by maximizing over the actions:

\[
V^{\pi^*}(s_t) = \max_a \sum_{s_{t+1}} p(s_{t+1} | s_t, a) \left[ r(s_t, a) + \gamma V^{\pi^*}(s_{t+1}) \right].
\]

The optimal action-value for taking an action \( a_t \) in state \( s_t \) can be obtained from the following equation:

\[
Q^{\pi^*}(s_t, a_t) = \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \left[ r(s_t, a_t) + \gamma \max_{a'} Q^{\pi^*}(s_{t+1}, a') \right].
\]

Methods used to obtain \( \pi^* \) from the above equations are called value-based methods. In real-world settings, exploring the entire state and action space is unfeasible, hence finding the true probability distribution \( p(s_{t+1} | s_t, a_t) \) for every state and action is intractable. For this reason, a function \( J \) parameterized by \( \theta \) is used to approximate the state-value, action-value or the policy itself. The parameters \( \theta \) can function as a compression of the full probability distribution being approximated. While a simple linear function can work as a function approximator, the complexities of the environment can be better captured by neural networks, which has attracted a tremendous amount of recent research [Mnih et al., 2013] [Mnih et al., 2015] [Gao et al., 2019].

### 2.2 Multi-Agent Reinforcement Learning

Following from the framework of Single-Agent RL, where the presence of other agents would simply be modelled as part of the environment, MARL explicitly acknowledges the presence of other agents by modelling separate individual state observations, actions and rewards. In particular, Single-Agent RL fails to take into account that the Markov property (future state transitions and rewards only depend on the present state) becomes invalid when multiple learning agents are present in the environment and create a non-stationary scenario due to their private learning experience. Applying algorithms for Single-Agent RL in Multi-Agent settings could work in practice [Matignon et al., 2012], but the lack of guarantees can make the learning process highly unstable or fail all together [Shoham et al., 2007].

In general, MARL can be described using two frameworks: Markov Games or Extensive-Form Games. The former captures the intertwining of multiple agents, but is limited to settings of full observation, i.e., every agent has perfect information with regards to state \( s_t \) and action \( a_t \) at every timestep \( t \). This assumption does not hold in a plethora of cases and thus every agent only has a partial view of the state \( s_t \), i.e., the game exhibits imperfect-information. For such scenarios, the framework of Extensive-Form Games is usually a more appropriate choice. However, only the former is formally defined in this review.

A Markov Game consists of a tuple \( \langle S, N, A, T, R \rangle \) where \( S \) represents the set of states, \( N \geq 2 \) signifies the number of agents in the game and \( A \) represents the set of actions available in the game. \( T \) represents the transition probabilities and it is a function that is defined on the Cartesian product between the state and the action of every agent. Similarly, the reward function \( R \) depends on the actions of every agent.

For a learning agent \( i \) and the set of every other agent \( -i = N \setminus \{i\} \), the value function now depends on the joint action \( \alpha = (\alpha_i, \alpha_{-i}) \) and the joint policy \( \pi(s, \alpha) = \prod_j \pi_j(s, \alpha_j) \):
\[ V^\pi_i(s_t) = \sum_{a \in A} \pi_i(s_t, a) \sum_{s_{t+1} \in S} T(s_t, a_i, a_{-i}, s_{t+1})[R_i(s_t, a_i, a_{-i}, s_{t+1}) + \gamma V_i(s_{t+1})]. \] (5)

As a consequence, the optimal policy depends on the policy \( \pi_{-i}(s, a_{-i}) \) of all the other agents, which can be non-stationary:

\[
\begin{align*}
\pi^*_i(s_t, a_i, \pi_{-i}) &= \arg \max_{\pi_i} V^{(\pi_i, \pi_{-i})}_i(s_t) = \\
&= \arg \max_{\pi_i} \sum_{a \in A} \pi_i(s_t, a_i) \pi_{-i}(s_t, a_{-i}) \sum_{s_{t+1} \in S} T(s_t, a_i, a_{-i}, s_{t+1})[R_i(s_t, a_i, a_{-i}, s_{t+1}) + \gamma V^{(\pi_i, \pi_{-i})}_i(s_{t+1})].
\end{align*}
\] (6)

A few examples of recent successful MARL algorithms are MADDPG [Lowe et al., 2017] and EAQR [Zhang and Wang, 2018]. More examples can be found in Zhang et al. [2019b].

In contrast with Markov Games, Extensive-Form Games define a game tree and attribute each node in the tree to either one of the players or to the environment. Each state corresponds to a node and each action is modelled as an edge between a node and its children. The key distinction from Markov Games is that an agent which has to take a decision in state \( s_t \) does not know the full history of states up to that point in time, since the trajectory through the game tree includes actions that other players have taken in private. Algorithms such as counterfactual regret minimization have been applied to this framework with great success, for both two-player [Brown and Sandholm, 2018] and multi-player games [Brown and Sandholm, 2019].

