On the Agenda(s) of Research on Multi-Agent Learning

Yoav Shoham and **Rob Powers** and **Trond Grenager**
Computer Science Department
Stanford University
Stanford, CA 94305
{shoham, powers, grenager}@cs.stanford.edu

Abstract

We survey the recent work in AI on multi-agent reinforcement learning (that is, learning in stochastic games). After tracing a representative sample of the recent literature, we argue that, while exciting, much of this work suffers from a fundamental lack of clarity about the problem or problems being addressed. We then propose five well-defined problems in multi-agent reinforcement learning and single out one that in our view is both well-suited for AI and has not yet been adequately addressed. We conclude with some remarks about how we believe progress is to be made on this problem.

Introduction

Reinforcement learning (RL) has been an active research area in AI for many years. Recently there has been growing interest in extending RL to the multi-agent domain. From the technical point of view, this has taken the community from the realm of Markov Decision Problems (MDPs) to the realm of game theory, and in particular stochastic (or Markov) games (SGs).

The body of work in AI on multi-agent RL is still small, with only a couple of dozen papers on the topic as of the time of writing. This contrasts with the literature on single-agent learning in AI, as well as the literature on learning in game theory – in both cases one finds hundreds if not thousands of articles, and several books. Despite the small number we still cannot discuss each of these papers. Instead we will trace a representative historical path through this literature. We will concentrate on what might be called the "Bellman heritage" in multi-agent RL – work that is based on Q-learning (Watkins & Dayan 1992), and through it on the Bellman equations (Bellman 1957). Specifically, we will discuss (Littman 1994; Claus & Boutilier 1998; Hu & Wellman 1998; Bowling & Veloso 2001; Littman 2001; Greenwald, Hall, & Serrano 2002), and in the course of analyzing these papers will mention several more.

In the next section we trace the "Bellman heritage", and summarize the results obtained there. These results are unproblematic for the cases of zero-sum SGs and common-payoff (aka 'team', or pure-coordination) SGs, but the attempt to extend them to general-sum SGs is problematic. In section 3 we trace back the technical awkwardness of the results to what we view as a misguided focus on the Nash equilibrium as an ingredient in both the learning algorithm and the evaluation criterion. But we believe the problem runs deeper and has to do with a basic lack of clarity about the exact problem being addressed. In section 4 we argue that there are (at least) five distinct well-defined problems to be addressed, and attempt to map the existing work into these categories. We identify one of the five that we feel is the most interesting for AI, and note that it has barely been addressed in that line of research. Finally, in section 5 we make some comments on how we think one might go about tackling it.

Bellman's Heritage in Multi-Agent RL

In this section we review a representative sample of the literature. We start with the algorithms, and then summarize the results reported.

Throughout, we use the following terminology and notation. An (n-agent) stochastic game (SG) is a tuple $(N, S, \vec{A}, \vec{R}, T)$. N is a set of agents indexed $1, \ldots, n$. S is a set of n-agent stage games (usually thought of as games in normal form, although see (Jehiel & Samet 2001) for an exception). $\vec{A} = A_1, \ldots, A_n$, with A_i the set of actions (or pure strategies) of agent i (note that we assume the agent has the same strategy space in all games; this is a notational convenience, but not a substantive restriction). $\vec{R} = R_1, \ldots, R_n$, with $R_i : S \times \vec{A} \rightarrow \mathcal{R}$ the immediate reward function of agent i. $T : S \times \vec{A} \rightarrow \Pi(S)$ is a stochastic transition function, specifying the probability of the next game to be played based on the game just played and the actions taken in it. A Markov Decision Problem (MDP) is a 1-agent SG; an MDP thus has the simpler structure (S, A, R, T).

From Minimax-Q to Nash-Q and Beyond

We start with the *(single-agent) Q-learning* algorithm (Watkins & Dayan 1992) for computing an optimal policy

in an MDP with unknown reward and transition functions:[1]

$$\begin{aligned} Q(s,a) &\leftarrow (1-\alpha)Q(s,a) + \alpha[R(s,a) + \gamma V(s')] \\ V(s) &\leftarrow \max_{a \in A} Q(s,a) \end{aligned}$$

As is well known, with certain assumptions about the way in which actions are selected at each state over time, Q-learning converges to the optimal value function V^*.

