If multi-agent learning is the answer, what is the question?

Yoav Shoham, Rob Powers, and Trond Grenager
Stanford University
{shoham,powers,grenager}@cs.stanford.edu

February 15, 2006

Abstract

The area of learning in multi-agent systems is today one of the most fertile grounds for interaction between game theory and artificial intelligence. We focus on the foundational questions in this interdisciplinary area, and identify several distinct agendas that ought to, we argue, be separated. The goal of this article is to start a discussion in the research community that will result in firmer foundations for the area.[1]

1 Introduction

The topic of learning in multi-agent systems, or multi-agent learning (MAL henceforth), has a long history in game theory, almost as long as the history of game theory itself.[2] As early as 1951, fictitious play [Brown, 1951] was proposed as a learning algorithm for computing equilibria in games and there have been proposals for how to evaluate the success of learning rules going back to [Hannan, 1957] and [Blackwell, 1956]. Since that time hundreds, if not thousands, of articles have been published on the topic, and at least two books ([Fudenberg and Levine, 1998] and [Young, 2004]).

In Artificial Intelligence (AI) the history of *single*-agent learning is as rich if not richer, with thousands of articles, many books, and some very compelling applications in a variety of fields (for some examples see [Kaelbling et al., 1996], [Mitchell, 1997], or [Sutton and Barto, 1998]). While it is only in recent years

[1] This article has a long history and owes many debts. A first version was presented at the NIPS workshop, Multi-Agent Learning: Theory and Practice, in 2002. A later version was presented at the AAAI Fall Symposium in 2004 [Shoham et al., 2004]. Over time it has gradually evolved into the current form, as a result of our own work in the area as well as the feedback of many colleagues. We thank them all collectively, with special thanks to members of the multi-agent group at Stanford in the past three years. Rakesh Vohra and Michael Wellman provided detailed comments on the latest draft which resulted in substantive improvements, although we alone are responsible for the views put forward. This work was supported by NSF ITR grant IIS-0205633 and DARPA grant HR0011-05-1.

[2] Another more recent term for the area within game theory is *interactive learning*.

that AI has branched into the multi-agent aspects of learning, it has done so with something of a vengeance. If in 2003 one could describe the AI literature on MAL by enumerating the relevant articles, today this is no longer possible. The leading conferences routinely feature articles on MAL, as do the journals.[3]

While the AI literature maintains a certain flavor that distinguishes it from the game theoretic literature, the commonalities are greater than the differences. Indeed, alongside the area of mechanism design, and perhaps the computational questions surrounding solution concepts such as the Nash equilibrium, MAL is today arguably one of the most fertile interaction grounds between computer science and game theory.

The MAL research in both fields has produced some inspiring results. We will not repeat them here, since we cannot be comprehensive in this article, but nothing we say subsequently should be interpreted as belittling the achievements in the area. Yet alongside these successes there are some indications that it could be useful to take a step back and ask a few basic questions about the area of MAL. One surface indication is the presence of quite a number of frustrating dead ends. For example, the AI literature attempting to extend Bellman-style single-agent reinforcement learning techniques (in particular, Q-learning [Watkins and Dayan, 1992]) to the multi-agent setting, has fared well in zero-sum repeated games (e.g., [Littman, 1994] and [Littman and Szepesvari, 1996]) as well as common-payoff (or 'team') repeated games (e.g., [Claus and Boutilier, 1998], [Kapetanakis and Kudenko, 2004], and [Wang and Sandholm, 2002]), but less well in general-sum stochastic games (e.g., [Hu and Wellman, 1998], [Littman, 2001] and [Greenwald and Hall, 2003]) (for the reader unfamiliar with this line of work, we cover it briefly in Section 4). Indeed, upon close examination, it becomes clear that the very foundations of MAL could benefit from explicit discussion. What exact question or questions is MAL addressing? What are the yardsticks by which to measure answers to these questions? The present article focuses on these foundational questions.

To start with the punch line, following an extensive look at the literature we have reached two conclusions:

- There are several different agendas being pursued in the MAL literature. They are often left implicit and conflated; the result is that it is hard to evaluate and compare results.
- We ourselves can identify and make sense of five distinct research agendas.

Not all work in the field falls into one of the five agendas we identify. This is not necessarily a critique of work that doesn't; it simply means that one must identify yet other well-motivated and well defined problems addressed by that work. We expect that as a result of our throwing down the gauntlet additional such problems will be defined, but also that some past work will

[3] We acknowledge a simplification of history here. There is definitely MAL work in AI that predates the last few years, though the relative deluge is indeed recent. Similarly, we focus on AI since this is where most of the action is these days, but there are also other areas in computer science that feature MAL material; we mean to include that literature here as well.

be re-evaluated and reconstructed. Certainly we hope that future work will always be conducted and evaluated against well-defined criteria, guided by this article and the discussion engendered by it among our colleagues in AI and game theory. In general we view this article not as a final statement but as the start of a discussion.

In order to get to the punch line outlined above, we proceed as follows. In the next section we define the formal setting on which we focus. In Section 3 we illustrate why the question of learning in multi-agent settings is inherently more complex than in the single-agent setting, and why it places a stress on basic game theoretic notions. In Section 4 we provide some concrete examples of MAL approaches from both game theory and AI. This is anything but a comprehensive coverage of the area, and the selection is not a value judgment. Our intention is to anchor the discussion in something concrete for the benefit of the reader who is not familiar with the area, and – within the formal confines we discuss in Section 2 – the examples span the space of MAL reasonably well. In Section 5 we identify five different agendas that we see (usually) implicit in the literature, and which we argue should be made explicit and teased apart. We end in Section 6 with a summary of the main points made in this article.