In addition to the above definitions, the reward function is constrained depending on the incentive alignment structure of the game. In pure-motive cooperative settings the reward function is commonly shared between every agent, i.e., \( \mathcal{R}^1 = \mathcal{R}^2 = \ldots = \mathcal{R}^N \). Within pure-motive settings, but in purely competitive environments, the game would typically be modeled as a zero-sum Markov Game, i.e., \( \sum_{i \in N} \mathcal{R}(s_t, a_t, s_{t+1}) = 0 \). In contrast, mixed-motive settings have no restriction upon the shape of the reward function attributed to each agent. Since each agent is self-interested, however, the reward of one agent could be in conflict with that of the others, but this is not a necessity since agents could find policies that jointly benefit their cumulative rewards.

There are numerous works concerning open problems in MARL which will not be discussed in this research review but are worth mentioning due to their importance. A non-exhaustive list is as follows: convergence properties [Song et al., 2019], scalability [Zhou et al., 2020], the curse of dimensionality [Busoniu et al., 2008], non-stationarity [Papoudakis et al., 2019] and credit assignment [Agogino and Tumer, 2004].

### 2.3 Social Learning

Social Learning is a powerful component in human and animal intelligence, wherein an individual learns from the observations and interactions with others. By using collective intelligence, an individual is able to ‘stand on the shoulders of giants’. One example in the animal kingdom is fish that are eating food that has already been discovered [Laland, 2004]. Evidently, humans can also teach each-other new behaviours and are able to observe experts in order to improve their own abilities [Boyd et al., 2011]. Beyond simple imitation, humans can form coalitions around an issue and either reward or punish each-other based on reciprocal agreements and established norms.

In this review, Social Learning is expanded to encompass Multi-Agent Reinforcement Learning algorithms that enable agents to reshape the reward function attributed to other agents. Some of the recent examples include the ability to directly reward others from a private budget [Lupu and Precup, 2020], learning an incentive function that explicitly accounts for its impact on the behaviours of its recipients [Yang et al., 2020a], learning cues from new experts to achieve high zero-shot performance [Ndousse et al., 2020], amending the Markov Game by allowing a benevolent agent to hand out additional rewards and punishments with the scope of maximizing total social welfare [Baumann et al., 2020] or rewarding agents for having causal influence over others through counterfactual reasoning [Jaques et al., 2019].
While expanding upon the methods of each one of the above-mentioned papers is out of the scope of this review, several notable changes are going to be presented. First of all, most of the modifications surrounding Social Learning concern the reward function $R$. For instance, in Jaques et al. [2019] an agent’s immediate reward is modified so that it becomes

$$r^k_t = \alpha e^k_t + \beta c^k_t,$$

with $e^k_t$ being the classical extrinsic, environmental reward and $c^k_t$ being the causal influence reward.

A similar change is made in Yang et al. [2020a], where the reward for agent $j \in N$ becomes

$$r^j(s_t, a_t, \eta^{-j}) := r^j_{\text{env}}(s_t, a_t) + \sum_{i \neq j} r^j_{\eta^i}(o^i_t, a^{-i}_t),$$

where $r^j_{\eta^i}: \mathcal{O} \times \mathcal{A}^{-i} \mapsto \mathbb{R}^{N-1}$ is an incentive function parameterized by $\eta^i \in \mathbb{R}^n$ for player $i \in N$, that maps the observation $o^i$ and all the other agents’ actions $a^{-i}$ to a vector of rewards for the other $N \setminus \{i\}$ agents and $r^j_{\text{env}}$ is the extrinsic, environmental reward.

Secondly, the source of the rewards received by other agents is important, as noticed by Lupu and Precup [2020]. In their work, the authors find that the most successful strategy is one in which the rewards given to others are subtracted from the rewards of the giver. The authors conjecture that this is due to an early decrease in peace (agents can time-out each other as punishment) and due to agents learning restraint by experiencing altruistic behaviour earlier in their training (altruistic in the sense that an action is taken by an agent to decrease its own immediate reward to increase the reward of others, with the expectation of future reciprocity). Other works such as Yang et al. [2020a] separate the behaviours concerned with maximizing own rewards from those concerned with giving incentives to shape other agent’s learning. The choice regarding how to budget and structure the incentives, depending on the particular algorithm and environment being used, is still an open research problem.

Social Learning addresses a key issue in Reinforcement Learning: safe exploration. Since agents can learn to avoid unsafe states by observing the experience of others, Social Learning allows agents to manage the degree of risk they involve themselves in. Similarly, in a cooperative setting, being incentivized by others to avoid an unsafe action would help an agent avoid unnecessary actions that would hurt the welfare of the agent population.

