International Game Theory Review, Vol. 3, Nos. 2 & 3 (2001) 159–174

ASPIRATION-BASED REINFORCEMENT LEARNING IN REPEATED INTERACTION GAMES: AN OVERVIEW

JONATHAN BENDOR*

Graduate School of Business, Stanford University, 518 Memorial Way, Stanford, CA 94305-5015, USA

DILIP MOOKHERJEE†

Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, USA

DEBRAJ RAY‡

Department of Economics, New York University, 269 Mercer St, NY 10003, USA

In models of aspiration-based reinforcement learning, agents adapt by comparing payoffs achieved from actions chosen in the past with an aspiration level. Though such models are well-established in behavioural psychology, only recently have they begun to receive attention in game theory and its applications to economics and politics. This paper provides an informal overview of a range of such theories applied to repeated interaction games. We describe different models of aspiration formation: where (1) aspirations are fixed but required to be consistent with longrun average payoffs; (2) aspirations evolve based on past personal experience or of previous generations of players; and (3) aspirations are based on the experience of peers. Convergence to non-Nash outcomes may result in either of these formulations. Indeed, cooperative behaviour can emerge and survive in the long run, even though it may be a strictly dominated strategy in the stage game, and despite the myopic adaptation of stage game strategies. Differences between reinforcement learning and evolutionary game theory are also discussed.

1. Introduction

Traditional game theory is based on strong assumptions concerning rationality and common knowledge [Aumann and Brandenburger (1995) and Tan and Werlang (1988)] that appear implausible as descriptions of the behaviour of real world players. Many experimental settings produce substantive departures from the predictions of the theory [Roth (1995)]. Accordingly, recent developments of the

*E-mail: bendor_jonathan@gsb.stanford.edu
†E-mail: dilipm@bu.edu
‡E-mail: debraj.ray@nyu.edu

theory have explored boundedly rational behaviour of various kinds, such as models of learning and experimentation, and evolutionary processes; surveys of this literature appear in Fudenberg and Levine (1998), Marimon (1997), Samuelson (1997), Young (1998) and Vega-Redondo (1996).

However, relatively little theoretical attention has been directed towards one particular form of boundedly rational behaviour supported by a number of experimental studies — *reinforcement learning*, according to which players *satisfice* rather than *maximise* payoffs. Specifically, players select current actions based on past experience with playing the same game. Actions judged to have resulted in satisfactory payoffs are more likely to be selected, while those resulting in unsatisfactory payoffs tend to be dropped. The standard by which experiences are judged to be satisfactory is defined by an aspiration level, against which achieved payoffs are evaluated. Aspirations are socially inherited, or evolve slowly as a player gains experience. Such behaviour rules are informationally and cognitively less demanding than those entailing maximisation of some utility function. They appear particularly plausible when players are unable to form a coherent "model" of their environment. For instance, they might lack information concerning the strategic structure of the game or the actions selected by others in the past. Alternatively, they may not process such information even when it is available.

To illustrate the contrast beween the different approaches, consider the context of a repeated oligopoly, where a given number of firms set prices or quantities of their respective products at successive dates. Traditional game theory assumes that firms know demand and cost functions (or have Bayesian priors over these), and thereafter select supergame strategies which condition their actions on the past history of actions taken by all firms. These strategies, information and rationality of all firms is common knowledge among them, and each firm selects a best response to the strategies of others. Models based on an intermediate level of rationality, such as "belief learning",[a] drops the assumption that strategies are common knowledge. It is postulated instead that each firm predicts the future actions of other firms based on observation of past plays, and then selects a myopic best response to these beliefs by maximising their expected payoff function. In contrast, reinforcement learning models allow firms to be ignorant of demand and cost functions, and adapt their actions over time in line with their profit experience with different actions in the past. In particular, firms do not attempt to maximise some measure of expected profit, owing to the difficulty of articulating probabilistic beliefs concerning demand patterns or competitors' actions, and the complexity of the associated optimisation problem. Instead, actions that generated high (respectively low) profits in the past are apt to be repeated (dropped). Actions thus get "selected" on the basis of their past "fitness", suggesting some similarity with evolutionary game theory. As explained in further detail below,

[a]See Selten (1991) for an exposition of the distinction between belief learning and reinforcement learning.

however, the results of this theory are markedly distinct from those of evolutionary game theory, partly owing to the different interaction patterns underlying the two theories.[b]

Reinforcement learning models first appeared in the mathematical psychology literature: see Estes (1954), Bush, *et al.* (1954), Bush and Mosteller (1955), and Suppes and Atkinson (1960). Similar models were studied for their normative properties as representing modes of automata learning in the computer science and electrical engineering literature: see Lakshmivarahann (1981), Narendra and Mars (1983), Narendra and Thathachar (1989), and Papavassilapoulos (1989). Amongst economists, early pioneers of these models include Simon (1955, 1957, 1959), Cross (1973) and Nelson and Winter (1982). More recently, Gilboa and Schmeidler (1995) have developed an axiomatic basis for such an approach. Experimental support of the reinforcement learning hypothesis *vis-á-vis* the traditional rational play hypothesis and belief learning is discussed in Selten and Stoecker (1986), Selten (1991), Mookherjee and Sopher (1994, 1997), Roth and Erev (1995), Kim (1995a, b), Erev and Roth (1998) and Camerer and Ho (1999).

