Multi-Agent Reinforcement Learning: a critical survey

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

May 16, 2003

Abstract

We survey the recent work in AI on multi-agent reinforcement learning (that is, learning in stochastic games). We then argue that, while exciting, this work is flawed. The fundamental flaw is unclarity about the problem or problems being addressed. After tracing a representative sample of the recent literature, we identify four well-defined problems in multi-agent reinforcement learning, single out the problem that in our view is most suitable for AI, and make some remarks about how we believe progress is to be made on this problem.

1 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 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 and Dayan1992], and through it on the Bellman equations [Bellman1957]. Specifically, we will discuss [Littman1994, Claus and Boutilier1998, Hu and Wellman1998, Bowling and Veloso2001, Littman2001,

Greenwald *et al.*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 the problem runs deeper, we believe, and has to do with a basic unclarity about the problem being addressed. In section 4 we argue that there are (at least) four distinct well-defined problems to be addressed, and that the tail end of the "Bellman heritage" does not fit in any of them. We identify one of the four as the most interesting for AI, and that 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.

2 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 and Samet2001] for an exception). $\vec{A} = A_1, \ldots, A_n$, with A_i the set of actions (or pure strategies) of agent i (note 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 $R_i : S \times \vec{A} \to \mathcal{R}$ the immediate reward function of agent i. $T : S \times \vec{A} \to \Pi(S)$ 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).

2.1 From Minimax-Q to Nash-Q and beyond

We start with the *(single-agent) Q-learning* algorithm [Watkins and Dayan1992] for computing an optimal policy in an MDP with unknown reward and transition functions:[1]

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

[1]This procedure is based directly on the Bellman equations [Bellman1957] and the dynamic programming procedures based on them for MDPs with known reward and transition functions.

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 *et al.*1994]). However, this approach is unmotivated for two reasons. First, the definition of the Q-values assumes incorrectly that they are independent of the actions selected by the other agents. Second, it is no longer sensible to use the maximum of the Q-values to update V.

The cure to the first problem is to simply 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 [Littman1994]:

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

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 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* [Brown1951] and *rational learning* [Kalai and Lehrer1993], is pursued by Claus and Boutilier [Claus and Boutilier1998]. In their work they 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. The pivotal contribution here is *Nash-Q* learning [Hu and Wellman1998], 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:

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

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[Littman2001]. 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 *et al.*2002]:

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

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

2.2 Convergence results

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 and Szepesvari1996]

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 and Boutilier1998] 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 *uncorrelated* Nash equilibria in a number of empirical experiments, although there are no formal results presented.

3 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.[3] Furthermore, note that the conditions are in fact quite restrictive, since they must hold for the games defined by the intermediate Q-values throughout the execution of the protocol. So it is extremely unlikely that a game will satisfy this condition, and in any case hard to verify at the outset whether it does.

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 equilibria 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 [Bowling2000], Littman [Littman2001], and by Hu and Wellman themselves in the journal version of their article [Hu and Wellman2002].

[3] 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 is guaranteed in but not limited to common-payoff games. However, it is hard to find natural cases in which the conditions hold other than in the special cases.

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.[4]

3. Even if by magic one could pick out a unique equilibrium at each stage game, why is that relevant in light of the fact that one is playing an extended SG?

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 and Veloso2001]. 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 [Hannan1959]). In particular, their conditions require that during self-play, learning only terminate in a stationary Nash equilibrium. This is a useful criterion, but it is weak in that it ignores the fact that one is playing an extended SG.[5] 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.

[4]One can view the CE-Q learning as throwing in the towel, and admitting upfront that one must assume a correlating device for the agents. We don't have a quarrel with this stance, but we don't see the motivation for focusing on a sample correlated equilibrium. In particular, we are not swayed by the argument that a sample correlated equilibrium is easier to compute than a sample Nash equilibrium; there is no reason to think that the signal inherent in the sample equilibrium that is computed will be available to the agents in a given situation. In any event, we have already noted that Greenwald et al. report that, in simulations, CE-Q seems to converge to uncorrelated Nash equilibria in self-play.

[5]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 GT too is unclear on its motivation in doing so. We comment on this in the next section, but this is not our focus in this article.