Social Learning addresses another key issue in MARL: decentralization. By allowing agents to learn from each other, a centralized controller is no longer needed, greatly improving the scalability, flexibility and practicality of the algorithms. The paradigm of decentralized training has been addressed before [Zhang et al., 2019a], but without a mechanism designed to incentivize agents to maximize collective performance, there are difficulties in attaining high individual and collective return [Golembiewski and Olson, 1965].

Compelling directions for future research include understanding settings that encompass:

1. manipulation/deception
2. learning from others which pursue different goals
3. interleaving solitary with social learning
4. analysing the continuously modifying set of equilibria from the joint learning process
5. adapting the cost of the incentive function according to timely circumstances
6. better planning for the long-term effects of giving rewards
7. allowing agents to reject incentives based on the reputation of the giver

2.4 Social Dilemmas

In order to demonstrate the capacities of Social Learning, environments that stress the importance of collaboration are a key element in guiding this line of research. A Social Dilemma is a situation in which two concepts of rationality compete for the attention of each participant: their own individual rationality, concerned with their own well-being, and collective rationality, which is concerned with the well-being of everyone involved [Rapoport, 1974]. In a Social Dilemma, if all involved parties act in accordance with collective rationality, they are each better off than if they act upon individual rationality. The key challenge resides in the attractiveness of individual rationality above all else: it
Social Dilemmas can be modelled as general-sum matrix games, with properties that satisfy certain inequalities [Leibo et al., 2017], and they have been successfully applied to study a variety of phenomena in theoretical social science and biology [Trivers, 1971] [Barner-Brad, 1985] [Nowak and Sigmund, 1992]. However, by resorting to matrices, Social Dilemmas fail to capture a few critical aspects of real-world social dilemmas, the most important of which is the temporary nature [Leibo et al., 2017].

Formally, a Sequential Social Dilemma is a tuple \((\mathcal{M}, \Pi^C, \Pi^D)\) in which \(\mathcal{M}\) represents a Stochastic Game with a state space \(S\) and \(\Pi^C\) and \(\Pi^D\) are disjoint sets of policies that represent cooperative and defective behaviour. In general-sum matrix games, which are games defined on a matrix in which payoffs do not have restrictions and agents take actions simultaneously, there are four possible outcomes: \(R\) (reward for mutual cooperation), \(P\) (punishment from mutual defection), \(S\) (sucker for cooperating with a defector) and \(T\) (temptation for defecting against a cooperator). For state \(s \in S\), let the empirical matrix \((R(s), P(s), S(s), T(s))\), be the payoff matrix induced by following the policies \(\Pi^C\) and \(\Pi^D\). Then, the tuple \((\mathcal{M}, \Pi^C, \Pi^D)\) is an SSD when there exist states \(s \in S\) that induce a matrix that satisfies the following inequalities [Leibo et al., 2017]: \(R > P, R > S, R > \frac{T+S}{2}\) and either \(T > R\) or \(P > S\).

In order to address the above-mentioned issues, Sequential Social Dilemmas (SSD) have been proposed as a potential solution. An SSD is modelled as a general-sum simultaneous-move Markov game wherein the payoff matrix exhibits the same properties as a Social Dilemma [Leibo et al., 2017]. An important remark that is drawn from the properties of SSDs is that cooperativeness is not a binary property, but one which can be graded on a scale. The general-sum property of SSDs make them significantly more challenging than their zero-sum counterparts [Zinkevich et al., 2005].

A notable subset of SSDs are problems of common-pool resource (CPR) appropriation in which a common resource is shared between agents and it is difficult for them to exclude one another from accessing it. A couple of examples used in research include the game of Harvest [Perolat et al., 2019] and the game of Cleanup [Hughes et al., 2018], although pressing global issues such as climate change can also be framed in this way [Tavoni et al., 2011]. Successful algorithms that learn how to manage the common resources sustainably could therefore be applied to real-world policy-making [Caleiro et al., 2019].

### 2.5 Deception

While Social Learning offers numerous advantages when applied to SSDs, such as being a method that prevents the tragedy of the commons [Lupu and Precup, 2020], an open research problem identified so far has been the issue of preventing an agent from misusing their own incentive function. If an agent engages in deception tactics to mask their true intentions, their misuse could go undetected by the others. Being able to understand dishonesty and unethical behaviour is a crucial research avenue since machines that have reasons and capacity to deceive pose a serious threat to the relations between Artificial Intelligence and human society [Sarkadi et al., 2019].

Bond and Robinson [1988] define deception as a "false communication that tends to benefit the communicator". In the context of Social Learning and Sequential Social Dilemmas, deception can be thought of as a way for the communicator to establish a cooperative equilibrium that is sub-optimal from the perspective of total population welfare. Whales [1983] defines deception as a "distortion of perceived reality" which can be manifested in two different ways: hiding the reality or deliberately transmitting false information. For the purpose of this review, an agent may deceive others by hiding its true intentions or by deliberately convincing others that playing a certain sub-optimal strategy is in their best interest.