Despite the interest in this hypothesis in experimental settings, relatively little theoretical effort has been devoted to the study of its long run implications. This paper provides an informal overview of our recent work on this topic, as well as some related literature. The models differ chiefly in their treatment of the process by which aspirations get formed. Applications to games of cooperation, coordination, competition and bargaining are subsequently discussed. We focus particularly on repeated games, and emphasise some distinctive features of reinforcement learning processes that emerge from this literature:

(i) Dominated strategies can be frequently played (a phenomenon frequently observed in experimental settings) and can survive in the long run, a result difficult to reconcile with rational choice, belief learning or evolutionary models.
(ii) Players learn to cooperate in the Prisoners Dilemma (as well as games of common interest, where one outcome Pareto dominates all others); under particular circumstances, this is the unique long run outcome. In others, the set of possible long run outcomes can be narrowed down to the mutual defection and mutual cooperation outcomes.

[b]Borgers and Sarin (1997a) explore the connection between reinforcement learning and the replicator dynamic that underlies most evolutionary models within a special class of games, and show that they are equivalent only under special circumstances. Apart from the differences between discrete time and continuous time formulations in which these two hypotheses are respectively embedded, the replicator dynamic applies to the relative frequency of different strategies in a large population that "plays itself". Reinforcement learning is applied instead to strategy choices of a single player. Hence, individual players that are reinforcement learners may fail to converge to expected payoff maximising strategies, as they tend to shift away from them following occasional unsuccessful payoff experiences. In a large population, the strategy plays other strategies a large number of times, and so obtains a higher average score than others, causing it to grow relative to other strategies.

(iii) Players learn to coordinate on Pareto-efficient (rather than risk-dominant) outcomes in coordination games.
(iv) In competitive zero-sum games players learn to play maxmin pure strategies (if these happen to exist).
(v) Norms of fairness may emerge in "population" bargaining games.

The concluding section assesses the state of this literature and possible directions for future research.

2. Modeling Reinforcement Learning

Consider a game played repeatedly by two players denoted A and B. The stage game has finite sets of pure strategies $\mathcal{A}$ and $\mathcal{B}$ and payoff functions $f(a,b), g(a,b)$ for the two players, respectively. The game is played at successive dates $t = 1, 2, \ldots$. The *state* of a player at any date includes the player's psychological inclination to select different actions at that date, based on past experience, and is represented either by a pure strategy or a mixed strategy. Borgers, *et al.* (1998) analyse the class of learning rules that enable agents to learn to play expected payoff maximising actions in the long run in a decision environment, and show that this typically requires mixed strategy states. For both this normative reason, as well as descriptive plausibility, the mixed strategy formulation is more natural. The only justification for the pure strategy formulation is its analytic tractability.

The state is updated from one date to another on the basis of the player's payoff *experience*, which is judged relative to the player's *aspiration*. Aspirations themselves may be updated over time based on past experience, in which case the state space includes players' aspiration levels in addition to their strategies. Different models vary regarding their formulation of the state space, players' experience and the updating rules. But they all tend to embody the psychological principle that actions experiencing successful (respectively unsuccessful) payoffs (relative to aspirations) tend to be reinforced (respectively discouraged), i.e. the probability weight on those actions are revised upwards (respectively downwards). Most formulations also postulate that strategy revisions additionally incorporate some degree of *inertia*, i.e. a bias in favour of actions recently selected, and *experimentation*, i.e. some small probability weight is placed on every action, regardless of past payoff experience.

These updating rules define a Markov process over the state space, and the focus is on long run properties of this process. Below, we describe the principal results of alternative formulations of this model in our research, and then discuss related literature.

3. Constant Consistent Aspirations

Bendor, Mookherjee and Ray (hereafter denoted as BMR) (1992, 1995) adopted a mixed strategy formulation in which each player has a constant aspiration level,

which does not evolve in the course of play. Hence the players' mixed strategies constitute the state variable.[c] With small degrees of experimentation, the induced Markov process over players' strategies is ergodic. The constant aspiration level is required to be consistent with the induced long run average payoffs. This notion represents a steady state of an unmodelled dynamic process of aspiration revision, in which players inherit aspirations from payoff experiences of past generations of players. In any steady state of such a process, aspiration levels must be mirrored by the payoff experience of any given generation in order for succeeding generations to inherit the same aspirations. For any given degree of experimentation, such consistent aspiration-strategy distribution pairs were shown to generally exist, assuming that the strategy updating rule is continuous with respect to the (fixed) aspiration level. They defined an *Equilibrium with Consistent Aspirations* (*ECA*) to be the limit of consistent aspirations-strategy distribution pairs, as the degree of experimentation converges to zero.