The simplest way to extend this to the multi-agent SG setting is just to add a subscript to the formulation above; that is, to have the learning agent pretend that the environment is passive:

$$\begin{aligned} Q_i(s,a_i) &\leftarrow (1-\alpha)Q_i(s,a_i) + \alpha[R_i(s,\vec{a}) + \gamma V_i(s')] \\ V_i(s) &\leftarrow \max_{a_i \in A_i} Q_i(s,a_i) \end{aligned}$$

Several authors have tested variations of this algorithm (e.g., (Sen, Sekaran, & Hale 1994)). However, this approach ignores the multi-agent nature of the setting entirely. The Q-values are updated without regard for the actions selected by the other agents. While this can be justified when the opponents' choices of actions are stationary, it fails when an opponent may adapt its choice of actions based on the past history of the game.

A cure to this problem is to define the Q-values as a function of all agents' actions:

$$Q_i(s,\vec{a}) \leftarrow (1-\alpha)Q_i(s,\vec{a}) + \alpha[R_i(s,\vec{a}) + \gamma V_i(s')]$$

We are left with the question of how to update V, given the more complex nature of the Q-values.

For (by definition, two-player) zero-sum SGs, Littman suggests the *minimax-Q* learning algorithm, in which V is updated with the minimax of the Q values (Littman 1994):

$$V_1(s) \leftarrow \max_{P_1 \in \Pi(A_1)} \min_{a_2 \in A_2} \sum_{a_1 \in A_1} P_1(a_1) Q_1(s,(a_1,a_2)).$$

Although it can be extended to general-sum SGs, minimax-Q is no longer well motivated in those settings. One alternative is to try to explicitly maintain a belief regarding the likelihood of the other agents' policies, and update V based on the induced expectation of the Q values:

$$V_i(s) \leftarrow \max_{a_i} \sum_{a_{-i} \in A_{-i}} P_i(s,a_{-i}) Q_i(s,(a_i,a_{-i})).$$

This approach, which is in the spirit of the belief-based procedures in game theory such as *fictitious play* (Brown 1951) and *rational learning* (Kalai & Lehrer 1993), is pursued by Claus and Boutilier (Claus & Boutilier 1998). In this work their joint-action learners specifically adopt the belief-maintenance procedures of fictitious play, in which the probability of a given action in the next stage game is assumed to be its past empirical frequency. Although this procedure is well defined for any general-sum game, Claus and Boutilier only consider it in the context of common-payoff (or 'team') games. A stage game is common-payoff if at each outcome all agents receive the same payoff. The payoff is in general different in different outcomes, and thus the agents' problem is that of coordination; indeed these are also called *games of pure coordination.*

Zero-sum and common-payoff SGs have very special properties, and, as we discuss in the next section, it is relatively straightforward to understand the problem of learning in them. The situation is different in general-sum games, which is where the picture becomes less pretty. An early contribution here is *Nash-Q* learning (Hu & Wellman 1998), another generalization of Q-learning to general-sum games. Nash-Q updates the V-values based on some Nash equilibrium in the game defined by the Q-values:

$$V_i(s) \leftarrow \text{Nash}_i(Q_1(s,\vec{a}),\ldots,Q_n(s,\vec{a})).$$

There is some abuse in the above notation; the expression represents a game in which $Q_i(s,\vec{a})$ denotes the payoff matrix to player i, and Nash_i denotes "the" Nash payoff to that player.

Of course in general there are many Nash equilibria, and therefore the Nash payoff may not be unique. If Nash-Q is taken to apply to all general-sum SGs, it must be interpreted as a nondeterministic procedure. However, the focus of Hu and Wellman has been again on a special class of SGs. Littman articulated it most explicitly, by reinterpreting Nash-Q as the *Friend-or-Foe (FoF)* algorithm(Littman 2001). Actually, it is more informative to view FoF as two algorithms, each applying in a different special class of SGs. The Friend class consists of SGs in which, throughout the execution of the algorithm, the Q-values of the players define a game in which there is a globally optimal action profile (meaning that the payoff to any agent under that joint action is no less than his payoff under any other joint action). The Foe class is the one in which (again, throughout the execution of the algorithm), the Q-values define a game with a saddle point. Although defined for any number of players, for simplicity we show how the Vs are updated in a two-player game:

$$\begin{array}{lll} \text{Friend:} & V_1(s) \leftarrow & \max_{a_1 \in A_1, a_2 \in A_2} Q_1(s,(a_1,a_2)) \\ \text{Foe:} & V_1(s) \leftarrow & \max_{P_1 \in \Pi(A_1)} \min_{a_2 \in A_2} \\ & & \sum_{a_1 \in A_1} P_1(a_1) Q_1(s,(a_1,a_2)) \end{array}$$

Thus Friend-Q updates V similarly to regular Q-learning, and Foe-Q updates as does minimax-Q.