While deception has been studied within MARL [Asgharnia et al., 2020] [Bontrager et al., 2019] [Li et al., 2020] [Ghiya and Sycara, 2020], it is usually only applied on algorithms that allow limited influence capabilities between agents. The possibilities for deception enabled by Social Learning remain a largely unexplored problem and judging from current research that suggests influencing others significantly affects their behaviour [Jaques et al., 2019], a likely hypothesis is that Social Learning, although highly compelling, brings unprecedented challenges in regards to controlling and preventing deception.
3 Evidence of Deception in Social Learning

Recent developments provide evidence for the vulnerability of agents which rely on the signals of others to guide their own learning. While most research highlights the improvements created by social learning, it also exposes its unexplored risks. A non-exhaustive list of key papers is identified, and their premise, methodology, and conclusions are explained and analysed:

- **Yang et al. [2020]** provide evidence that learning from incentives can enable agents to achieve near-optimal collective rewards, in contrast with various baselines which led to sub-optimal behaviours. Furthermore, social learning as implemented in this paper enables decentralized training, as opposed to the baselines which are unable to find cooperative solutions. Despite the advantages, the authors note that an agent might misuse the incentive function to exploit others. This claim forms the basis of the argument that before this new capability becomes deployed, its safety must be explored and ensured.

- **Lupu and Precup [2020]** investigate how a gifting mechanism can enable agents to learn in social dilemmas and withstand robustly against the tragedy of the commons. The environment used, ‘Harvest’, spawns apples which agents compete for to collect. Apples regenerate based upon the number of remaining apples in the immediate vicinity. If agents do not learn how to collaborate in collecting apples sustainably, they will deplete the environment altogether, resulting in no future reward for any of them. Their methods enable agents to learn this type of collaboration, however, they also expose the need for a very large number of opportunities for altruism, something which can be unrealistic in real-world settings, and indicate the need for a mechanism to identify free-riders who exploit the altruism of the group.

- **Jaques et al. [2019]** discuss a unified mechanism for achieving coordination and cooperation by causal influence under counterfactual reasoning, and present evidence for its higher collective return under Sequential Social Dilemmas. Their results suggest that ‘influence’ rewards consistently lead to higher collective returns, making a case for their desirability. However, they note that their methods introduce the additional risk of deception, in ways which have not been researched before.

- **Hughes et al. [2018]** extend the idea of inequity aversion to agents interacting in SSDs and find that it mitigates the problem of ‘collective action’ (coordinating against socially deficient equilibria) and improves temporal credit assignment. However, the authors note that agents which exhibit guilt can be exploited by others that lack it. This result suggests that in the context of Social Learning, a population of agents needs to be homogeneous. If an agent is incapable of exhibiting a behaviour, such as deception, it should, therefore, be vulnerable against it. Hence, researching inequity aversion in heterogeneous teams should be a priority for protecting against deception.

- **Ndousse et al. [2020]** investigate whether independent reinforcement learning agents can take cues from the behaviour of experts in the environment. In this paper, expert agents shape novice’s learning process by augmenting their loss function to predict the expert’s future behaviour. While slightly different from Social Learning as defined in this review, the methods described by Ndousse et al. [2020] still allow agents to significantly affect each other’s future rewards, if the expert is knowingly demonstrating behaviour meant to deceive the novice into adopting sub-optimal policies. One of the key assumptions in this paper is that both the experts and the novices pursue the same goals. Future research directions should look at safely learning from experts that have divergent goals.

4 Open Problems

As seen in the previous section, Social Learning provides attractive properties which should lead to its wide-spread adoption in future Multi-Agent Systems. Therefore, it stands to reason that its exploitative properties have to be explored in order to facilitate safely deploying this technology in the future. The following is a list of open problems that should to be addressed in order to make significant progress in this regard:

- **Why is Social Learning desirable?** Most of the evidence provided in works such as Yang et al. [2020], Lupu and Precup [2020], Jaques et al. [2019] highlights the advantages
of Social Learning such as avoiding the tragedy of the commons in SSDs and achieving higher collective return. The evidence indicates that Social Learning creates decentralized, independent interactions which can greatly speed-up the emergence of desirable collective behaviour, such as cooperation and sustainably managing resources in CPR problems. These properties are highly sought after in Multi-Agent RL research, but the proposed solutions have no safety guarantees against deception.

• **What are the ways in which Social Learning can facilitate deception?** The evidence points to numerous ways in which an agent can deceive another one which trusts it enough to never reject its influence. For instance, in the SSD called 'Harvest', an agent with an unlimited reward-giving mechanism could deceive the other agents into adopting sub-optimal policies by gifting them to under-explore the environment. The agent engaging in deception would keep the common resource mostly to itself and would be free to even deplete it. Such a scenario would suggest that Social Learning presents safety concerns, since an agent that is deceived would take actions that could significantly reduce its own collective reward. For this reason, methods to ensure Safety in Social Learning are highly desirable.