The key result of BMR (1992, 1995) was the existence of ECAs concentrated entirely on cooperative strategies in symmetric games. Consider for instance the following class of 2×2 games:

	C	D
C	(σ, σ)	$(0, \theta)$
D	$(\theta, 0)$	(δ, δ)

where $\sigma > \delta > 0$ and $0 \leq \theta \neq \delta$. This includes games of common interest ($\sigma > \theta$), coordination ($\theta = 0$) and cooperation (it is a Prisoners Dilemma if $\theta > \sigma$). In this game there is an ECA concentrated entirely on (C, C), the pure strategy pair involving cooperation. The outline of the argument in the case of a Prisoners Dilemma is as follows. Suppose that both players have an aspiration slightly below σ. Mutual cooperation is the only strategy pair where both attain their aspirations. In the absence of any experimentation, this is the only absorbing state. In all other states at least one player is dissatisfied with positive probability, causing a revision of the state. In the presence of small degrees of experimentation, starting from the mutually cooperative state, one player may occasionally deviate from the cooperative strategy. While the deviator benefits from this change, the other player obtains a payoff below her aspiration, and thereafter deviates as well. With neither cooperating, both are dissatisfied and thus inclined to try the cooperative strategy again. Sooner or later they switch simultaneously to cooperation, whence they are both satisfied, and tend to get absorbed into this state again. The long run outcome is thus concentrated on the mutually cooperative state, *given* that their aspirations are throughout held fixed slightly below (σ, σ). And these aspirations are justified

[c]BMR (1992) restricted probabilities to lie on a finite grid, thus ensuring that the state space is finite. BMR (1995) this restriction, so the state space for each player is the entire probability simplex. Besides this aspect, the 1995 paper extended and generalised the results of the 1992 paper in a number of directions.

in turn by the long run outcome of the process, i.e. cooperation most of the time, with occasional forays of experimenting with deviations by either or both players.

This example illustrates how long run outcomes may be concentrated on non-Nash outcomes of the stage game, which may even involve play of strictly dominated pure strategies. The contrast with belief learning or the replicator dynamic is apparent. In belief learning, players consciously maximise their payoff function, and so can never play a strictly dominated strategy. Similarly, the replicator dynamic cannot converge to any distribution assigning positive weight to a strictly dominated strategy, as it would grow less quickly than the corresponding strategy that dominates it, thus tending to become extinct in the long run. Only evolutionary settings where players select automata or supergame strategies permit the evolution of cooperative behaviour [Axelrod (1984), Bendor and Swistak (1997), Binmore and Samuelson (1992)].

4. Evolving Aspirations Based on Personal Experience

The problem with the preceding theory is that it does not model the dynamics of aspiration adjustment. Karandikar, Mookherjee, Ray and Vega-Redondo (hereafter denoted KMRV) (1998) explored the consequences of aspirations that evolve in the course of play.[d] Then players' aspiration levels become part of the state. They assume that if a player's aspiration at date t is A_t, then it is updated according to $A_{t+1} = \lambda A_t + (1 - \lambda)\pi_t$, where $\lambda \in (0, 1)$ is a persistence parameter, and π_t is the payoff experienced at t.[e] Hence players' aspirations are based on their personal payoff experience in past plays. In addition, aspirations of either player are subject to small idiosyncratic shocks: with probability $1 - \eta$ they evolve according to the deterministic rule above. Otherwise, it is perturbed according to some positive density parametrised continuously by the previous aspiration level A_t with a compact support contained within the range of feasible payoffs. These random aspiration perturbations are independent across time and across players, and prevent players from ever settling down at some pure strategy state (with aspirations equal to the corresponding payoffs).

To keep the analysis tractable, KMRV focused on the class of 2×2 games described in the previous section. Moreover they assumed that players behaviour states are represented by pure strategies. The strategy updating rule was as follows: if a player is satisfied ($\pi_t \geq A_t$) at date t, then she sticks to the same strategy at the following date. Otherwise, she switches to the alternate strategy with a probability that depends continuously on the extent of dissatisfaction, but is bounded away from 1. The latter assumption incorporates *inertia*, while *experimentation* is induced by the aspiration perturbations. The main result of their paper is that the

[d]Borgers and Sarin (1997b) study a model of endogenous aspirations in a single person decision context, where the behavioural state variable is represented by a mixed strategy.

[e]For an early model of aspiration adjustment that used just these assumptions, see Cyert and March (1963).

long run outcome involves both players cooperating most of the time, if the speed of updating aspirations is sufficiently slow (λ is sufficiently close to 1), and the probability of aspiration trembles η is sufficiently low.[f]

Hence, the result concerning the play of dominated cooperative strategies is upheld when aspirations evolve in the course of the game. In fact, the result is even sharper: it is the *only* long run outcome of the game, regardless of initial conditions. In BMR (1992, 1995) in contrast, it was one possible ECA, and other possible ECAs of the game could not be ruled out, even with special assumptions about the nature of the updating rules. In particular, aspirations in KMRV tend to lie at or slightly below the cooperative payoffs most of the time. Aspirations tend to drift downwards only following unfavourable payoff experiences resulting from deviations by one or both players. But as players remain dissatisfied with such experience, they are inclined to switch back to cooperative strategies. Once they do, they are both satisfied, so they tend to stick with these strategies, returning aspirations slowly to their previous level (until the next aspiration perturbation causes another deviation). And when players start with low aspirations, they are prevented from settling into a low aspiration "trap" by the occasional aspiration perturbations, which motivates them to experiment with different strategies, including the cooperative one. Once they do simultaneously cooperate, they tend to be satisfied, and hence stick to these actions, serving to raise their aspirations over time. Of course this heuristic discussion omits many complications; the reader is invited to consult KMRV (Sec. 4) for an informal discussion of further details concerning the aspiration dynamic.

