Learning Strategies*

Nobuyuki Hanaki† Rajiv Sethi‡ Ido Erev§ Alexander Peterhansl¶

November 8, 2002

Abstract

Adaptive learning models that have been tested against experimental data typically share two features: (i) initial attractions (or beliefs) are given exogenously, and (ii) learning is based on the performance of stage-game actions rather than repeated game strategies. We develop a model of learning which endogenizes initial attractions and allows for the learning of repeated game strategies. Learning occurs in two phases. In an initial long-run 'pre-experimental' phase, we allow players to explore a complete set of repeated game strategies that satisfy a complexity constraint. The limiting attractions from the first phase are then used as initial attractions in the second, short-run phase, which can be tested against experimental data. We find that, relative to existing adaptive models, we are better able to account for the behavior of subjects in environments where fairness and reciprocity appear to play a significant role.

*We thank Atila Abdulkadiroglu, Alessandra Casella, Peter S. Dodds, Duncan J. Watts, and seminar participants at Columbia University for their comments and suggestions, and Gueorgi Kossinets and Sibel Sirakaya for helpful advice in translating earlier Mathematica code into C.

†Department of Economics, Columbia University (nh85@columbia.edu).

‡Department of Economics, Barnard College, Columbia University (rs328@columbia.edu).

§Faculty of Industrial Engineering and Management, Technion, Israel (erev@tx.technion.ac.il).

¶Department of Economics, Columbia University (ap11@columbia.edu).

1 Introduction

Within the literature on learning in games, two distinct strands may be identified. One deals with the abstract question of the long-run convergence properties of learning models, with particular attention paid to the conditions under which learning leads to Nash equilibrium. The second deals with the more empirical task of describing the manner in which human subjects learn in laboratory interactions. The latter class of learning models can be further subdivided into those that are belief-based, such as fictitious play, and those based on reinforcement. In belief-based models, subjects use observed histories of opponent actions to predict future play, and respond optimally to such beliefs. Reinforcement learning, in contrast, is based on the hypothesis that the propensity to chose an action increases or decreases in response to the payoff experience resulting from the choice of that action. Both belief-based and reinforcement learning models are special cases of experience-weighted attraction (EWA) learning, which allows for the reinforcement not only of actions taken, but also of actions that were *not* taken, based on the imagined payoffs that such actions would have yielded.[1]

Adaptive learning models have been reasonably successful in accounting for observed behavior in certain strategic environments, such as games with unique mixed strategy equilibrium and some coordination games, while failing dramatically to replicate human behavior in others. For example, when experimental subjects are paired to play a Prisoner's Dilemma for a finite number of periods under conditions of full information, convergence to mutual cooperation occurs frequently for many payoff configurations. In contrast, fictitious play predicts convergence to mutual defection for all parameter values. Similarly, in the Battle of the Sexes, fixed subject pairs frequently alternate between the two pure-strategy equilibria of the stage game, thus managing to achieve payoff profiles that are both equitable

[1]Fudenberg and Levine (1998) provide a detailed survey of the theoretical literature. The experimental learning literature is vast; see, for instance, Crawford (1995), Cheung and Friedman (1997), Mookherjee and Sopher (1997), and Erev and Roth (1998). EWA learning was developed by Camerer and Ho (1999), who also show that it generalizes both fictitious play and reinforcement learning. Stahl (1999, 2000) has developed a model allowing for the learning of behavioral rules, defined broadly as mappings from games and histories to probability distributions over actions.

and efficient. Neither reinforcement learning nor fictitious play can account for this, with both models predicting convergence to the repeated play of one or the other pure-strategy stage-game equilibria.[2]

One could conceivably account for the disparity between experimental findings and the predictions of learning models by arguing that subjects care not just about their own monetary payoffs, but also about the payoffs obtained by those with whom they interact. Several recent attempts have been made to identify a richer class of preferences which are able to take such interdependencies into account in a manner consistent with experimental behavior.[3] From this perspective, payoff functions must be appropriately transformed before learning models can properly be tested or compared. While this is an important and promising direction for research, there is as yet no consensus on the precise manner in which monetary payoffs should be transformed in order to conform to 'social preferences'. Moreover, as we argue below, behavior that appears to be motivated by a concern for fairness and efficiency can in fact be the consequence of an entirely orthodox process of learning in which material payoffs are the driving force.