A final remark is in order. The reader may find some of the material in the next three sections basic or obvious; different readers will probably find different parts so. We don't mean to insult anyone's intelligence, but we err on the side of explicitness for two reasons. First, this article is addressed to at least two different communities with somewhat different backgrounds. Second, our goal is to contribute to the clarification of foundational issues; we don't want to be guilty of vagueness ourselves.

2 The formal setting

We will couch our discussion in the formal setting of *stochastic games* (aka *Markov games*). Most of the MAL literature adopts this setting, and indeed most of it focuses on the even more narrow class of *repeated games.* Furthermore, stochastic games also generalize *Markov Decision Problems (MDPs)*, the setting from which much of the relevant learning literature in AI originates. These are defined as follows.

A stochastic game can be represented as 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. $\vec{A} = A_1, \ldots, A_n$, 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 $R_i : S \times \vec{A} \rightarrow \mathcal{R}$ giving the immediate reward function of agent i for stage game S. $T : S \times \vec{A} \rightarrow \Pi(S)$ is a stochastic transition function, specifying the probability of the next stage game to be played based on the game just played and the actions taken in it.

We also need to define a way for each agent to aggregate the set of immediate rewards received in each state. For finitely repeated games we can simply use the sum or average, while for infinite games the most common approaches are to use either the limit average or the sum of discounted awards $\sum_{t=1}^{\infty} \delta^t r_t$, where r_t is the reward received at time t.

A repeated game is a stochastic game with only one stage game, while an MDP is a stochastic game with only one agent.

While most of the MAL literature lives happily in this setting, we would be remiss not to acknowledge the literature that does not. Certainly one could discuss learning in the context of extensive-form games of incomplete and/or imperfect information (cf. [Jehiel and Samet, 2001]). We don't dwell on those since it would distract from the main discussion, and since the lessons we draw from our setting will apply there as well.

Although we will not specifically include them, we also intend our comments to apply at a general level to large population games and evolutionary models, and particularly *replicator dynamics (RD)* [Schuster and Sigmund, 1983] and *evolutionary stable strategies (ESS)* [Smith, 1982]. These are defined as follows. The replicator dynamic model assumes a population of homogenous agents each of which continuously plays a two-player game against every other agent. Formally the setting can be expressed as a tuple $(A, \vec{P_0}, R)$. A is the set of possible pure strategies/actions for the agents indexed $1, \ldots, m$. P_0 is the initial distribution of agents across possible strategies, $\sum_{i=1}^{m} P_0(i) = 1$. $R : A \times A \to \mathcal{R}$ is the immediate reward function for each agent with $R(a, a')$ giving the reward for an agent playing strategy a against another agent playing strategy a'. The population then changes proportions according to how the reward for each strategy compares to the average reward: $dt(P_t(a)) = P_t(a)[u_t(a) - u_t^*]$, where $u_t(a) = \sum_{a'} P_t(a')R(a, a')$ and $u_t^* = \sum_a P_t(a)u_t(a)$. A strategy a is then defined to be an evolutionary stable strategy if and only if for some $\epsilon > 0$ and for all other strategies a', $R(a, (1-\epsilon)a + \epsilon a') > R(a', (1-\epsilon)a + \epsilon a')$.

As the names suggest, one way to interpret these settings is as building on population genetics, that is, as representing a large population undergoing frequent pairwise interactions. An alternative interpretation however is as a repeated game between two agents, with the distribution of strategies in the population representing the agent's mixed strategy (in the homogenous definition above the two agents have the same mixed strategy, but there exist more general definitions with more than two agents and with non-identical strategies). The second interpretation reduces the setting to the one we discuss. The first bears more discussion, and we do it briefly in Section 4.

And so we stay with the framework of stochastic games. What is there to learn in these games? Here we need to be explicit about some aspects of stochastic games that were glossed over so far. Do the agents know the stochastic game, including the stage games and the transition probabilities? If not, do they at least know the specific game being played at each stage, or only the actions available to them? What do they see after each stage game has been played – only their own rewards, or also the actions played by the other agent(s)? Do they perhaps magically see the other agent(s)' mixed strategy in the stage game? And so on.

In general, games may be known or not, play may be observable or not, and so on. We will focus on known, fully observable games, where the other agent's strategy (or agents' strategies) is not known a priori (though in some case there is a prior distribution over it). In our restricted setting there two possible things to learn. First, the agent can learn the opponent's (or opponents') strategy (or strategies), so that the agent can then devise a best (or at least a good) response. Alternatively, the agent can learn a strategy of his own that does well against the opponents, without explicitly learning the opponent's strategy. The first is sometimes called *model-based learning*, and the second *model-free learning*.

In broader settings there is more to learn. In particular, with unknown games, one can learn the game itself. Some will argue the restricted setting is not a true learning setting, but (a) much of the current work on MAL, particularly in game theory, takes place in this setting, and (b) the foundational issues we wish to tackle surface already here. In particular, our comments are intended to also apply to the work in the AI literature on games with unknown payoffs, work which builds on the success of learning in unknown MDPs. We will have more to say about the nature of 'learning' in the setting of stochastic games in the following sections.