• **What are the ways in which an agent can detect deception?** One way to detect whether an agent would engage in deception is by repeatedly probing its policy, classifying its behaviour into different types, and relying on the assumption that once identified, a transgressor will always remain one [Zhou et al., 2019]. However, such a method would be highly unethical, since it would require creating scenarios that are highly enticing for engaging in deception simply in order to see if an agent would fall into the temptation. Instead of actively probing an agent, another method would be to classify an agent as being deceptive or not based on an engagement via a cheap communication channel, which would be inconspicuous and less inquisitive [Azaria et al., 2014].

• **What are the ways in which an agent can protect itself from deception?** Hughes et al. [2018] suggest that inequity aversion should discourage an agent from engaging in deceptive tactics, if those tactics end up benefiting itself. An agent desiring to protect itself against deception should, therefore, punish those which gain an unfair advantage. However, this approach requires that an agent is able to distinguish whether another agent’s success was achieved through deception or not. It has been shown that a Markov Games can be carefully designed such that a single player can unilaterally manipulate the evolution of the game so that any other player can maximize their payoff only via global cooperation [Li and Hao, 2019]. Such a scenario could be used to establish good social norms and should disincentivize the formation of any colluding alliance looking to gain the upper hand through deception.

• **To which extent does social learning open up the risk of deception by malevolent agents?** Lin et al. [2020] argue that since cooperative Multi-Agent Reinforcement Learning algorithms have the potential to be included in critical infrastructure, researching their robustness to adversarial manipulation is necessary to their deployment in production. Critical infrastructure, such as fleets of self-driving cars or market trading bots, could stand to benefit tremendously from the benefits described by social learning. They are also attractive targets in cyber-attacks. The extent to which these algorithms can enable deception is, therefore, quite large, and the damage done by deploying them unsafely can be catastrophic.

5 Varying Assumptions

So far, this review has evaluated the evidence under the assumption that agents can reshape the reward function of others. However, other assumption are also worth investigating, since they might bring additional insight into dealing with deception in Social Learning. The literature on MARL is vast; for instance, the delicacy of collaborative equilibria has been studied before in settings such as Prisoner’s Dilemma [Badjatiya et al., 2020]. The exploitation of others has also been studied extensively in imperfect-information games [Ganzfried and Sun, 2018]. Non-stationarity is another key element of MARL, with recent surveys drawing attention to its importance [Papoudakis et al., 2019] [Hernandez-Leal et al., 2019]. Following is a list of pivotal assumptions that can change how deception is addressed, with potential directions for future research:

• **Assuming centralized training:** Lin et al. [2020] present the first work that investigates the adversarial exploitation of cooperative teams. By manipulating the observations of a sin-
gle agent, the results show that an entire team’s overall reward can degrade so drastically that their win rate drops from 98.9% to 0% in StarCraft II. While the algorithm used for testing the vulnerability hypothesis employs centralised training, which offers reduced manipulation opportunities compared to Social Learning, the results highlight the importance of protecting against deception and provides evidence for the potential consequences when failing to do so.

• Assuming the option to form teams: Ghiya and Sycara [2020] address both the challenge of learning collaborative policies in the presence of an adversary and provide evidence for the capacity of a group of agents to learn how to deceive an adversary. Despite their limitation of assuming centralised training, their methods provide an example of successful deception in multi-agent environments, strengthening the argument that and can form an incipient argument for investigating deception in Social Learning.

• Assuming that an agent can choose whom they interact with: Anastassacos et al. [2019] present evidence for the impact of partner selection in eventual cooperative outcomes. Their work suggests the importance of developing trust with cooperative agents and exposes the risks of being exploited by defecting agents. A malevolent agent could pretend to cooperate to deliberately gain the trust of the learning agent in order to retaliate more costly in the future. The authors note that detecting and avoiding such manipulative behaviours has not been investigated yet.

• Assuming imperfect information: Peysakhovich and Lerer [2018] present methods for learning cooperation in social dilemmas under the additional challenge of imperfect information, and explore the limitations of relying purely on past outcomes. Their work brings light on the differences between evaluating consequences and intentions, and could be a critical point in classifying deception.

• Assuming benevolence from a centralized coordinator: Baumann et al. [2020] present a mechanism designed to allow a planning agent to succeed in promoting significantly higher levels of cooperation in Social Dilemmas. However, the mechanism assumes benevolence on behalf of a planning agent, which distributes rewards and punishments according to a model about how other agent’s learning changes according to its own incentives. As designed, this mechanism could presents significant ethical issues if the models of the planning agent are not transparent and agents cannot refuse the planner’s rewards or punishments. Allowing a competition between planning agents could incentivize learning better predictive models, but at the same time, it could also become a competition of deception if each planning agent believes their own model is better, and the only way to enact their own policy is by deceiving the agents that decide between which candidate planning agent to follow.