Related results were obtained by Kim (1995a) and Pazgal (1995), both of whom apply the Gilboa-Schmeidler case-based theory of reinforcement learning. Kim focused on a similar class of 2×2 games, whereas Pazgal examined a larger class of games of mutual interest, in which there is an outcome which strictly Pareto dominates all others. Actions are scored on the basis of their cumulative past payoff relative to the current aspiration, and players select only those actions with the highest score. Aspirations evolve in the course of play: aspirations average *maximal* experienced payoffs in past plays (in contrast to KMRV who assumed that they equal a geometric average of past payoffs). Both Kim and Pazgal show that cooperation necessarily results in the long run if initial aspiration levels lie in pre-specified ranges. Pazgal requires initial aspirations to be sufficiently high relative to the cooperative payoffs; Kim assumes that they are not much lower than these payoffs. With high aspirations, players will not be satisfied by non-cooperative strategies and thus experiment with cooperation. Once they do cooperate, their assumption concerning aspiration formation implies that they will aspire to cooperative payoffs thereafter. Given this, long run strategies must be entirely concentrated on cooperation.

[f]Specifically, limits of the resulting long run distribution are first taken with respect to $\eta \to 0$ for a given λ, and then λ is taken to 1.

Finally, Bendor, *et al.* (2000) build a model of turnout — electoral participation by citizens — in which the agents' aspirations adjust as in KMRV but mixed strategies are allowed. Given dynamical aspirations and mixed strategies, a complete analytical solution of the model is infeasible, and so computational modelling (supplemented by partial analytical results) is required. The main finding of their model is that the "paradox of voting" vanishes: citizens turn out in substantial numbers (in a typical simulation about half the population votes) even in large electorates ($N = 1,000,000$) and even when everyone has strictly positive costs of voting.

5. Aspirations Evolving on the Basis of Experience of Previous Generations of Players

BMR (2001) explore an extension of their earlier work by explicitly modelling the dynamics of aspirations. The game is assumed to be played by successive generations of players. Within any generation, a pair of players are picked to play the repeated game. The key assumption is that players of any given generation inherit a fixed aspiration level from the payoff experiences of previous generations; these aspirations are not modified during their *own* lifetime (though of course their payoff experiences will affect the aspirations their children will inherit). This represents the notion that while behaviour is conditioned by aspirations, the latter is subject to a more gradual process of "cultural" evolution.

The analytical benefit of this approach is that the dynamic of strategies and aspirations is sequenced rather than simultaneous, thus limiting the dimensionality of the state space, which renders the analysis tractable. In particular, the theory allows players' states to be represented as mixed strategies following very general updating rules, and can be applied to arbitrary two-person finite games. Moreover, as explained below, it permits a sharp and simple characterisation of long run outcomes, allowing the theory to be applied to a wider range of contexts than is represented by 2×2 games of cooperation or coordination.

Within any given generation with a fixed (inherited) aspiration, strategy updating rules follow three simple properties: an action which is positively reinforced (i.e. generates a payoff which exceeds the aspiration) has its probability weight increased; negative reinforcement (payoff below aspiration) induces experimentation with other actions, and inertia (which assigns an additional weight on the action currently chosen at the next round, and totally mixed strategies are mapped into totally mixed strategies). These updating rules generate a Markov process (over mixed strategy pairs), which we shall denote $\mathcal{S}(A)$ corresponding to aspirations A.

The dynamic over aspirations updates aspirations across rounds (or generations) on the basis of the discrepancy between aspirations and average payoffs achieved in the previous round. Specifically, if A_T denotes the aspiration in round T, then $A_{T+1} = \lambda A_T + (1 - \lambda)\Pi_T$, where $\lambda \in (0, 1)$ and Π_T is the average payoff of $\mathcal{S}(A_T)$. However, the process $\mathcal{S}(A_T)$ may not be ergodic, so the average payoff may

not be uniquely defined. For instance, in the Prisoners Dilemma game depicted above, if both players have aspirations below δ there are two long run distributions, which are respectively concentrated on the pure strategy pairs (D, D) and (C, C). In that case, the average payoff of a generation with these aspirations is path dependent, hence inherently unpredictable. Accordingly, Π_T is a random variable. Only one restriction is imposed on Π_T: it assigns positive probability to the average payoff of any invariant distribution of $\mathcal{S}(A_T)$ which is stable against small doses of experimentation by players.

Specifically, suppose the process $\mathcal{S}(A_T)$ is perturbed to $\mathcal{S}_\epsilon(A_T)$, where players' strategies are subject to i.i.d. trembles with probability η.[g] The perturbed process $\mathcal{S}_\epsilon(A_T)$ is ergodic. The limit of the corresponding ergodic distribution along any convergent subsequence $\eta \to 0$ is an invariant distribution of the original unperturbed process $\mathcal{S}(A_T)$, and thus a mixture of different non-communicating invariant distributions of the latter process (in each of which all states in the support communicate with one another). These are the "stable" invariant distributions of $\mathcal{S}(A_T)$, and each of them is postulated to receive positive weight in Π_T.