3 On some special characteristics of multi-agent learning

Before launching into specifics, we wish to highlight the special nature of MAL. There are two messages we would like to get across, one aimed at AI researchers specifically and one more broadly. Both lessons can be gleaned from simple and well-known examples.

	Left	*Right*
Up	$1,0$	$3,2$
Down	$2,1$	$4,0$

Figure 1: Stackelberg stage game: The payoff for the row player is given first in each cell, with the payoff for the column player following.

Consider the game described in Figure 1. In this game the row player has a strictly dominant strategy, *Down*, and so seemingly there is not much more to say about this game. But now imagine a repeated version of this game. If the row player indeed repeatedly plays *Down*, assuming the column player is paying any attention, he (the column player) will start responding with *Left*, and the two will end up with a repeated (*Down*,*Left*) play. If, however, the row player starts repeatedly playing *Up*, and again assuming the column player is awake, he may instead start responding by playing *Right*, and the two players will end up with a repeated (*Up*,*Right*) play. The lesson from this is simple yet profound: In a multi-agent setting one cannot separate *learning* from *teaching*. In this example, by playing his dominated strategy, the row player taught the column player to play in a way that benefits both. Indeed, for this reason it might be more appropriate to speak more neutrally about multi-agent *adaptation* rather than *learning*. We will not fight this linguistic battle, but the point remains important, especially for computer scientists who are less accustomed to thinking about interactive considerations than game theorists. In particular, it follows there is no a priori reason to expect that machine learning techniques that have proved successful in AI for single-agent settings will also prove relevant in the multi-agent setting.

The second lesson we draw from the well-known game of Rochambeau, or Rock-Paper-Scissors, given in Figure 2.

	Rock	*Paper*	*Scissors*
Rock	$0,0$	$-1,1$	$1,-1$
Paper	$-1,1$	$0,0$	$1,-1$
Scissors	$-1,1$	$1,-1$	$0,0$

Figure 2: Rock-Paper-Scissors

As is well known, this zero-sum game has a unique Nash equilibrium in which each player randomizes uniformly among the three strategies. One could conclude that there is not much more to say about the game. But suppose you entered a Rochambeau tournament. Would you simply adopt the equilibrium strategy?

If you did, you would not win the competition. This is no idle speculation; such competitions take place routinely. For example, starting in 2002, the World Rock Papers Scissors Society (WRPS) standardized a set of rules for international play and has overseen annual International World Championships as well as many regional and national events throughout the year. These championships have been attended by players from around the world and have attracted widespread international media attention. The winners are never equilibrium players. For example, on October 25th, 2005, 495 people entered the competition in Toronto from countries as diverse as Norway, Northern Ireland, the Cayman Islands, Australia, New Zealand and the UK. The winner was Toronto Lawyer Andrew Bergel, who beat Californian Stan Long in the finals. His strategy? "[I] read the minds of my competitors and figure out what they were thinking. I don't believe in planning your throws before you meet your opponent."

These tournaments of course are not a perfect match with the formal model of repeated games. However, we include the example not only for entertainment value. The rules of the RPS tournaments call for 'matches' between players, each match consisting of several 'games', where a game is a single play of RPS. The early matches adopted a "best of three of three" format, meaning that the player who wins a best of three set garners one point and requires two points to take the match. The Semi-Finals and the Final Match used the "best of three of five" format, meaning that the player who wins a best of three set garners one point and requires three points to take the match. And so the competition really consisted of a series of repeated games, some of them longer than others.[4]

Entertainment aside, what do we learn from this? We believe that this is a cautioning tale regarding the predictive or prescriptive role of equilibria in complex games, and in particular in repeated games. There are many examples of games with complex strategy spaces, in which equilibrium analysis plays little or no role – including familiar parlor games, or the Trading Agent Competition (TAC), a computerized trading competition[5]. The strategy space in a repeated game (or more generally a stochastic game) is immense – all mappings from past history to mixed strategies in the stage game. In such complex games it is not reasonable to expect that players contemplate the entire strategy space – their own or that of the opponent(s). Thus, (e.g., Nash) equilibria don't play here as great a predictive or prescriptive role.

Our cautioning words should be viewed as countering the default blind adoption of equilibria as the driving concept in complex games, but not as a sweeping statement against the relevance of equilibria in some cases. The simpler the stage game, and the longer its repetition, the more instructive are the equilibria. Indeed, despite our example above, we do believe that if only two players play a repeated RPS game for long enough, they will tend to converge to the equilibrium strategy (this is particularly true of computer programs, that don't share the human difficulty with throwing a mental die). Even in more complex games there are examples where computer calculation of approximate equilibria within a restricted strategy space provided valuable guidance in constructing effective strategies. This includes the game of Poker ([Koller and Pfeffer, 1997] and [Billings et al., 2003]), and even, as an exception to the general rule we mentioned, one program that competed in the Trading Agent Competition [Cheng et al., 2005]. Our point has only been that in the context of complex games, so-called "bounded rationality", or the deviation from the ideal behavior of omniscient agents, is not an esoteric phenomenon to be brushed aside.