• Assuming non-stationarity: Agents learning concurrently create a highly non-stationary joint policy. Deceptive behaviours might be circumstantial and methods that seek to prevent deception need to take this aspect into account. Papoudakis et al. [2019] and Hernandez-Leal et al. [2019] cast insights into dealing with non-nonstationarity in Multi-Agent Reinforcement Learning, which could be applied in future work when studying deception in Social Learning.

• Assuming information-asymmetry: Shen and How [2020] show how to construct robust policies in asymmetric imperfect-information games, in which a protagonist agent of a certain publicly known type has to infer the privately known type of an opponent agent. By learning an opponent model through self-play, a protagonist agent can robustly respond to a variety of different opponents. One can extend this work by considering the response of a protagonist trained using self-play against an opponent which is hiding whether if it is engaging in deception or not. The protagonist would learn how to robustly respond to deception only if it would be able to perfectly re-create it using self-play.

• Assuming self-play cannot create robust policies that protect against deception: Multi-Agent Reinforcement Learning algorithms that rely on self-play can create highly idiosyncratic policies which, although able to learn coordination policies with global optimum reward, do not generalize well to novel partners [Hu et al., 2020]. If robust strategies in a coordination game should exploit the presence of known asymmetries in the policy space, as shown in [Hu et al., 2020], then so should algorithms that learn how to protect against deception coming from novel players.
• Assuming cooperative agents have different goals: [Hostallero et al. 2020] describe how reward reshaping can promote semi-cooperation between agents that have different goals. Their proposed method gradually reshapes the rewards such that every agent’s perception of the equilibrium is geared towards optimizing for the total population welfare. The contribution is a mechanism design that, similar to Social Learning, allows agents to exchange peer evaluation signals which guide each-other’s myopic best-response strategies to a joint strategy that optimizes for global reward. Future research could investigate the effect of this mechanism at reducing the effectiveness of deception in a population of self-interested agents that use deception to promote their own, divergent goals.

6 Conclusion

Learning how to optimally behave in an environment populated by other intelligent actors is a difficult problem, both for humans and animals, as well as for Artificial Intelligence agents. Learning solely from the perspective of rewards gained by individually performed actions is a myopic strategy that fails to guide towards globally optimum rewards. Cooperation is a difficult equilibrium to achieve, especially if the dividends of coordination are not immediately apparent. Games designed to highlight the need for more socially responsible algorithms have been used to show the advantages created by Social Learning, a class of Multi-Agent Reinforcement Learning algorithms that allows agents to reshape each-other’s reward functions. However attractive, this new opportunity opens up questions concerning risks of manipulation, since by directly interfering with the learning process of another, an agent has unprecedented possibilities to deceive another agent into learning policies which it knows are not in that agent’s best interest.

To summarize the contributions, this research review motivates the problem statement, outlines the necessary background, introduces Social Learning and then defines both Social Dilemmas and Deception. Then, it identifies the evidence that shows Deception in Social Learning is an unexplored, yet important research area. Finally, this review identifies at least five critical open questions concerning Deception in Social Learning, and nine core assumptions that can shape how future solutions are implemented.

References

A. Agogino and Kagan Tumer. Unifying temporal and structural credit assignment problems. Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004., pages 980–987, 2004.

Yoav Alon and Huiyu Zhou. Multi-agent reinforcement learning for unmanned aerial vehicle coordination by multi-critic policy gradient optimization. arXiv preprint arXiv:2012.15472, 2020.

Nicolas Anastassacos, S. Hailes, and Mirco Musolesi. Understanding the impact of partner choice on cooperation and social norms by means of multi-agent reinforcement learning. ArXiv, abs/1902.03185, 2019.

A. Asgharnia, H. M. Schwartz, and M. Atia. Deception in the game of guarding multiple territories: A machine learning approach. In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 381–388, 2020. doi: 10.1109/SMC42975.2020.9283173.

Amos Azaria, A. Richardson, and S. Kraus. An agent for deception detection in discussion based environments. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2014.

Pinkesh Badjatiya, Mausoom Sarkar, A. Sinha, Siddharth Singh, Nikaash Puri, Jayakumar Subramanian, and Balaji Krishnamurthy. Inducing cooperative behaviour in sequential-social dilemmas through multi-agent reinforcement learning using status-quo loss. arXiv: Artificial Intelligence, 2020.

Carol Barner-Barry. The evolution of cooperation robert axelrod new york: Basic books, inc., 1984. Politics and the Life Sciences, 4:76–76, 1985.

T. Baumann, T. Graepel, and J. Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation. 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–7, 2020.
Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957. ISSN 00959057, 19435274. URL [http://www.jstor.org/stable/24900506](http://www.jstor.org/stable/24900506).