Aspirations now follow a Markov process. In two relevant cases, they can be shown to converge to a degenerate distribution, i.e. a limit which is deterministic with respect to both behaviour and aspirations: (i) if the game is symmetric, with a symmetric Pareto-efficient pure payoff point, and players of any given generation inherit the same aspiration; or in more general (possibly asymmetric) games if (ii) aspirations of the first generation are *intermediate*, i.e. bounded above by some pure strategy payoff, and are individually rational (i.e. bounded below by pure strategy maxmin payoffs).

The long run solution of the game can thus be identified with a deterministic steady state aspiration of this process (where in addition both players select a particular pair of pure strategies). This bears a close relation to the (pure strategy version of) ECA notion in BMR (1992, 1995). Such steady states are shown to be equivalently represented as follows. Say that a distribution μ is a long run outcome from initial state γ and aspirations A, if starting from γ the sequence of probability measures over the state space (of strategies) generated by $\mathcal{S}(A)$ converges weakly to μ. Then distribution μ is said to be *consistent* if (i) from every state γ in its support, μ is a long run outcome from γ and aspirations A which equals the expected payoff under μ; and (ii) average payoffs converge almost surely to A.

In addition to consistency, the solution must be stable with respect to small doses of experimentation. It suffices to check for the following forms of experimentation, which we call a *single random perturbation*. Take any state (α, β) in the support of the distribution μ, select one of the two players randomly, and alter the behaviour state of that player to a neighbouring (totally mixed) state. Checking for stability requires that we go through the following steps:

[g]With probability ϵ, the player's strategy is set equal to a totally mixed strategy in the neighbourhood of the strategy that would have been chosen otherwise.

(1) Find the set of all long run distributions μ' that can be reached from μ (with aspirations fixed at A) following a single random perturbation.
(2) Check that every such μ' generates an average payoff vector of A.
(3) Finally, starting from any such μ', it is possible to return to μ following a sequence of single random perturbations (i.e. there is a sequence of long run distributions $\mu^1, \mu^2, \ldots, \mu^N$ with $\mu^1 = \mu'$ and $\mu^N = \mu$, such that μ^k can be reached from μ^{k-1} following a single random perturbation).

Of interest, therefore, are stable distributions with the additional property that they are concentrated entirely on a pair of pure strategies, since these correspond to the deterministic steady states of the process. These are called *pure stable outcomes* (*pso*). BMR (2001) subsequently provide an almost complete characterisation of pso's. To explain this characterisation, the following definitions are necessary.

A pure strategy pair is said to be *individually rational* if each player earns a payoff at least as large as his (pure strategy) maxmin payoff. It is said to be *strictly individually rational* if each player's payoff strictly exceeds the (pure strategy) maxmin payoff. It is a *protected Nash equilibrium* if it is a Nash equilibrium with the additional property that unilateral deviations by any player can neither hurt the other player, nor generate a (weak) Pareto improvement.[h] It is a *protected strict Nash equilibrium* if it is a protected Nash equilibrium which is also a strict Nash equilibrium. Finally, a pure strategy pair is *efficient* if there does not exist any other pure strategy pair which weakly Pareto dominates it.

The characterisation of pso's is the following. *Every pso is individually rational, and* **either** *a protected Nash equilibrium*, **or** *efficient*. Conversely, a pure strategy pair is a pso if it is **either** strictly individually rational and efficient, **or** a strict Nash equilibrium. This implies that a pso is inefficient only if it is a protected Nash equilibrium, and inefficient pso's can arise.

An example of an inefficient protected (strict) Nash equilibrium is mutual defection (D, D) in the Prisoners Dilemma. Here a deviation by any player to C benefits the opponent and hurts the deviator. Hence the opponent will be satisfied to stick to the pure strategy D while the deviator experiments with the deviation. In turn, this ensures that the deviator must be dissatisfied with the deviation, and must eventually return to D. The outcome is therefore stable with respect to unilateral deviations. In contrast if the game is one of coordination, so $\theta = 0$, (D, D) is no longer protected. Then a deviation by one player hurts *both* agents, motivating both to experiment with other actions. With positive probability, they will simultaneously experiment with C. Once they do so, they will be positively reinforced; with positive probability they will stick to (C, C) forever after, resulting in higher payoffs for both. Consequently (D, D), while a stable outcome in the Prisoners Dilemma, is unstable in the coordination game.

[h]By a weak Pareto improvement we mean that neither player is worse off, and at least one player is strictly better off.

Pareto efficient outcomes such as (C, C) are nevertheless stable pure outcomes in both games, since they are strictly individually rational in either game. Hence, players may continue to converge to dominated cooperative actions. But they may also converge to an inefficient outcome, such as (D, D) in the Prisoners Dilemma, provided it is a protected Nash equilibrium. The prediction is different from the KMRV analysis, in which aspirations of any given generation were permitted to evolve within their own lifetime. In that model, (D, D) is not stable in the Prisoners Dilemma because deviations by one player to C raise the payoff (hence aspiration) of the other player, so play cannot return to (D, D) as the latter would now be dissatisfied with this outcome.