4 A (very partial) sample of MAL work

To make the discussion concrete, it is useful to look at MAL work over the years. The selection that follows is representative but very partial; no value

[4] We do acknowledge some degree of humor in the example. The detailed rules in `http://www.rpschamps.com/rules.html` make for additional entertaining reading; of note is the restriction of the strategy space to Rock, Papers and Scissors, and explicitly ruling out others: "Any use of Dynamite, Bird, Well, Spock, Water, Match, Fire, God, Lightning, Bomb, Texas Longhorn, or other non-sanctioned throws, will result in automatic disqualification." The overview of the RPS society and its tournaments is adapted from the inimitable Wikipedia, the collaborative online encyclopedia, as available on January 2, 2006. Wikipedia goes on to list the champions since 2002; we note without comment that they are all male Torontonians. The results of the specific competition cited are drawn from the online edition of the Boise Weekly dated November 2, 2005. The Boise Weekly starts the piece with "If it weren't true, we wouldn't report on it".

[5] http://tac.eecs.umich.edu

judgment or other bias are intended by this selection. The reader familiar with the literature may wish to skip to Section 4.3, where we make some general subjective comments.

Unless we indicate otherwise, our examples are drawn from the special case of repeated, two-person games (as opposed to stochastic, n-player games). We do this both for ease of exposition, and because the bulk of the literature indeed focuses on this special case.

We divide the coverage into three parts: techniques, results, and commentary.

4.1 Some MAL techniques

We will discuss three classes of techniques – one representative of work in game theory, one more typical of work in AI, and one that seems to have drawn equal attention from both communities.

4.1.1 Model-based approaches

The first approach to learning we discuss, which is common in the game theory literature, is the model-based one. It adopts the following general scheme:

1. Start with some model of the opponent's strategy.
2. Compute and play the best response.
3. Observe the opponent's play and update your model of her strategy.
4. Goto step 2.

Among the earliest, and probably the best-known, instance of this scheme is *fictitious play* [Brown, 1951]. The model is simply a count of the plays by the opponent in the past. The opponent is assumed to be playing a stationary strategy, and the observed frequencies are taken to represent the opponent's mixed strategy. Thus after five repetitions of the Rochambeau game in which the opponent played (R, S, P, R, P), the current model of her mixed strategy is $(R = .4, P = .4, S = .2)$.

There exist many variants of the general scheme, for example those in which one does not play the exact best response in step 2. This is typically accomplished by assigning a probability of playing each pure strategy, assigning the best response the highest probability, but allowing some chance of playing any of the strategies. A number of proposals have been made of different ways to assign these probabilities such as *smooth fictitious play* [Fudenberg and Kreps, 1993] and *exponential fictitious play* [Fudenberg and Levine, 1995].

A more sophisticated version of the same scheme is seen in *rational learning* [Kalai and Lehrer, 1993]. The model is a distribution over the repeated-game strategies. One starts with some prior distribution; for example, in a repeated Rochambeau game, the prior could state that with probability .5 the opponent repeatedly plays the equilibrium strategy of the stage game, and, for all $k > 1$, with probability 2^{-k} she plays R k times and then reverts to the repeated equilibrium strategy. After each play, the model is updated to be the posterior obtained by Bayesian conditioning of the previous model. For instance, in our example, after the first non-R play of the opponent, the posterior places probability 1 on the repeated equilibrium play.

4.1.2 Model-free approaches

An entirely different approach that has been commonly pursued in the AI literature [Kaelbling et al., 1996], is the model-free one, which avoids building an explicit model of the opponent's strategy. Instead, over time one learns how well one's own various possible actions fare. This work takes place under the general heading of *reinforcement learning*[6], and most approaches have their roots in the Bellman equations [Bellman, 1957]. The basic algorithm for solving for the best policy in a known MDP starts by initializing a value function, $V_0 : S \rightarrow \mathcal{R}$, with a value for each state in the MDP. The value function can then be iteratively updated using the Bellman equation:

$$V_{k+1} \leftarrow R(s) + \gamma max_a \sum_{s'} T(s,a,s')V_k(s')$$

The optimal policy can then be obtained by selecting the action, a, at each state, s, that maximizes the expected value: $\sum_{s'} T(s,a,s')V_k(s')$. Much of the work in AI has focused strategies for rapid convergence, on very large MDPs, and in particular on unknown and partially observable MDPs. While this is not our focus, we do briefly discuss the unknown case, since this is where the literature leading to many of the current approaches for stochastic games originated.

For MDPs with unknown reward and transition functions, the *Q-learning* algorithm [Watkins and Dayan, 1992] can be used to compute an optimal policy.

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

As is well known, with certain assumptions about the way in which actions are selected at each state over time and constraints on the learning rate schedule, α_t, Q-learning can be shown to converge to the optimal value function V^*.

The Q-learning algorithm can be extended to the multi-agent stochastic game setting by having each agent simply ignore the other agents and pretend that the environment is passive:

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

[6] We note that the term is used somewhat differently in the game theory literature.

Several authors have tested variations of the basic Q-learning algorithm for MAL (e.g., [Sen et al., 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' distributions of actions are stationary, it can fail when an opponent may adapt its choice of actions based on the past history of the game.