C. Bond and M. Robinson. The evolution of deception. *Journal of Nonverbal Behavior*, 12:295–307, 1988.

Philip Bontrager, Ahmed Khalifa, D. Anderson, Matthew Stephenson, C. Salge, and J. Togelius. Superstition in the network: Deep reinforcement learning plays deceptive games. *ArXiv*, abs/1908.04436, 2019.

R. Boyd, P. Richerson, and J. Henrich. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108:10918 – 10925, 2011.

Daniel S. Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences. *arXiv preprint arXiv:2002.09089*, 2020.

Noam Brown and T. Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359:418 – 424, 2018.

Noam Brown and T. Sandholm. Superhuman ai for multiplayer poker. *Science*, 365:885 – 890, 2019.

L. Busoni, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008. doi: 10.1109/TSMCC.2007.913919.

A. Caleiro, M. Sousa, and Ingo Andrade de Oliveira. Global development and climate change: A game theory approach. 2019.

Gang Chen. A new framework for multi-agent reinforcement learning – centralized training and exploration with decentralized execution via policy distillation. *arXiv preprint arXiv:1910.09152*, 2019.

S. Chen, Ying-Fang Kao, and R. Venkatachalam. Computational behavioural economics. 2017.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020.

Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS ’14, page 165–172, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450327381.

Tom Everitt and Marcus Hutter. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *arXiv preprint arXiv:1908.04734*, 2019.

S. Ganzfried and Q. Sun. Bayesian opponent exploitation in imperfect-information games. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8, 2018. doi: 10.1109/CIG.2018.8490452.

Qitong Gao, Davood Hajinezhad, Yan Zhang, Yiannis Kantaros, and Michael M. Zavlanos. Reduced variance deep reinforcement learning with temporal logic specifications. In *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, ICCPS ’19, page 237–248, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362856. doi: 10.1145/3302509.3311053. URL [https://doi.org/10.1145/3302509.3311053](https://doi.org/10.1145/3302509.3311053).

Siddharth Ghiya and K. Sycara. Learning complex multi-agent policies in presence of an adversary. *ArXiv*, abs/2008.07698, 2020.

Mevludin Glavic, Raphael Fonteneau, and Damien Ernst. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC-PapersOnLine*, 50(1):6918 – 6927, 2017. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2017.08.1217. URL [http://www.sciencedirect.com/science/article/pii/S2405896317317238](http://www.sciencedirect.com/science/article/pii/S2405896317317238). 20th IFAC World Congress.
R. Golembiewski and M. Olson. The logic of collective action. 1965.

M. Grzes. Reward shaping in episodic reinforcement learning. In AAMAS, 2017.

Christopher J. Hazard. Trust and Reputation in Multiagent Systems: Strategies and Dynamics with Reference to Electronic Commerce. PhD thesis, 2010. AA13425980.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity, 2019.

D. Hostallero, Dae-Woo Kim, Sangwoo Moon, Kyunghwan Son, W. Kang, and Y. Yi. Inducing cooperation through reward reshaping based on peer evaluations in deep multi-agent reinforcement learning. In AAMAS, 2020.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. Other-play for zero-shot coordination, 2020.

E. Hughes, Joel Z. Leibo, Matthew G. Philips, K. Tuyls, Edgar A. Duñez-Guzmán, A. Castañeda, Iain Dunning, T. Zhu, Kevin R. McKee, R. Koster, Heather Roff, and T. Graepel. Inequity aversion resolves intertemporal social dilemmas. ArXiv, abs/1803.08884, 2018.

M. M. Islam and E. Al-Shaer. Active deception framework: An extensible development environment for adaptive cyber deception. In 2020 IEEE Secure Development (SecDev), pages 41–48, 2020. doi: 10.1109/SecDev45635.2020.00023.

N. Jaques, A. Lazaridou, E. Hughes, Çağlar Gülçehre, Pedro A. Ortega, D. Strouse, Joel Z. Leibo, and N. D. Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In ICML, 2019.

K. Laland. Social learning strategies. Animal Learning & Behavior, 32:4–14, 2004.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv:1702.03037, 2017.

Chunmao Li, Xuanguang Wei, Yinliang Zhao, and Xupeng Geng. An effective maximum entropy exploration approach for deceptive game in reinforcement learning. Neurocomputing, 403:98 – 108, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.04.068. URL http://www.sciencedirect.com/science/article/pii/S0925231220306536

K. Li and Dong Hao. Cooperation enforcement and collusion resistance in repeated public goods games. In AAAI, 2019.

J. Lin, Kristina Dzeparoska, S. Zhang, A. Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning. ArXiv, abs/2003.03722, 2020.

Ryan Lowe, Yi Wu, A. Tamar, J. Harb, P. Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In NIPS, 2017.

Andrei Lupu and Doina Precup. Gifting in multi-agent reinforcement learning. In AAMAS, 2020.