Finally, Greenwald et al.'s *CE-Q* learning is similar to Nash-Q, but instead uses the value of a correlated equilibrium to update V (Greenwald, Hall, & Serrano 2002):

$$V_i(s) \leftarrow \text{CE}_i(Q_1(s,\vec{a}),\ldots,Q_n(s,\vec{a})).$$

Like Nash-Q, it requires agents to select a unique equilibrium payoff, an issue that the authors address explicitly by suggesting several possible selection mechanisms.

Convergence Results

In the work referenced above, the main criteria used to measure the performance of the above algorithms was its ability to converge to an equilibrium in self-play. In (Littman

[1]This procedure is based directly on the Bellman equations (Bellman 1957) and the dynamic programming procedures based on them for MDPs with known reward and transition functions.

& Szepesvari 1996) minimax-Q learning is proven to converge in the limit to the correct Q-values for any zero-sum game, guaranteeing convergence to a Nash equilibrium in self-play. These results make the standard assumptions of infinite exploration and the conditions on learning rates used in proofs of convergence for single-agent Q-learning. Claus and Boutilier (Claus & Boutilier 1998) conjecture that both independent Q-learners and the belief-based joint action learners mentioned above will converge to an equilibrium in common payoff games under the conditions of self-play and decreasing exploration, but do not offer a formal proof. Nash-Q learning was shown to converge to the correct Q-values for the classes of games defined earlier as Friend games and Foe games.[2] Finally, *CE-Q* learning is shown to converge to Nash equilibria (a subset of the set of correlated equilibria) in a number of empirical experiments, although there are no formal results presented.

Why Focus on Equilibria?

In the previous section we summarized the developments in multi-agent RL without editorial comments. Here we begin to discuss that work more critically.

The results concerning convergence of Nash-Q are quite awkward. Nash-Q attempted to treat general-sum SGs, but the convergence results are constrained to the cases that bear strong similarity to the already known cases of zero-sum games and common-payoff games. The analysis is interesting in that it generalizes both conditions: The existence of a saddle point is guaranteed in, but not limited to, zero-sum games, and the existence of a globally optimal Nash equilibrium payoff is guaranteed in, but not limited to, common-payoff games. However, the conditions on the payoffs are quite restrictive, since they must hold for the games defined by the intermediate Q-values throughout the execution of the protocol. This makes it hard to find any natural classes of games that satisfy these properties beyond the two special cases, and also difficult to verify at the outset if a given game satisfies the properties.

Note that like the original work on single agent Q-learning, Nash-Q concentrates on learning the correct Q-values, in this case for a Nash equilibrium of the game. However, it is not obvious how to turn this into a procedure for guiding play beyond zero-sum games. If multiple optimal equilibria exist the players need an oracle to coordinate their choices in order for play to converge to a Nash equilibrium, which begs the question of why to use learning for coordination at all.

In our view, these unsatisfying aspects of the Bellman heritage from Nash-Q onwards – the weak/awkward convergence assurances, the limited applicability, the assumption of an oracle – manifest a deeper set of issues. Many of these can be summarized by the following question: What justifies the focus on (e.g., Nash) equilibrium?

[2] A certain local debate ensued regarding the initial formulation of these results, which was resolved in the papers by Bowling (Bowling 2000), Littman (Littman 2001), and by Hu and Wellman themselves in the journal version of their article (Hu & Wellman 2002).

Nash-Q appeals to the Nash equilibrium in two ways. First, it uses it in the execution of the algorithm. Second, it uses convergence to it as the yardstick for evaluating the algorithm. The former is troubling in several ways:

1. Unlike the max-min strategy, employed in minimax-Q, a Nash-equilibrium strategy has no prescriptive force. At best the equilibrium identifies conditions under which learning can or should stop (more on this below), but it does not purport to say anything prior to that.
2. One manifestation of the lack of prescriptive force is the existence of multiple equilibria; this is a thorny problem in game theory, and limiting the focus to games with a uniquely identified equilibrium – or assuming an oracle – merely sweeps the problem under the rug.
3. Finally, the argument for playing an equilibrium strategy in many games often seems dependent on the rather circular assumption that one's opponents will also seek an equilibrium strategy. While one might be able to justify such an assumption if players had unbounded computational ability, even calculating a Nash equilibrium for a large game can prove intractable.