A first step in addressing this problem is to define the Q-values as a function of all the agents' actions:

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

We are however 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]:

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

Later work (such as the joint-action learners in [Claus and Boutilier, 1998] and the Friend-or-Foe Q algorithm in [Littman, 2001]) proposed other update rules for the Q and V functions focusing on the special case 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.*

The work on zero-sum and common-payoff games continues to be refined and extended (e.g., [Kearns and Singh, 1998], [Brafman and Tennenholtz, 2000], [Lauer and Riedmiller, 2000], and [Wang and Sandholm, 2002]). Much of this work has concentrated on provably optimal tradeoffs between exploration and exploitation in unknown, zero-sum games; this is a fascinating topic, but not germane to our focus. More relevant are the most recent efforts in this line of research to extend the "Bellman heritage" to general-sum games (e.g., Nash-Q by [Hu and Wellman, 2003] and CE-Q by [Greenwald and Hall, 2003]). We do not cover these for two reasons: The description is more involved, and the results have been less satisfactory; more on the latter below.

4.1.3 Regret minimization approaches

Our third and final example of prior work in MAL is no-regret learning. It is an interesting example for two reasons. First, it has some unique properties that distinguish it from the work above. Second, both the AI and game theory communities appear to have converged on it independently. The basic idea goes back to early work on how to evaluate the success of learning rules [Hannan, 1957] and [Blackwell, 1956], and has since been extended and rediscovered numerous times over the years under the names of universal consistency, no-regret learning, and the Bayes envelope (see [Foster and Vohra, 1999] for an overview of this history). We will describe the algorithm proposed in [Hart and Mas-Colell, 2000] as a representative of this body of work. We start by defining the *regret*, $r_i^t(a_j, s_i)$ of agent i for playing the sequence of actions s_i instead of playing action a_j, given that the opponents played the sequence s_{-i}.

$$r_i^t(a_j, s_i|s_{-i}) = \sum_{k=1}^{t} R(a_j, s_{-i}^k) - R(s_i^k, s_{-i}^k)$$

The agent then selects each of its actions with probability proportional to $\max(r_i^t(a_j, s_i), 0)$ at each time step $t+1$.

Recently, these ideas have also been adopted by researchers in the computer science community (e.g., [Freund and Schapire, 1995], [Jafari et al., 2001], and [Zinkevich, 2003]).

Note that the application of approaches based on regret minimization has been restricted to the case of repeated games. The difficulties of extending this concept to stochastic games are discussed in [Mannor and Shimkin, 2003].

4.2 Some typical results

One sees at least three kinds of results in the literature regarding the learning algorithms presented above, and others like them. These are:

1. Convergence of the strategy profile to an (e.g., Nash) equilibrium of the stage game in self play (that is, when all agents adopt the learning procedure under consideration).

2. Successful learning of an opponent's strategy (or opponents' strategies).

3. Obtaining payoffs that exceed a specified threshold.

Each of these types comes in many flavors; here are some examples. The first type is perhaps the most common in the literature, in both game theory and AI. For example, while fictitious play does not in general converge to a Nash equilibrium of the stage game, the distribution of its play can be shown to converge to an equilibrium in zero-sum games [Robinson, 1951], 2x2 games with generic payoffs [Miyasawa, 1961], or games that can be solved by iterated elimination of strictly dominated strategies [Nachbar, 1990].

Similarly in AI, in [Littman and 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. This result makes 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 Boutilier, 1998] conjecture that both single-agent Q-learners and the belief-based joint action learners they proposed converge to an equilibrium in common payoff games under the conditions of self-play and decreasing exploration, but do not offer a formal proof. Friend-or-Foe Q and Nash-Q were both shown to converge to a Nash equilibrium in a set of games that are a slight generalization on the set of zero-sum and common payoff games.

Rational learning exemplifies results of the second type. The convergence shown is to correct beliefs about the opponent's repeated game strategy; thus it follows that, since each agent adopts a best response to their beliefs about the other agent, in the limit the agents will converge to a Nash equilibrium of the repeated game. This is an impressive result, but it is limited by two factors; the convergence depends on a very strong assumption of absolute continuity, and the beliefs converged to are only correct with respect to the aspects of history that are observable given the strategies of the agents. This is an involved topic, and the reader is referred to the literature for more details.

The literature on no-regret learning provides an example of the third type of result, and has perhaps been the most explicit about criteria for evaluating learning rules. For example, in [Fudenberg and Levine, 1995] two criteria are suggested. The first is that the learning rule be 'safe', which is defined as the requirement that the learning rule guarantee at least the minimax payoff of the game. (The minimax payoff is the maximum expected value a player can guarantee against any possible opponent.) The second criterion is that the rule should be 'consistent'. In order to be 'consistent', the learning rule must guarantee that it does at least as well as the best response to the empirical distribution of play when playing against an opponent whose play is governed by independent draws from a fixed distribution. They then define 'universal consistency' as the requirement that a learning rule do at least as well as the best response to the empirical distribution regardless of the actual strategy the opponent is employing (this implies both safety and consistency) and show that a modification of the fictitious play algorithm achieves this requirement. In [Fudenberg and Levine, 1998] they strengthen their requirement by requiring that the learning rule also adapt to simple patterns in the play of its opponent. The requirement of 'universal consistency' is in fact equivalent to requiring that an algorithm exhibit *no-regret*, generally defined as follows, against all opponents.

$$\forall \epsilon > 0, (lim_{t\rightarrow \inf}[\frac{1}{t}\max_{a_j \in A_i} r_i^t(a_j, s_i|s_{-i})] < \epsilon)$$

In both game theory and artificial intelligence, a large number of algorithms have been show to satisfy universal consistency or no-regret requirements. In addition, recent work [Bowling, 2005] has tried to combine these criteria resulting in GIGA-WoLF, a no-regret algorithm that provably achieves convergence to a Nash equilibrium in self-play for games with two players and two actions per player. Meanwhile, the regret matching algorithm [Hart and Mas-Colell, 2000] described earlier guarantees that the empirical distributions of play converge to the set of correlated equilibria of the game.