In this paper we maintain the hypothesis that subjects are motivated primarily by a concern with their own monetary payoffs, but allow for the possibility that subjects learn not just among actions but among repeated game strategies. This general approach has been suggested by Erev and Roth (1999), on the basis of an extensive review of the evidence. Camerer and Ho (1999, p.871) have also noted that "stage game strategies are not always the most natural candidates for the strategies that players learn about," and McKelvey and Palfrey (2001, p.19) have argued for the development of "strategic learning" models in which players learn not about the performance of actions but rather of strategies.[4] There are two potential problems with this approach. First, the size of the strategy space precludes experimentation

[2]These and other failures of existing adaptive learning models are discussed further in Section 2 below. McKelvey and Palfrey (2001) identify additional weaknesses of standard learning models, such as their insensitivity to variations in information and matching conditions.

[3]See, in particular, Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Levine (1998), and Charness and Rabin (2002). Such preference interdependence is predicted by several evolutionary models, including Güth and Yaari (1992), Huck and Oechssler (1999), Gintis (2000), and Sethi and Somanathan (2001).

[4]See also Erev and Roth (2001) and Stahl and Haruvy (2002).

with all but a few strategies in any given interaction. In fact, any learning rule can itself be interpreted as a single repeated game strategy. This problem can be overcome, as McKelvey and Palfrey point out, by restricting the complexity of repeated game strategies. The second difficulty arises from the fact that if players are learning among repeated game strategies, it becomes impossible to compute the hypothetical payoffs that would have been obtained had a *different* strategy been chosen. Hence, even with observable actions and stage-game payoff functions, neither fictitious play nor the general version of experience-weighted attraction learning can be implemented. As McKelvey and Palfrey observe, "players face an inference problem going from histories to beliefs" about the strategies of their opponents.

When learning responds only to payoffs obtained by strategies actually chosen by the subject, *this inference problem does not arise.* A much maligned attribute of reinforcement learning, therefore, turns out to be an advantage in developing models of learning among repeated game strategies. In a straightforward extension of their earlier models of reinforcement learning, Erev and Roth (2001) have studied the Prisoner's Dilemma while allowing for players to choose among the two stage-game actions as well as the "tit-for-tat" repeated game strategy. Allowing for the possibility that subjects can learn to reciprocate significantly improves the predictive power of the model. It does so at a cost, however. Erev and Roth assume, in effect, that "tit-for-tat" is the only repeated game strategy to have positive probability weight when the process of learning begins. This choice is fundamentally arbitrary, and raises the question of why other strategies cannot also have positive initial weight. More generally, one would like a theory of initial attractions that identifies the set of repeated game strategies which experimental subjects explore.

Developing such a theory is the principal aim of this paper. Our contribution can be thought of as a reinforcement based approach to learning over long horizons in a 'pre-experimental' phase, that determines which repeated game strategies are salient when subjects enter the laboratory. It is in this sense a theory of the initial attractions that appear as parameters in standard learning models. In our model learning occurs over a long horizon, and begins with positive weight on each repeated game strategy that satisfies a bounded complexity constraint. Specifically, we consider all strategies that can be represented by

automata having no more than two states.[5]