Laëtitia Matignon, G. Laurent, and N. L. Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. Knowl. Eng. Rev., 27:1–31, 2012.

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

V. Mnih, K. Kavukcuoglu, D. Silver, Andrei A. Rusu, J. Veness, Marc G. Bellemare, A. Graves, Martin A. Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, S. Petersen, C. Beattie, A. Sadik, Ioannis Antonoglou, H. King, D. Kumaran, Daan Wierstra, S. Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. Nature, 518:529–533, 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:undefined*, 2013.

Mohammadreza Nazari, Afshin Oroojlooy, Lawrence V. Snyder, and Martin Takáč. Reinforcement learning for solving the vehicle routing problem. *arXiv preprint arXiv:1802.04240*, 2018.

Kamal Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Learning social learning, 2020.

Thanh Thi Nguyen and Vijay Janapa Reddi. Deep reinforcement learning for cyber security. *arXiv preprint arXiv:1906.05799*, 2019.

M. Nowak and K. Sigmund. Tit for tat in heterogeneous populations. *Nature*, 355:250–253, 1992.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik’s cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Afshin OroojlooyJadid and Davood Hajimezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V. Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

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. *arXiv preprint arXiv:1707.06600*, 2017.

Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information, 2018.

A. Rapoport. Prisoner’s dilemma — recollections and observations. 1974.

Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter John McBurney, Simon Dominic Parsons, and Martin David Chapman. Modelling deception using theory of mind in multi-agent systems. *AI COMMUNICATIONS*, 32(4):287–302, 2019. ISSN 0921-7126. doi: 10.3233/AIC-190615.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

S. Shalev-Shwartz, Shaked Shamhah, and A. Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv*, abs/1610.03295, 2016.

Macheng Shen and Jonathan P. How. Robust opponent modeling via adversarial ensemble reinforcement learning in asymmetric imperfect-information games, 2020.

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. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2006.02.006. URL http://www.sciencedirect.com/science/article/pii/S0004370207000495. Foundations of Multi-Agent Learning.

Michael Slum, Max Kleiman-Weiner, Michael L. Littman, and Joshua B. Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. *arXiv preprint arXiv:1901.06085*, 2019.

Xinliang Song, Tonghan Wang, and Chongjie Zhang. Convergence of multi-agent learning with a finite step size in general-sum games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS ’19*, page 935–943, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.
Alessandro Tavoni, Astrid Dannenberg, Giorgos Kallis, and Andreas Löschel. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, 108(29):11825–11829, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1102493108. URL https://www.pnas.org/content/108/29/11825.

R. Trivers. The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46:35 – 57, 1971.

B. Whaley. Toward a general theory of deception. *Journal of Strategic Studies*, 5:178–192, 1982.

Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents, 2020a.

Jiachen Yang, A. Nakhaei, David Isele, H. Zha, and K. Fujimura. Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. *ICLR 2020 Poster*, abs/1809.05188, 2020b.

Luyao Yuan, Zipeng Fu, Linqi Zhou, Kexin Yang, and Song-Chun Zhu. Emergence of theory of mind collaboration in multiagent systems. 2019.

K. Zhang, Zhuoran Yang, and T. Başar. Decentralized multi-agent reinforcement learning with networked agents: Recent advances. *ArXiv*, abs/1912.03821, 2019a.

K. Zhang, Zhuoran Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *ArXiv*, abs/1911.10635, 2019b.

Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H. Ballard, and Peter Stone. Leveraging human guidance for deep reinforcement learning tasks. *arXiv preprint arXiv:1909.09906*, 2019c.

Z. Zhang and D. Wang. Eaqr: A multiagent q-learning algorithm for coordination of multiple agents. *Complex.*, 2018:7172614:1–7172614:14, 2018.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, and Richard Socher. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. *arXiv preprint arXiv:1804.06459*, 2018.

Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, Aurora Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Dong Chen, Zhengbang Zhu, Nhat Nguyen, Mohamed Elsayed, Kun Shao, Sanjeevan Ahilan, Baokuan Zhang, Jiannan Wu, Zhengang Fu, Kasra Rezaee, Peyman Yadmelat, Mohsen Rohani, Nicolás Perez Nieves, Yihan Ni, Seyyedeh Baniamali, Alexander Cowen Rivers, Zheng Tian, Daniel Palencic, Haitham bou Ammar, Hongbo Zhang, Wulong Liu, Jianye Hao, and Jun Wang. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *arXiv preprint arXiv:2010.09776*, 2020.

Wenxuan Zhou, Lerrel Pinto, and Abhinav Gupta. Environment probing interaction policies. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryl8-3acFX.

Q. Zhu and J. Oh. Deep reinforcement learning for fairness in distributed robotic multi-type resource allocation. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 460–466, 2018. doi: 10.1109/ICMLA.2018.00075.

Martin Zinkevich, A. Greenwald, and M. Littman. Cyclic equilibria in markov games. In *NIPS*, 2005.