4 Four well defined problems 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. If (e.g.,) Nash-Q is the answer, what is the question? In this section we identify what we think is a coherent research agenda on multi-agent RL. In fact, we generously offer four 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 and Roth1998, Camerer *et al.*2002]). The name of the game here is to show experimentally that a certain formal model of learning agrees with people's behavior (typically, in laboratory experiments). This line of work is as legitimate and coherent as any other experimental work in psychology, and we have no further comment on it. Much of the work on learning in game theory has adopted this stance, if often implicitly. Researchers propose various dynamics that are a perceived as plausible in one sense or another, 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 play a central role in non-cooperative game theory.[6] The main limitation of that 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 might be called the 'distributed AI (DAI) agenda'. This is a problem of distributed control; a central designer controls multiple agents, but cannot or will not design an optimal policy for them. Instead it assigns them each an adaptive procedure that converges to an optimal policy. In this case there is no role for equilibrium analysis; the agents have no freedom to deviate from the prescribed algorithm.

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 is not one studied in game theory, and in fact is explicitly rejected in at least one place [Fudenberg and Kreps1993]; we have only seen it pursued recently, outside game theory [Tennenholtz2002]. The agenda can be described as follows. Since, in the traditional view of non-

[6] 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.

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 final prescriptive agenda is one we call 'the AI agenda'. 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 unglamorous. 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 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.

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 and Wellman2001, Hu and Wellman2002, Stone and Littman2001]) 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 and Kaelbling2001], 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 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., [Rubinstein1998]), 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 modelling 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 parameter-

ize 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. However, while in principle one could fold in this evolutionary element into a meta-learning algorithm that includes both the short-term learning and long-term evolution, this theoretical construct will in general not provide any useful insight.

The case of the Trading Agent Competition (TAC) serves to illustrate the point.[7] 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 modelling 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 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 fits nicely in the 'AI agenda', in the highly specialized case of zero-sum games. 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. Near as we can tell, however, Nash-Q and its descendants do not fit any of the agendas we discussed.

5 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.

[7] TAC [Wellman and Wurman1999] is a series of competitions in which computerized agents trade in a non-trivial set of interacting markets.

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 and Zame1996]. We however will concentrate on the question of how to parameterize the space of agents, which itself is 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 JAL learners in the sense of [Claus and Boutilier1998]. 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.[8] 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 and Kaelbling2001], and to some extent [Stone and Littman2001], 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 and Lehrer1993], and to place restrictions on its prior about our 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 and Zamir1985]), but the interaction between type spaces and bounded rationality is unchartered territory (though see [Gmytrasiewicz *et al.*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 let us make a final note regarding the temptation to label some agent types learning as 'weak' and others as 'strong' with respect to any taxonomy. In a multi-agent setting, learning and teaching are inseparable. Any choice i makes is both informed by j's past behavior and impacts $j's$ 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':

[8]This is the model pursued in the work on 'bounded rationality' (e.g., [Neyman1985, Papadimitriou and Yannakakis1994, Rubinstein1998]. 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.'

	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 and Boutilier1998]), the best strategy for an agent is to play the stationary policy of always daring; the other agent will learn to always yield. This is the "watch out I'm crazy" policy, Stone and Littman's "bully strategy" [Stone and Littman2001], or Oscar Wilde's "tyranny of the weak".

6 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. Granted, when you stand on the shoulders of giants, sometimes it can be uncomfortable for the giants. Our own request is that should our colleagues ever decide to stand on our shoulders, they refrain from wearing spiked heels.

References

[Bellman1957] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[Bowling and Veloso2001] Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

[Bowling2000] Michael Bowling. Convergence problems of general-sum multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 89–94, 2000.

[Brown1951] G. Brown. Iterative solution of games by fictitious play. In *Activity Analysis of Production and Allocation*. John Wiley and Sons, New York, 1951.

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

[Chang and Kaelbling2001] Yu-Han Chang and Leslie Pack Kaelbling. Playing is believing: The role of beliefs in multi-agent learning. In *Proceedings of NIPS*, 2001.

[Claus and Boutilier1998] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, 1998.