The model may be described briefly as follows. A large, fixed population is divided into subject pairs. There are two phases of learning. During the first, 'pre-experimental' phase, subjects engage in a lengthy process of learning among repeated game strategies while being occasionally rematched with other members of the population. There is a finite set of simple repeated game strategies from which subjects choose. At the start of the first phase of learning, each of the repeated game strategies has equal attraction, and hence equal probability of being chosen. Attractions are updated over time as the payoffs resulting from strategy choices are observed. Subjects maintain their chosen strategies for several repetitions of the stage game, with the length of this period determined stochastically. Specifically, at each stage, there is some small and constant probability that attractions will be updated and strategy revision will occur. Only strategies which are actually chosen are updated, based on their observed payoff consequences. If strategy revision occurs, the (possibly) new strategy is chosen on the basis of updated attractions. There is also a small probability that at any stage, subjects pairs are dispersed and individuals are re-matched with other subjects drawn from the population. Over the course of this process some strategies decline in use while others are observed with greater frequency. The process continues until convergence to a limiting distribution is approximated, and this ends the first phase of learning. The limiting attractions from the first phase are then used as initial attractions in the second, which consists of a fixed-pair matching for a small number of periods. Learning also occurs in the second phase, but without rematching. This corresponds to the conditions of an experiment, and enables us to compare our results with reported experimental data. We find that several patterns of behavior which are difficult to reconcile with action-learning models, such as cooperation in the Prisoner's Dilemma and alternation between pure-strategy equilibria in the Battle of the Sexes, emerge as outcomes of our learning procedure.

[5]This seems unrestrictive in an analysis of 2×2 games and, as we show below, promising results can be obtained without considering a larger strategy space. In an evolutionary model of the repeated Prisoners' Dilemma, Miller (1996) considered automata having up to 16 states, and found an endogenous decline in complexity, with the survival of strategies similar to 'tit-for-tat'.

2 Two Examples

In Arifovic, McKelvey and Pevnitskaya (2002, henceforth AMP), a number of well-known learning models are presented side-by-side with the experimental results from an earlier paper, McKelvey and Palfrey (2001). In one of the treatments reported, the setting was a fixed matching of two players playing a repeated game. The human experiments consisted of 48 subjects paired up for 24 matches. Each match consisted of paired subjects playing a game for 50 rounds. The subjects in each pairing both saw the complete payoff matrix, and observed their opponent's choice of action after each round. Each match produced one binary-tuple of data: the average payoffs of each of two subjects over the course of the match. The resulting 24 data points were plotted in payoff space, with the row-player's payoffs on the horizontal axis and the column-player's payoffs on the vertical axis. A variety of learning models, including fictitious play, reinforcement learning, and EWA learning, were then simulated under literally the same 'experimental' conditions, using the initial conditions and parameter values obtained in prior studies. Data in this form was produced for eight different games, including the Prisoner's Dilemma, Chicken, Battle of the Sexes, 2×2 and 3×3 Stag Hunt games, and strategic form versions of Ultimatum bargaining and the Centipede games. There were systematic deviations between the simulated and the experimental results. The differences were, in fact, so large that the authors have called for new learning models firmly rooted in the experimental evidence and for new methodologies for evaluating them.

To get a feel for the differences in the outcomes of the learning models versus the outcomes of the experiments, consider their findings for the Prisoner's Dilemma (Figure 1) and the Battle of the Sexes (Figure 2).

	A	B
A	$8, 8$	$1, 9$
B	$9, 1$	$2, 2$

Figure 1: Prisoner's Dilemma

Although AMP report only the average payoff profile obtained by each subject pair, it is

possible to make some clear inferences about the path of actions chosen. In the Prisoner's Dilemma, over half of the human data are tightly grouped around the payoffs of $(8, 8)$, implying that the subjects have coordinated on the non-equilibrium action profile (A, A). The rest of the data points imply a mix of actions, such as exploiting a cooperator by defecting, as well as being exploited when cooperating. Fictitious play immediately converges to the unique Nash equilibrium with payoffs of $(2, 2)$. Most reinforcement learning data points are scattered in an area relatively close to the Nash equilibrium. Although better results might be achieved with a recalibration of the models (as indeed we show below) there are clear qualitative differences between the predictions of the learning models and the behavior of most human subjects.