Beside being concerned with the specific details of Nash-Q and its descendants, we are also concerned with the use of convergence to Nash equilibrium as the evaluation criterion. Bowling and Veloso articulate this yardstick most clearly (Bowling & Veloso 2001). They put forward two criteria for any learning algorithm in a multi-agent setting: (1) The learning should always converge to a stationary policy, and (2) it should only terminate with a best response to the play by the other agent(s) (a property called Hannan-consistency in game theory (Hannan 1959)). In particular, their conditions require that during self-play, learning only terminate in a stationary Nash equilibrium. This is a useful criterion, but it seems weak in that it ignores the fact that one is playing an extended SG.[3] We again confront the centrality of the Nash equilibrium to game theory, and the question of whether it should play the same central role in AI. We return to this in the next section, but briefly, in our view the answer is no.

Five Well-Defined Agendas in Multi-Agent Learning

In our view the root of the difficulties with the recent work is that the field has lacked a clearly defined problem statement. In this section we identify what we think is a coherent research agenda on multi-agent RL. In fact, we generously offer five such agendas. We also identify one of them as being, in our view, the most appropriate for AI, and the most heretical from the game theoretic point of view.

The first agenda is descriptive – it asks how humans learn in the context of other learners (see, e.g., (Erev & Roth 1998; Camerer, Ho, & Chong 2002)). The name of the game

[3] It should be said that the literature on learning in game theory (mostly in repeated games, a special case of SGs) revolves almost entirely around the question of whether this or that learning procedure leads to a Nash equilibrium. In our opinion game theory is also unclear on its motivation in doing so. We comment on this in the next section, but this is not our focus in this article.

here is to show experimentally that a certain formal model of learning agrees with people's behavior (typically, in laboratory experiments). This work is typically undertaken by psychologists, experimental game theorists, or other experimentally-inclined social scientists.

The second agenda is computational in nature. It views learning algorithms as an iterative way to compute solution concepts. Fictitious play was originally proposed as a way of computing a sample Nash equilibrium, and other adaptive procedures have been proposed more recently for computing other solution concepts (for example, computing equilibria in local-effect games (Leyton-Brown & Tennenholtz 2003)). These tend not to be the most efficient computation methods, but they do sometimes constitute quick-and-dirty methods that can easily be understood and implemented.

The above two agendas are often intertwined within much of the work on learning in game theory as researchers propose various dynamics that are perceived as plausible in one sense or another (often by explaining human behavior), and proceed to investigate whether those converge to equilibria. This is a key concern for game theory, since a successful theory would support the notion of Nash (and other kinds of) equilibrium, which plays a central role in non-cooperative game theory.[4] The main limitation of this line of research is that, as of now, there is no agreed-upon objective criterion by which to judge the reasonableness of any given dynamics.

The other three agendas are prescriptive. They ask how agents – people or programs - *should* learn. The first of these involves distributed control in dynamic systems. There is sometimes a need or desire to decentralize the control of a system operating in a dynamic environment, and in this case the local controllers must adapt to each other's choices. This direction, which is most naturally modeled as a repeated or stochastic common-payoff (or 'team') game, has attracted much attention in AI in recent years. Proposed approaches can be evaluated based on the value achieved by the joint policy and the resources required, whether in terms of computation, communication, or time required to learn the policy. In this case there is no role for equilibrium analysis; the agents have no freedom to deviate from the prescribed algorithm. Researchers interested in this agenda have access to a large body of existing work both within AI and other fields such as control theory and distributed processing/computation.

The two remaining prescriptive agendas both assume that the learning takes place by self-interested agents. To understand the relationship between these two agendas, it is worthwhile to explicitly note the following obvious fact: reinforcement learning – whether in a single- or multi-agent setting – is nothing but a specific form of acting in which the actions are conditioned on runtime observations about the world. Thus the question of "how best to learn" is a specialized version of the general question "how best to act".