These results permit applications to contexts of economic and political competition. Consider a duopoly involving two firms producing a homogenous or differentiated product, each with constant unit cost. The two firms either set quantities (the Cournot game) or prices (the Bertrand game) of their respective products. Assume that (i) both firms are free to exit the market and earn zero profits; (ii) there exist collusive actions (e.g. lower quantity or higher price than the competitive outcome) that generate positive profit for both; and (iii) given any such action pair there exists a deviation by one firm (e.g. sufficiently large quantity or sufficiently low price) that drives the other firm to a loss. Then maxmin profits are zero for both firms. Then every collusive (i.e. efficient from the standpoint of the two firms) outcome will be strictly individually rational, and hence constitute a pso. If there is any other pso, it must be a zero profit protected Nash equilibrium. An example of this is a competitive Bertrand equilibrium in the price-setting game with homogenous products, in which both firms price at unit cost. If such a zero profit equilibrium does not exist, as is typically the case in the Cournot game, or in the Bertrand game with differentiated products, then *every pso must involve maximal collusion*. This is despite the low order of rationality exhibited by the two firms and the lack of any conscious effort to collude.

Alternatively, consider a Downsian model of electoral competition between two political parties. A party's strategy is a selection of a policy platform from a finite set of points on the real line. There are a large number of voters, each with single peaked preferences over the set of platforms. If both parties select the same platform, they split the vote equally; otherwise citizens vote for the party closer to their ideal policy platform. Parties's payoffs are monotonically increasing in their share of the vote. Then the Downsian outcome, i.e. where both parties select the median voter's ideal policy, is the unique pso. The reason is that this is a zero-sum game in which the Downsian outcome is the unique Nash equilibrium. Hence, it has the saddle point property: if one party deviates from the median ideal point, the opponent's vote share must rise (as its own share falls). So it is a protected strict Nash equilibrium, implying that it constitutes a pso. Moreover, no other individually rational pure strategy pair is stable, as a single random perturbation of the behaviour state of one party can cause an increase in the long run vote share of that party (e.g. if it deviates to the Downsian platform and thereafter converges to

the pure strategy concentrated on that platform). It can additionally be shown to be the unique stable outcome, even when mixed strategy distributions are allowed. The Downsian insight concerning the "centrist" tendency inherent in two-party electoral competition is thus upheld even when parties exhibit the low order of rationality embodied in reinforcement learning behaviour. Extension of this result to alternative formulations of the nature of electoral competition and policy spaces is explored in BMR (2000).

6. Aspirations Based on Observed Experience of Peers, and Alternative Interaction Patterns

Aspirations may be formed also on the payoff experiences of one's own peers.[i] Dixon (2000) considers a set of identical (but separated) duopoly markets; in each market a given pair of firms repeatedly interact. The aspirations of any firm evolve in the course of the game, but are based on the profit experiences of firms across *all* the markets. Firms also randomly experiment with different actions. If the current action meets aspirations, then experimentation tends to disappear over time; otherwise they are bounded away from zero. In this model, play converges to joint profit maximising actions in all markets, regardless of initial conditions.

Palomino and Vega-Redondo (1999) consider a non-repeated game setting (akin to those studied in evolutionary game theory) where pairs are randomly selected from a large population in every period to play the Prisoners Dilemma. Aspirations of each player are based on the payoff experiences of the entire population. They show that in the long run, a positive fraction of the population will cooperate.

Ghosh (1998) studies a similar interaction pattern in which pairs are randomly selected from the population in every period to play a Nash demand game. Once a pair is selected, they must split one dollar. They independently submit minimum demands. If these demands add up to more than one, each gets nothing. Otherwise the dollar is split halfway between their respective demands. Ghosh assumes that players demand their current aspiration level. Each player updates his aspirations on the basis of his own experience at the previous round: it moves down if there was no agreement, otherwise it is increased. Irrespective of the distribution of initial aspirations in the population, Ghosh shows that the distribution converges to the degenerate one concentrated at the "fair" outcome involving a fifty-fifty split of the dollar. This is the Nash bargaining solution in this setting, one of a continuum of Nash equilibria of the game. The striking feature of this result is that it is obtained without presuming either that players experiment with alternative strategies or that the game itself is perturbed.

[i]Sociologists, who have studied this phenomenon empirically for quite some time, call a set of such peers a *reference group* [Merton and Rossi (1950)].

7. Concluding Comments

The small but growing literature surveyed here suggests that reinforcement learning is worthy of interest not just for its relevance in interpreting experimental evidence. Theoretical analysis of its long run properties indicates that in repeated interaction settings, these learning processes are marked by a number of distinct features, even relative to evolutionary models.[j] The most striking of these is the general possibility — indeed inevitability in certain settings — of the emergence of cooperative behaviour. This is despite the possibility that cooperative strategies may be strongly dominated in the stage game (as in the Prisoners Dilemma), and despite the fact that players learn from experience in an adaptive and myopic fashion. In particular, non-Nash outcomes will often occur in the long run, despite the propensity of players to constantly experiment with different actions. The reason is simple yet fundamental: profitable unilateral deviations rarely remain unilateral for long periods of time if the deviation hurts the opponent, since this provokes the latter to respond by switching to alternative strategies.