[Erev and Roth1998] Ido Erev and Alvin E. Roth. Predicting how people play games: reinforcement leaning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, 88(4):848–881, September 1998.

[Fudenberg and Kreps1993] D. Fudenberg and David Kreps. Learning mixed equilibria. *Games and Economic Behavior*, 5:320–367, 1993.

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

[Greenwald *et al.*2002] Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated-Q learning. In *NIPS Workshop on Multiagent Learning*, 2002.

[Hannan1959] J. F. Hannan. Approximation to bayes risk in repeated plays. *Contributions to the Theory of Games*, 3:97–139, 1959.

[Hu and Wellman1998] J. Hu and P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250, 1998.

[Hu and Wellman2001] Junling Hu and Michael Wellman. Learning about other agents in a dynamic multiagent system. *Journal of Cognitive Systems Research*, 2:67–69, 2001.

[Hu and Wellman2002] Junling Hu and Michael Wellman. Multiagent Q-learning. *Journal of Machine Learning*, 2002.

[Jehiel and Samet2001] Phillipe Jehiel and Dov Samet. Learning to play games in extensive form by valuation. *NAJ Economics*, 3, 2001.

[Kalai and Lehrer1993] Ehud Kalai and Ehud Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61(5):1019–1045, 1993.

[Littman and Szepesvari1996] Michael L. Littman and C. Szepesvari. A generalized reinforcement-learning model: Convergence and applications. In *Proceedings of the 13th International Conference on Machine Learning*, pages 310–318, 1996.

[Littman1994] Michael L. Littman. Markov games as a framework for multiagent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 157–163, 1994.

[Littman2001] Michael L. Littman. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[Mertens and Zamir1985] J-F. Mertens and S. Zamir. Formulation of bayesian analysis for games with incomplete information. *International Journal of Game Theory*, 14:1–29, 1985.

[Nachbar and Zame1996] John H. Nachbar and William R. Zame. Non-computable strategies and discounted repeated games. *Economic Theory*, 8:103–122, 1996.

[Neyman1985] Abraham Neyman. Bounded complexity justifies cooperation in finitely repeated prisoner's dilemma. *Economic Letters*, pages 227–229, 1985.

[Papadimitriou and Yannakakis1994] C.H. Papadimitriou and M. Yannakakis. On complexity as bounded rationality. In *STOC-94*, pages 726–733, 1994.

[Rubinstein1998] Ariel Rubinstein. *Modeling Bounded Rationality*. MIT Press, 1998.

[Sen *et al.*1994] Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 426–431, Seattle, WA, 1994.

[Stone and Littman2001] Peter Stone and Michael L. Littman. Implicit negotiation in repeated games. In John-Jules Meyer and Milind Tambe, editors, *Pre-proceedings of the Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, pages 96–105, 2001.

[Tennenholtz2002] Moshe Tennenholtz. Efficient learning equilibrium. In *Proceedings of NIPS*, 2002.

[Watkins and Dayan1992] C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8(3/4):279–292, May 1992.

[Wellman and Wurman1999] Michael P. Wellman and Peter R. Wurman. A trading agent competition for the research community. In *IJCAI-99 Workshop on Agent-Mediated Electronic Trading*, Stockholm, August 1999.

@inproceedings{Shoham2003MultiAgentRL, title={Multi-Agent Reinforcement Learning:a critical survey}, author={Yoav Shoham and Rob Powers and Trond Grenager}, year={2003} }. Yoav Shoham, Rob Powers, Trond Grenager. Published 2003. We survey the recent work in AI on multi-agent reinforcement learning (that is, learning in stochastic games). We then argue that, while exciting, this work is flawed. The fundamental flaw is unclarity about the problem or problems being addressed. After tracing a representative sample of the recent literature, we identify four well-defined problems in multi-agent rein... Reinforcement learning Multi-agent systems Partially observable Markov decision processes Shaping Policy-gradient. Abbreviations. RL. Reinforcement learning. MAS. Multi-agent system.Â Multi-agent reinforcement learning: A critical survey. Technical report, Stanford.Google Scholar. 39.Â Multiagent systems: A survey from a machine learning perspective. Autonomous Robotics, 8(3).Google Scholar. 44.