The two remaining prescriptive agendas diverge on how they interpret 'best'. We call the first the 'equilibrium agenda'. Although one could have expected a game theory purist to adopt this perspective, it differs from what is commonly studied in game theory, and in fact is explicitly rejected in at least one place (Fudenberg & Kreps 1993). We have only seen it pursued recently, outside game theory (Tennenholtz 2002). The agenda can be described as follows. Since, in the traditional view of non-cooperative game theory, the notion of optimal strategy is meaningless and is replaced by the notions of best response and (predominantly, Nash) equilibrium, and since a learning strategy is after all just a strategy in an extended game, one should ask when a vector of learning strategies (one for each agent) forms an equilibrium. Of course, for this to be meaningful, one has to be precise about the game being played – including the payoff function and the information structure. In particular, in the context of SGs, one has to specify whether the aggregate payoff to an agent is the limit average, the sum of future discounted rewards, or something else. The focus of this agenda would most naturally seem to focus on identifying what classes of learning strategies form an equilibria for different classes of stochastic games.

The final prescriptive agenda is one that we shall call the 'AI agenda', pending a more descriptive title. Again the name could be viewed as a bit ironic since for the most part it is not the approach taken in AI, but we do believe it is the one that makes the most sense for the field. This agenda might seem somewhat less than glamorous; it asks what the best learning strategy is for a given agent *for a fixed class of the other agents in the game*. It thus retains the design stance of AI, asking how to design an optimal (or at least effective) agent for a given environment. It just so happens that this environment is characterized by the types of agents inhabiting it. This does raise the question of how to parameterize the space of environments, and we return to that in the next section. The objective of this agenda is to identify effective strategies for environments of interest. A more effective strategy is one that achieves a better payoff in its environment, the selected class of opponents. The class of opponents should itself be motivated as being reasonable and containing problems of interest. Convergence to an equilibrium is valuable if and only if it serves the goal of maximizing payoff (again we need to be careful when discussing the payoff for a stochastic game to specify how to aggregate the payoffs from the individual matrix games).

We should say that the 'AI agenda' is in fact not as alien to past work in multi-agent RL in AI as our discussion implies. While most of the work cited earlier concentrates on comparing convergence rates between algorithms in self-play, we can see some preliminary analysis comparing the performance of algorithms in environments consisting of other learning agents (e.g. (Hu & Wellman 2001; 2002; Stone & Littman 2001)) However, these experimental strands were not tied to a formal research agenda, and in particular not to the convergence analyses. One striking exception is the work by Chang and Kaelbling (Chang & Kaelbling 2001), to which we return in the next section.

The 'AI agenda', however, is quite antithetical to the prevailing spirit of game theory. This is precisely because it adopts the 'optimal agent design' perspective and does not

[4] It has been noted that game theory is somewhat unusual, if not unique, in having the notion of an equilibrium without associated dynamics that give rise to the equilibrium.

consider the equilibrium concept to be central or even necessarily relevant at all. The essential divergence between the two approaches lies in their attitude towards 'bounded rationality'. Traditional game theory assumed it away at the outset, positing perfect reasoning and infinite mutual modeling of agents. It has been struggling ever since with ways to gracefully back off from these assumptions when appropriate. It's fair to say that despite notable exceptions (cf., (Rubinstein 1998)), bounded rationality is a largely unsolved problem for game theory. In contrast, the AI approach embraces bounded rationality as the starting point, and only adds elements of mutual modeling when appropriate. The result is fewer elegant theorems in general, but perhaps a greater degree of applicability in certain cases. This applies in general to situations with complex strategy spaces, and in particular to multi-agent learning settings.

It should be said that although the "equilibrium agenda" and the "AI agenda" are quite different, there are still some areas of overlap once one looks more closely. First, as we discuss in the next section, in order to parameterize the space of environments one must start to grapple with traditional game theoretic notions such as type spaces. Furthermore, when one imagines how learning algorithms might evolve over time, one can well imagine that the algorithms evolve towards an equilibrium, validating the 'game theory agenda' after all. While this may advise thoughts about the long-term outcome of such evolution, it stills provides no guidance for how to behave in the short-term, prior to such a convergence.

The case of the Trading Agent Competition (TAC) serves to illustrate the point. TAC (Wellman & Wurman 1999) is a series of competitions in which computerized agents trade in a non-trivial set of interacting markets. You would think that the TAC setting would allow for application of game theoretic ideas. In fact, while the teams certainly gave thought to how other teams might behave – that is, to their class of opponents – the programs engaged in no computation of Nash equilibria, no modeling of the beliefs of other agents, nor for the most part any sophisticated attempts to send specific signals to the other agents. The situation was sufficiently complex that programs concentrated on simpler tasks such as predicting future prices in the different markets, treating them as external events as opposed to something influenced by the program itself. One could reasonably argue that after each competition each team will continue to improve its TAC agent, and eventually the agents will settle on an equilibrium of learning strategies. Although we believe this to be true in principle, this argument is only compelling when the game is fairly simple and/or is played over a long time horizon. For TAC the strategy space is so rich that this convergence is unlikely to happen in our lifetime. In any case, it provides no guidance on how to win the next competition.