At least two aspects of these models differentiate them from evolutionary models. One is the interaction pattern: repeated interaction between a few players, rather than random matching of players from a large population. The other is the importance of aspirations in adapting strategies to past experience. The assumptions underlying most evolutionary models are influenced by the biological context in which such theories originally arose, and so may be less appropriate in many contexts involving human agents.

Obviously, the contexts studied so far in the literature on reinforcement learning are also stylised with a limited range of applications. This necessitates extending these models to a richer range of contexts, e.g. involving more than two players, and non-repeated patterns of interaction. This is particularly so as reinforcement learning processes seem more appropriate to contexts involving large populations where the complexity of modelling and tracking the behaviour of all the other participants is too great to justify devoting significant cognitive resources to such activities. This may include buying, selling or search behaviour of "small" traders in markets. For instance, it is a plausible description of buying and search behaviour by individual households for products involving a moderate fraction of total expenditures. It may also be a plausible description of pricing or product decisions made by firms that lack information concerning search or demand patterns of households, or the cost structure of other firms, as studied in Vishwanath (1994). In such settings it is difficult to imagine most market participants acting on the basis of a coherent "model" of the entire market and the incentives of all participants. Similar arguments can be made concerning behaviour of voters or citizens in political processes,

[j]It is worth clarifying this is so when both models are applied to the stage game strategies, rather than to automata or supergame strategies. While evolutionary analyses of the latter have been explored by Axelrod (1984), Bendor and Swistak (1997) and Binmore and Samuelson (1992), we are not familiar with any analysis of reinforcement learning applied to supergame strategies.

collective action problems (e.g. involving environmental issues) or other forms of social interaction such as traffic behaviour.

Alternative formulations of strategies and their revision processes incorporating different degrees of sophistication also need to be studied. For instance, players may revise strategies on the basis of more information concerning past payoff experience. The strategies themselves may include rules for reacting to past actions of others. Eventually, the degree of sophistication may itself be chosen on the basis of a higher order reinforcement learning process, as suggested by Selten (1978) and Stahl (1996).

References

Aumann, R. and A. Brandenburger (1995). "Epistemic Conditions for Nash Equilibrium". *Econometrica*, Vol. 63, 1161–1180.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Bendor, J., D. Mookherjee and D. Ray (1992). *Aspirations, Adaptive Learning and Cooperation in Repeated Games*. Discussion Paper, Planning Unit, Indian Statistical Institute, New Delhi.

——— (1995). *Aspirations, Adaptive Learning and Cooperation in Repeated Games*. Discussion Paper No. 9442, Center for Economic Research, Tilburg University (May 1994). Revised, mimeo, Department of Economics, Boston University.

——— (2001). *Reinforcement Learning in Repeated Interaction Games. Advances in Economic Theory*, Vol. 1, No. 1, Article 3.

——— (2000). *Adaptive Parties and Downsian Competition*. Mimeo, Graduate School of Business, Stanford University.

Bendor, J. and P. Swistak (1997). "The Evolutionary Stability of Cooperation". *American Political Science Review*, Vol. 91, 290–307.

Bendor, J., D. Diermeier and M. Ting (2000). *A Behavioural Model of Turnout*. Research Paper No. 1627, Graduate School of Business, Stanford University (March 2000).

Binmore, K. and L. Samuelson (1992). "Evolutionary Stability in Games Played by Finite Automata". *Journal of Economic Theory*, Vol. 57, 278–302.

——— (1997). "Muddling Through: Noisy Equilibrium Selection". *Journal of Economic Theory*, Vol. 74, 235–265.

Börgers, T. and R. Sarin (1997a). "Learning Through Reinforcement and Replicator Dynamics". *Journal of Economic Theory*, Vol. 77, 1–14.

——— (1997b). *Naive Reinforcement Learning With Endogenous Aspirations*. Mimeo, Texas A&M University.

Börgers, T., A. Morales and R. Sarin (1998). *Simple Behaviour Rules Which Lead to Expected Payoff Maxmising Choices*. Mimeo, University College, London.

Bush, R. and F. Mosteller (1955). *Stochastic Models of Learning*. New York: John Wiley and Sons.

Bush, R., F. Mosteller and G. Thompson (1954). "A Formal Structure For Multiple Choice Situations". In R. M. Thrall, C. H. Coombs and R. L. Davis (eds.), *Decision Processes*. New York: Wiley.

Camerer, C. and T. Ho (1999). "Experience-Weighted Attraction Learning in Normal Form Games". *Econometrica*, Vol. 67, No. 4, 827–874.

Cross, J. (1973). "A Stochastic Learning Model of Economic Behaviour". *Quarterly Journal of Economics*, Vol. 87, 239–266.

Cyert, R. and J. March (1963). *A Behavioural Theory of the Firm*. Englewood-Cliffs, NJ: Prentice-Hall.