Other recent work by [Banerjee and Peng, 2005] has addressed concerns with only requiring guarantees about the behavior in the limit. Their algorithm is guaranteed to achieve ϵ-no-regret payoff guarantees with small polynomial bounds on time and uses only the agent's ability to observe what payoff it receives for each action.

4.3 Some observations and questions

We have so far described the work without comment; here we take a step back and ask some questions about this representative work.

Our first comment concerns the settings in which the results are presented. While the learning procedures apply broadly, the results for the most part focus on self play (that is, when all agents adopt the learning procedure under consideration). They also tend to focus on games with only two agents. Why does most of the work have this particular focus? Is it technical convenience, or is learning among more than two agents, or among agents using different learning procedures, less relevant for some reason?

Our second comment pertains to the nature of the results. With the exception of the work on no-regret learning, the results we described investigate convergence to equilibrium play of the stage game (albeit with various twists). Is this the pertinent yardstick? If the process (say of self play between two agents) does not converge to equilibrium play, should we be disturbed? More generally, and again with the exception of no-regret learning, the work focuses on the play to which the agents converge, not on the payoffs they obtain. Which is the right focus?

No-regret learning is distinguished by its starting with criteria for successful learning, rather than a learning procedure. The question one might ask is whether the particular criteria are adequate. In particular, the requirement of consistency ignores the basic lesson regarding learning-vs.-teaching discussed in Section 3. By measuring the performance only against stationary opponents, we do not allow for the possibility of teaching opponents. Thus, for example, in an infinitely repeated Prisoners' Dilemma game, no-regret dictates the strategy of always defecting, precluding the possibility of cooperation (for example, by the mutually reinforcing Tit-For-Tat strategies).

Our goal here is not to critique the existing work, but rather to shine a spotlight on the assumptions that have been made, and ask some questions that get at the basic issues addressed, questions which we feel have not been discussed as clearly and as explicitly as they deserve. In the next section we propose an organized way of thinking about these questions.

5 Five distinct agendas in multi-agent learning

After examining the MAL literature – the work surveyed here and much else – we have reached the conclusion that there are several distinct agendas at play, which are often left implicit and conflated. We believe that a prerequisite for success in the field is to be very explicit about the problem being addressed. We ourselves can identify five distinct possible goals of MAL research. There may well be others, but these are the ones we can identify. They each have a clear motivation and a success criterion that will allow researchers to evaluate new contributions, even if people's judgments may diverge regarding their relative importance or success to date. They can be caricatured as follows:

1. Computational
2. Descriptive
3. Normative
4. Prescriptive, cooperative
5. Prescriptive, non-cooperative

We can now consider each of the five in turn.

The first agenda is computational in nature. It views learning algorithms as an iterative way to compute properties of the game, such as solution concepts. As an example, fictitious play was originally proposed as a way of computing a sample Nash equilibrium for zero-sum games [Brown, 1951], and replicator dynamics has been proposed for computing a sample Nash equilibrium in symmetric games. Other adaptive procedures have been proposed more recently for computing other solution concepts (for example, computing equilibria in local-effect games [Leyton-Brown and 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 second agenda is descriptive – it asks how natural agents learn in the context of other learners. The goal here is to investigate formal models of learning that agree with people's behavior (typically, in laboratory experiments), or possibly with the behaviors of other agents (for example, animals or organizations). This same agenda could also be taken to apply to large-population models, if those are indeed interpreted as representing populations. This problem is clearly an important one, and when taken seriously calls for strong justification of the learning dynamics being studied. One approach is to apply the experimental methodology of the social sciences. There are several good examples of this approach in economics and game theory, for example [Erev and Roth, 1998] and [Camerer et al., 2002]. There could be other supports for studying a given learning process. For example, to the extent that one accepts the Bayesian model as at least an idealized model of human decision making, one could justify Kalai and Lehrer's model of rational learning.[7] However, it seems to us that sometimes there is a rush to investigate the convergence properties, motivated by the wish to anchor the central notion of game theory in some process, at the expense of motivating that process rigorously.[8] In this connection see also a recent critique [Rubinstein, 2005] of some work carried out more generally in behavioral economics.

The centrality of equilibria in game theory underlies the third agenda we identify in MAL, which for lack of a better term we called normative, and which focuses on determining which sets of learning rules are in equilibrium with each other. More precisely, we ask which repeated-game strategies are in equilibrium; it just so happens that in repeated games, most strategies embody a learning rule of some sort. For example, we can ask whether fictitious play and Q-learning, appropriately initialized, are in equilibrium with each other in a repeated Prisoner's Dilemma game. Although one might expect that game theory purists might flock to this approach, there are very few examples of it. In fact, the only example we know originates in AI rather than game theory [Brafman and Tennenholtz, 2002], and it is explicitly rejected by at least some game theorists [Fudenberg and Kreps, 1993]. We consider it a legitimate normative theory. Its practicality depends on the complexity of the stage game being played and the length of play; in this connection see our discussion of the problematic role of equilibria in Section 3.