Before we say a few words about the 'AI agenda', let us reconsider the "Bellman heritage" discussed earlier; how does it fit into this categorization? Minimax-Q can be fit into the 'AI agenda', for the highly specialized case of zero-sum games and the objective of minimizing the worst case payoff against the set of all possible opponents. The work on self-play in common-payoff SGs, although superficially reminiscent of the 'AI agenda', probably fits better with the 'DAI agenda', with the payoff function interpreted as the payoff of the agents' designer. In general, when evaluating performance in self-play, a separate argument would seem to be required as to why that would be a reasonable class of opponents to expect. Nash-Q and its descendants feel somewhat like followers of the 'equilibrium agenda' although for a restricted set of the equilibria in stochastic games. They fail to resonate with the 'AI agenda' since it is unclear what class of environments they might achieve a good payoff within.

Pursuing the 'AI agenda'

The 'AI agenda' calls for categorizing strategic environments, that is, populations of agent types with which the agent being designed might interact. These agent types may come with a distribution over them, in which case one can hope to design an agent with maximal expected payoff, or without such a distribution, in which case a different objective is called for (for example, an agent with maximal minimum payoff). In either case we need a way to speak about agent types. The question is how to best represent meaningful classes of agents, and then use this representation to calculate a best response.

We won't have much to say about the best-response calculation, except to note that it is computationally a hard problem. For example, it is known that in general the best response in even a two-player SG is non-computable (Nachbar & Zame 1996). We will however touch on the question of how to parameterize the space of agents, which is itself a challenge. Our objective is not to propose a specific taxonomy of agent types, but instead to provide guidance for the construction of useful taxonomies for different settings.

Agents are categorized by their strategy space. Since the space of all strategies is complex, this categorization is not trivial. One coarse way of limiting a strategy space is to simply restrict it to a family. For example, we might assume that the agent belongs to the class of joint-action learners in the sense of (Claus & Boutilier 1998). Another, in principle orthogonal, way of restricting the strategy space is to place computational limitations on the agents. For example, we might constrain them to be finite automata with a bounded number of states.[5] Even after these kinds of limitations we might still be left with too large a space to reason about, but there are further disciplined approaches to winnowing down the space. In particular, when the strategies of the opponent are a function of its beliefs, we can make restricting assumptions about those beliefs. This is the approach taken by Chang and Kaelbling (Chang & Kaelbling 2001), and to some extent (Stone & Littman 2001), although they both look at a rather limited set of possible strategies and beliefs. A more general example would be to assume that the opponent is a 'rational learner' in the sense of (Kalai & Lehrer 1993), and to place restrictions on its prior about

[5]This is the model commonly pursued in the work on 'bounded rationality' (e.g., (Neyman 1985; Papadimitriou & Yannakakis 1994; Rubinstein 1998)). Most of that work however is concerned with how equilibrium analysis is impacted by these limitations, so it's not clear whether the technical results obtained there will directly contribute to the 'AI agenda.'

other agents' strategies. Note though that this is a slippery slope, since it asks not only about the second agent's computational limitations and strategy space, but also recursively about his beliefs about the first agent's computational powers, strategy space, and beliefs. This brings us into the realm of type spaces (e.g., (Mertens & Zamir 1985)), but the interaction between type spaces and bounded rationality is unchartered territory (though see (Gmytrasiewicz, Durfee, & Wehe 1991)).

There is much more research to be done on weaving these different considerations into a coherent and comprehensive agent taxonomy. We will not settle this open problem, but will instead focus briefly on the question of how best to evaluate competing methods once an environment has been defined. In recent work (Powers & Shoham 2005), we have attempted to define a set of criteria to guide further effort in designing learning algorithms for stochastic games. These criteria set requirements on the minimal payoff to be achieved against several classes of opponents. In particular we require that an algorithm define a target set against which it is required to achieve an ϵ-optimal payoff, while simultaneously guaranteeing that it achieves at least the payoff of the security level strategy minus ϵ against any opponent. We demonstrate an algorithm that provably meets these criteria when the target set is the class of opponents whose actions are independent of the game history. Though we are continuing to develop algorithms that perform optimally versus more intriguing sets of opponents, ultimately an environment of interest is the set of existing multi-agent learning algorithms. Although this set is too diverse to easily fit within a specific formally defined class of opponents, we can strive to approximate it by sampling. Towards this end, we implemented a wide variety of these algorithms, including many of those described in this paper as well as others from the game theory literature. Within this empirical environment, we were able to compare the performance of each of the algorithms and display both highly competitive performance for our new algorithm and a method for empirical testing in line with the 'AI agenda' as we see it.