Dixon, H. D. (2000). "Keeping up with the Joneses: Competition and the Evolution of Collusion." *Journal of Economic Behaviour and Organization*, forthcoming.

Erev, I. and A. Roth (1998). "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria". *American Economic Review*, Vol. 88, 848–881.

Estes, W. (1954). "Individual Behaviour in Uncertain Situations: An Interpretation in Terms of Statistical Association Theory". In R. M. Thrall, C. H. Coombs and R. L. Davis (eds.), *Decision Processes*. New York: Wiley.

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

Ghosh, P. (1998). "Bargaining, Aspirations and Fairness". In *Information and Strategic Interaction in Large Populations*, Ph.D. dissertation, Department of Economics, Boston University, Chapter 5.

Gilboa, I. and D. Schmeidler (1995). "Case-Based Decision Theory". *Quarterly Journal of Economics*, Vol. 110, 605–640.

Karandikar, R., D. Mookherjee, D. Ray and F. Vega-Redondo (1998). "Evolving Aspirations and Cooperation". *Journal of Economic Theory*, Vol. 80, 292–331.

Kim, Y. (1995a). *Satisficing, Cooperation and Coordination*. Mimeo, Department of Economics, Queen Mary and Westfield College, University of London.

——— (1995b). *A Satisficing Model of Learning in Extensive Form Games*. Mimeo, Department of Economics, Yonsei University, Seoul.

Marimon, R. (1997). "Learning From Learning in Economics". In D. Kreps and K. Wallis (eds.), *Advances in Economics and Econometrics*. Cambridge: Cambridge University Press.

Merton, R. and A. Rossi (1950). "Contributions to the Theory of Reference Group Behaviour". In R. Merton and P. Lazarsfeld (eds.), *Continuities in Social Research*. Glencoe, IL: Free Press.

Mookherjee, D. and B. Sopher (1994). "Learning Behaviour in an Experimental Matching Pennies Game". *Games and Economic Behavior*, Vol. 7, 62–91.

——— (1997). "Learning and Decision Costs in Experimental Constant Sum Games". *Games and Economic Behaviour*, Vol. 19, 97–132.

Narendra, K. and P. Mars (1983). "The Use of Learning Algorithms in Telephone Traffic Routing: A Methodology". *Automatica*, Vol. 19, No. 5, 495–502.

Narendra, K. and M. Thathachar (1989). *Learning Automata: An Introduction*. Englewood Cliffs: Prentice Hall.

Nelson, R. and S. Winter (1982). *An Evolutionary Theory of Economic Change*. Cambridge, Massachussetts: Harvard University Press.

Norman, M. F. (1972). *Markov Processes and Learning Models*. New York and London: Academic Press.

Palomino, F. and F. Vega-Redondo (1999). "Convergence of Aspirations and (Partial) Cooperation in the Prisoner's Dilemma". *International Journal of Game Theory*, Vol. 28, No. 4, 465–488.

Papvassilopoulos, G. (1989). "Learning Algorithms for Repeated Bimatrix Nash Games with Incomplete Information". *Journal of Optimisation Theory and Applications*, Vol. 62, No. 3, 467–488.

Pazgal, A. (1995). *Satisficing Leads to Cooperation in Mutual Interest Games*. Mimeo, Kellogg School of Management, Northwestern University.

Roth, A. (1995). "Introduction". In A. Roth and J. Kagel (eds.), *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Roth, A. and I. Erev (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term". *Games and Economic Behaviour*, Vol. 8, 164–212.

Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection.* Cambridge, MA: MIT Press.

Selten, R. (1978). "The Chain Store Paradox". *Theory and Decision*, Vol. 9, 127–159.

——— (1991). "Evolution, Learning and Economic Behaviour". *Games and Economic Behaviour*, Vol. 3, 3–24.

Selten, R. and R. Stoecker (1986). "End Behaviour in Sequences of Finite Prisoners Dilemma Supergames". *Journal of Economic Behaviour and Organization*, Vol. 7, 47–70.

Simon, H. (1955). "A Behavioural Model of Rational Choice". *Quarterly Journal of Economics*, Vol. 69, 99–118.

——— (1957). *Models of Man.* New York.

——— (1959). "Theories of Decision Making in Economics and Behavioural Science". *American Economic Review*, Vol. 49, No. 1, 253–283.

Stahl, D. (1996). "Boundedly Rational Rule Learning in a Guessing Game". *Games and Economic Behaviour*, Vol. 16, 303–330.

Suppes, P. and R. Atkinson (1960). *Markov Learning Models for Multiperson Interactions.* Stanford: Stanford University Press.

Tan, T. and S. R. Werlang (1988). "The Bayesian Foundations of Solution Concepts of Games". *Journal of Economic Theory*, Vol. 45, 370–391.

Young, P. (1998). *Individual Strategy and Social Structure.* Princeton, NJ: Princeton University Press.

Vega-Redondo, F. (1996). *Evolution, Games and Economic Behaviour.* Oxford: Oxford University Press.

Vishwanath, T. (1994). *Adaptive Learning and Search Market Equilibrium.* Mimeo, Department of Economics, Johns Hopkins University.

Weibull, J. (1995). *Evolutionary Game Theory.* Cambridge, MA: MIT Press.