The last two agendas are prescriptive; they ask how agents *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 modelled 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 rarely a role for equilibrium analysis; the agents have no freedom to deviate from the prescribed algorithm. Examples of this work include [Guestrin et al., 2001], [Claus and Boutilier, 1998], and [Chang et al., 2004] to name a small sample. 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 computing.

In our final agenda, termed 'prescriptive, non-cooperative', we ask how an agent should act to obtain high reward in the repeated (and more generally, stochastic) 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 hap-

[7] Although this is beyond the scope of this article, we note that the question of whether one can justify the Bayesian approach in an interactive setting goes beyond the familiar contravening experimental data; even the axiomatic justification of the expected-utility approach does not extend naturally to the multi-agent case.

[8] It has been noted that game theory is somewhat unusual in having the notion of an equilibrium without associated dynamics that give rise to the equilibrium.[Arrow, 1986]

pens that this environment is characterized by the types of agents inhabiting it, agents who may do some learning of their own. The objective of this agenda is to identify effective strategies for environments of interest. An effective strategy is one that achieves a high reward in its environment, where one of the main characteristics of this environment is the selected class of possible opponents. This class of opponents should itself be motivated as being reasonable and containing opponents of interest. Convergence to an equilibrium is not a goal in and of itself.

There are various possible instantiations of the term 'high reward'. One example is the no-regret line of work which we discussed. It clearly defines what it means for a reward to be high enough (namely, to exhibit no regret); we also discussed the limitations of this criterion. A more recent example, this one from AI, is [Bowling and Veloso, 2001]. This work puts forward two criteria for any learning algorithm in a multi-agent setting: (1) The learning should always converge to a stationary policy, and (2) if the opponent converges to a stationary policy, the algorithm must converge to a best response. There are possible critiques of these precise criteria. They can be too weak since in many cases of interest the opponents will not converge on a stationary strategy. And they can be too strong since attaining a precise best response, without a constraint on the opponent's strategy, is not feasible. But this work, to our knowledge, marks the first time a formal criterion was put forward in AI.

A third example of the last agenda is our own work in recent years. In [Powers and Shoham, 2005b] we define a criterion parameterized by a class of 'target opponents'; with this parameter we make three requirements of any learning algorithm: (1) (Targeted optimality) The algorithm must achieve an ϵ-optimal payoff against any 'target opponent', (2) (Safety) The algorithm must achieve at least the payoff of the security level strategy minus ϵ against any other opponent, and (3) (Auto-compatibility) The algorithm must perform well in self-play (the precise technical condition is omitted here). We then demonstrate an algorithm that provably meets these criteria when the target set is the set of stationary opponents in general-sum two-player repeated games. More recent work has extended these results to handle opponents whose play is conditional on the recent history of the game [Powers and Shoham, 2005a] and settings with more than two players [Vu et al., 2006].

6 Summary

In this article we have made the following points:

1. Learning in MAS is conceptually, not only technically, challenging.

2. One needs to be crystal clear about the problem being addressed and the associated evaluation criteria.

3. For the field to advance one cannot simply define arbitrary learning strategies, and analyze whether the resulting dynamics converge in certain cases

to a Nash equilibrium or some other solution concept of the stage game. This in and of itself is not well motivated.

4. We have identified five coherent agendas.

5. Not all work in the field falls into one of these buckets. This means that either we need more buckets, or some work needs to be revisited or reconstructed so as to be well grounded.

There is one last point we would like to make, which didn't have a natural home in the previous sections, but which in our view is important. It regards evaluation methodology. We have focused throughout the article on formal criteria, and indeed believe these to be essential. However, as is well known in computer science, many algorithms that meet formal criteria fail in practice, and vice versa. And so we advocate complementing the formal evaluation with an experimental one. We ourselves have always included a comprehensive bake-off between our proposed algorithms and the other leading contenders across a broad range of games. The algorithms we coded ourselves; the games were drawn from GAMUT, an existing testbed (see [Nudelman et al., 2004] and http://gamut.stanford.edu). GAMUT is available to the community at large. It would be useful to have a learning-algorithm repository as well.

To conclude, we re-emphasize the statement made at the beginning: This article is meant to be the beginning of a discussion in the field, not its end.

References

[Arrow, 1986] Arrow, K. (1986). Rationality of self and others in an economic system. *Journal of Business*, 59(4).

[Banerjee and Peng, 2005] Banerjee, B. and Peng, J. (2005). Efficient no-regret multiagent learning. In *Proceedings of the Twentieth National Conference on Artificial Intelligence.*

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

[Billings et al., 2003] Billings, D., Burch, N., Davidson, A., Holte, R., Schaeffer, J., Schauenberg, T., and Szafron, D. (2003). Approximating game-theoretic optimal strategies for full-scale poker. In *The Eighteenth International Joint Conference on Artificial Intelligence.*

[Blackwell, 1956] Blackwell, D. (1956). Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 336–338. North-Holland.

[Bowling, 2005] Bowling, M. (2005). Convergence and no-regret in multiagent learning. In *Advances in Neural Information Processing Systems 17.* MIT Press.

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

[Brafman and Tennenholtz, 2000] Brafman, R. and Tennenholtz, M. (2000). A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence*, 121(1-2):31–47.

[Brafman and Tennenholtz, 2002] Brafman, R. and Tennenholtz, M. (2002). Efficient learning equilibrium. In *Advances in Neural Information Processing Systems*, volume 15, Cambridge, Mass. MIT Press.