One of the interesting observations that came out of this tournament setting is that more complicated algorithms aren't necessarily more effective when evaluated in an environment of other adaptive agents. In a multi-agent setting, learning and teaching are inseparable. Any choice agent i makes is both informed by agent j's past behavior and impacts agent $j's$ future behavior. For this reason, the neutral term 'multi-agent adaptation' might have been more apt. It doesn't have quite the ring of 'multi-agent learning' so we will not wage that linguistic battle, but it is useful to keep the symmetric view in mind when thinking about how to pursue the 'AI agenda'. In particular, it helps explain why greater sophistication is not always an asset. For example, consider an infinitely repeated game of 'chicken':

	yield	dare
yield	2,2	1,3
dare	3,1	0,0

In the presence of any opponent who attempts to learn the other agent's strategy and play a best response (for example, using fictitious play or the system in (Claus & Boutilier 1998)), the best strategy for an agent is to play the stationary policy of always daring; the other agent will soon learn to always yield. This is the "watch out I'm crazy" policy, Stone and Littman's "bully strategy" (Stone & Littman 2001), or Oscar Wilde's "tyranny of the weak". But notice that the success of this simple strategy is a function of its environment, when it competes against other agents who are also using this same strategy it tends to fare abysmally, once again emphasizing the importance of specifying the class of opponents one wishes to perform well against.

Concluding Remarks

We have reviewed previous work in multi-agent RL and have argued for what we believe is a clear and fruitful research agenda in AI on multi-agent learning. Since we have made some critical remarks of previous work, this might give the impression that we don't appreciate it or the researchers behind it. Nothing could be further from the truth. Some of our best friends and colleagues belong to this group, and we have been greatly educated and inspired by their ideas. We look forward to the new and innovative results we are sure to see in the field and hope our comments may contribute to a healthy debate as we work together towards that goal.

Acknowledgements

This work was supported in part by DARPA grant F30602-00-2-0598 and a Benchmark Stanford Graduate Fellowship. We also want to thank the members of the Multi-Agent research group at Stanford University for their helpful advice and feedback throughout this project.

References

Bellman, R. 1957. *Dynamic Programming*. Princeton University Press.

Bowling, M., and Veloso, M. 2001. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*.

Bowling, M. 2000. Convergence problems of general-sum multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 89–94.

Brown, G. 1951. Iterative solution of games by fictitious play. In *Activity Analysis of Production and Allocation*. New York: John Wiley and Sons.

Camerer, C.; Ho, T.; and Chong, J. 2002. Sophisticated EWA learning and strategic teaching in repeated games. *Journal of Economic Theory* 104:137–188.

Chang, Y.-H., and Kaelbling, L. P. 2001. Playing is believing: The role of beliefs in multi-agent learning. In *Proceedings of NIPS*.

Claus, C., and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 746–752.

Erev, I., and Roth, A. E. 1998. Predicting how people play games: reinforcement leaning in experimental games with unique, mixed strategy equilibria. *The American Economic Review* 88(4):848–881.

Fudenberg, D., and Kreps, D. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5:320–367.

Gmytrasiewicz, P.; Durfee, E.; and Wehe, D. 1991. A decision-theoretic approach to coordinating multiagent interactions. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 62–68.

Greenwald, A.; Hall, K.; and Serrano, R. 2002. Correlated-Q learning. In *NIPS Workshop on Multiagent Learning*.

Hannan, J. F. 1959. Approximation to bayes risk in repeated plays. *Contributions to the Theory of Games* 3:97–139.

Hu, J., and Wellman, P. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 242–250.

Hu, J., and Wellman, M. 2001. Learning about other agents in a dynamic multiagent system. *Journal of Cognitive Systems Research* 2:67–69.

Hu, J., and Wellman, M. 2002. Multiagent Q-learning. *Journal of Machine Learning*.

Jehiel, P., and Samet, D. 2001. Learning to play games in extensive form by valuation. *NAJ Economics* 3.

Kalai, E., and Lehrer, E. 1993. Rational learning leads to nash equilibrium. *Econometrica* 61(5):1019–1045.

Leyton-Brown, K., and Tennenholtz, M. 2003. Local-effect games. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 772–780.

Littman, M. L., and Szepesvari, C. 1996. A generalized reinforcement-learning model: Convergence and applications. In *Proceedings of the 13th International Conference on Machine Learning*, 310–318.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, 157–163.

Littman, M. L. 2001. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Mertens, J.-F., and Zamir, S. 1985. Formulation of bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14:1–29.

Nachbar, J. H., and Zame, W. R. 1996. Non-computable strategies and discounted repeated games. *Economic Theory* 8:103–122.

Neyman, A. 1985. Bounded complexity justifies cooperation in finitely repeated prisoner's dilemma. *Economic Letters* 227–229.

Papadimitriou, C., and Yannakakis, M. 1994. On complexity as bounded rationality. In *STOC-94*, 726–733.

Powers, R., and Shoham, Y. 2005. New criteria and a new algorithm for learning in multi-agent systems. In *Advances in Neural Information Processing Systems*. Forthcoming.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. MIT Press.

Sen, S.; Sekaran, M.; and Hale, J. 1994. Learning to coordinate without sharing information. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 426–431.

Stone, P., and Littman, M. L. 2001. Implicit negotiation in repeated games. In Meyer, J.-J., and Tambe, M., eds., *Pre-proceedings of the Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, 96–105.

Tennenholtz, M. 2002. Efficient learning equilibrium. In *Advances in Neural Information Processing Systems*, volume 15. Cambridge, Mass.: MIT Press.

Watkins, C. J. C. H., and Dayan, P. 1992. Technical note: Q-learning. *Machine Learning* 8(3/4):279–292.

Wellman, M. P., and Wurman, P. R. 1999. A trading agent competition for the research community. In *IJCAI-99 Workshop on Agent-Mediated Electronic Trading*.

Multiagent learning â€¢ Learning is process of improving performance via experience â€¢ Can agents learn to coordinate actions with other agents? â€¢ What to learn? S. Albrecht, P. Stone. 6. Multiagent Learning. Multiagent learning â€¢ Learning is process of improving performance via experience â€¢ Can agents learn to coordinate actions with other agents? â€¢ What to learn? â‡’ How to select own actions.Â Research in Multiagent Learning. Multiagent learning studied in different communities â€¢ AI, game theory, robotics, psychology, â€¢ Some conferences & journals: AAMAS, AAAI, IJCAI, NIPS, UAI, ICMLÂ Forming temporary teams â€œon the yâ€ â€¢ Agents designed by different organisations. S. Albrecht, P. Stone. 64. Ad Hoc Teamwork. What if pre-coordination not possible? â€¢ During the process, a multi-agent grid environment is constructed based on characteristics of multi-agent systems and genetic algorithm (GA), and a corresponding neighbor interaction operator, a mutation operator based on neighborhood structure and a self-learning operator are designed. Then, combining tabu search algorithm with a MAGA, the algorithm MAGATS are presented. Finally, 43 benchmark instances are tested with the new algorithm. Compared with four other algorithms, the optimization performance of it is analyzed based on obtained test results.Â PLOS ONE promises fair, rigorous peer review, broad scope, and wide readership â€“ a perfect fit for your research every time. Learn More Submit Now. About. Why Publish with PLOS ONE. Cooperative Multi-Agent Reinforcement Learning. Shimon Whiteson Dept. of Computer Science. University of Oxford joint work with Jakob Foerster, Gregory FarquharÂ Speed learning with parameter sharing Dierent inputs, including a, induce dierent behaviour Still independent: critics condition only on Ï„ a and ua. Limitations: Nonstationary learning Hard to learn to coordinate Multi-agent credit assignment. Shimon Whiteson (Oxford). Cooperative Multi-Agent RL. July 4, 2018 14 / 27. Counterfactual Multi-Agent Policy Gradients. Centralised critic: stabilise learning to coordinate Counterfactual baseline: tackle multi-agent credit assignment Ecient critic representation: scale to large NNs. Shimon Whiteson (Oxford). Cooperative Multi-Agent RL. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. Kaiqing Zhang. Zhuoran Yang.Â Our overall goal with this chapter is, beyond providing an assessment of the current state of the eld on the mark, to identify fruitful future research directions on theoretical studies of MARL. We expect this chapter to serve as continuing stimulus for researchers interested in working on this exciting while challenging topic.