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

[Camerer et al., 2002] 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 et al., 2004] Chang, Y.-H., Ho, T., and Kaelbling, L. P. (2004). Mobilized ad-hoc networks: A reinforcement learning approach. In *1st International Conference on Autonomic Computing (ICAC 2004)*, pages 240–247.

[Cheng et al., 2005] Cheng, S.-F., Leung, E., Lochner, K. M., O'Malley, K., Reeves, D. M., Schvartzman, L. J., and Wellman, M. P. (2005). Walverine: A walrasian trading agent. *Decision Support Systems*, 39:169–184.

[Claus and Boutilier, 1998] 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*, pages 746–752.

[Erev and Roth, 1998] 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.

[Foster and Vohra, 1999] Foster, D. and Vohra, R. (1999). Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–36.

[Freund and Schapire, 1995] Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Proceedings of the Second European Conference*, pages 23–37. Springer-Verlag.

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

[Fudenberg and Levine, 1995] Fudenberg, D. and Levine, D. (1995). Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089.

[Fudenberg and Levine, 1998] Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games.* MIT Press, Cambridge, MA.

[Greenwald and Hall, 2003] Greenwald, A. and Hall, K. (2003). Correlated Q-learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 242–249.

[Guestrin et al., 2001] Guestrin, C., Koller, D., and Parr, R. (2001). Multiagent planning with factored mdps. In *Advances in Neural Information Processing Systems (NIPS-14).*

[Hannan, 1957] Hannan, J. F. (1957). Approximation to Bayes risk in repeated plays. *Contributions to the Theory of Games*, 3:97–139.

[Hart and Mas-Colell, 2000] Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150.

[Hu and Wellman, 2003] Hu, J. and Wellman, M. (2003). Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069.

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

[Jafari et al., 2001] Jafari, A., Greenwald, A., Gondek, D., and Ercal, G. (2001). On no-regret learning, fictitious play, and Nash equilibrium. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 226–223.

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

[Kaelbling et al., 1996] Kaelbling, L. P., Littman, M. L., and Moore, A. P. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285.

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

[Kapetanakis and Kudenko, 2004] Kapetanakis, S. and Kudenko, D. (2004). Reinforcement learning of coordination in heterogeneous cooperative multi-agent systems. In *Proceedings of the Third Autonomous Agents and Multi-Agent Systems conference.*

[Kearns and Singh, 1998] Kearns, M. and Singh, S. (1998). Near-optimal reinforcement learning in polynomial time. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 260–268.

[Koller and Pfeffer, 1997] Koller, D. and Pfeffer, A. (1997). Representations and solutions for game-theoretic problems. *Artificial Intelligence*, 94(1):167–215.

[Lauer and Riedmiller, 2000] Lauer, M. and Riedmiller, M. (2000). An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the 17th International Conference on Machine Learning*, pages 535–542. Morgan Kaufman.

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

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

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

[Littman and Szepesvari, 1996] 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*, pages 310–318.

[Mannor and Shimkin, 2003] Mannor, S. and Shimkin, N. (2003). The empirical bayes envelope and regret minimization in competitive markov decision processes. *Mathematics of Operations Research*, 28(2):327–345.

[Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

[Miyasawa, 1961] Miyasawa, K. (1961). On the convergence of learning processes in a 2x2 non-zero-person game. *Research Memo 33*.

[Nachbar, 1990] Nachbar, J. (1990). Evolutionary selection dynamics in games: Convergence and limit properties. *International Journal of Game Theory*, 19:59–89.

[Nudelman et al., 2004] Nudelman, E., Wortman, J., Leyton-Brown, K., and Shoham, Y. (2004). Run the GAMUT: A comprehensive approach to evaluating game-theoretic algorithms. *Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*.

[Powers and Shoham, 2005a] Powers, R. and Shoham, Y. (2005a). Learning against opponents with bounded memory. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*.

[Powers and Shoham, 2005b] Powers, R. and Shoham, Y. (2005b). New criteria and a new algorithm for learning in multi-agent systems. In *Advances in Neural Information Processing Systems 17*. MIT Press.

[Robinson, 1951] Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, 54:298–301.

[Rubinstein, 2005] Rubinstein, A. (2005). Comments on behavioral economics. In *Proceedings in Advances in Economic Theory (2005 World Congress of the Econometric Society)*.

[Schuster and Sigmund, 1983] Schuster, P. and Sigmund, K. (1983). Replicator dynamics. *Journal of Theoretical Biology*, 100:533–538.

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

[Shoham et al., 2004] Shoham, Y., Powers, R., and Grenager, T. (2004). On the agenda(s) of research on multi-agent learning. In *AAAI 2004 Symposium on Artificial Multi-Agent Learning [FS-04-02]*. AAAI Press.

[Smith, 1982] Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

[Vu et al., 2006] Vu, T., Powers, R., and Shoham, Y. (2006). Learning against multiple opponents. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multi Agent Systems*.

[Wang and Sandholm, 2002] Wang, X. and Sandholm, T. (2002). Reinforcement learning to play an optimal Nash equilibrium in team markov games. In *Advances in Neural Information Processing Systems*, volume 15.

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

[Young, 2004] Young, H. P. (2004). *Strategic Learning and Its Limits*. Oxford University Press.

[Zinkevich, 2003] Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*.