	A	B
A	$18, 6$	$3, 3$
B	$3, 3$	$6, 18$

Figure 2: Battle of the Sexes

In the Battle of the Sexes, the majority of the experimental data points are closely scattered around the payoff profile $(12, 12)$, implying that players coordinated by alternating between the two pure-strategy stage-game Nash equilibria. Fictitious play converges to one of the two pure-strategy equilibria, while the data points for reinforcement learning are scattered between the three equilibria. Similarly, in the 2×2 Stag Hunt and in the game of Chicken, the majority of experimental subjects coordinate their strategies so as to maximize joint payoffs while preserving an equitable distribution, a pattern of behavior that is difficult for action-learning models to consistently replicate.

3 Learning Among Repeated Game Strategies

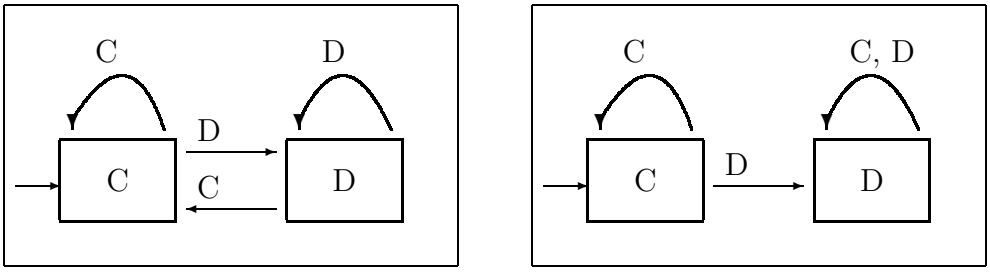
Any analysis of learning among repeated game strategies requires some restriction on the space of available strategies. This is achieved here by restricting the complexity of the strategies available to players. One way of assessing the complexity of a repeated game

strategy is on the basis of its representation as a finite automaton. The larger the number of states a strategy requires in automaton representation, the greater its complexity.[6] This section starts with a brief description of the manner in which a repeated game strategy can be represented as a finite automaton. We then proceed to discuss the learning model in some detail.

3.1 Representing Repeated Game Strategies with Automata

An automaton is described by four components: a *set of states*, an *initial state* that the automaton occupies at the outset, an *output function* that indicates which action is to be taken in each particular state, and a *transition function* that indicates which state will be reached in the next period given the current state and the current actions of the opponent. The current state of an automaton contains all information about the history of play that is relevant for the execution of the corresponding strategy.

The 'tit-for-tat' strategy in the Prisoner's Dilemma can be represented as a two-state automaton (Figure 3). The two states in this case are associated with the two available actions, cooperation and defection. The set of arrows are associated with the opponent's actions, and represent the transition function. The initial state is cooperation, and the automaton stays in (or returns to) this state each time its opponent cooperates. It enters (or remains in) the defection state each time its opponent defects.



tit for tat grim trigger

Figure 3: Two examples of two-state automata

[6]This approach to strategic complexity is also utilized, for instance, by Binmore and Samuelson (1992). Chapter 8 of Osborne and Rubinstein (1994) provides a good general introduction to the automaton representation of a repeated game strategy.

Another strategy that can be represented as a two-state automaton is 'grim trigger', also shown in Figure 3. Again, the two states are associated with the two available actions, cooperation and defection. The initial state is again cooperation, but the transition function is different. A defection by the opponent triggers a move to the defection state, which is absorbing (the automation never leaves this state).

Each of the above strategies require a memory of at most one action by the opponent. A slightly more complex strategy is 'tit-for-two-tats', illustrated in Figure 4. This strategy starts with cooperation and defects only if the opponent defects twice in a row. In order to implement this strategy, a three-state automaton is required. In the figure, the i-th state is denoted by S_i and the action to be taken in each state follows the colon after the state. The initial state is S_0 in which the player cooperates. If the opponent defects, state S_1 is reached. The player still cooperates in this state but remembers that the last action by the opponent was defection. Cooperation by the opponent when the player is in this state leads to a return to S_0. On the other hand, if the opponent defects again, S_2 is reached and the player defects. Cooperation by the opponent when the player is in S_2 induces a return to S_0.

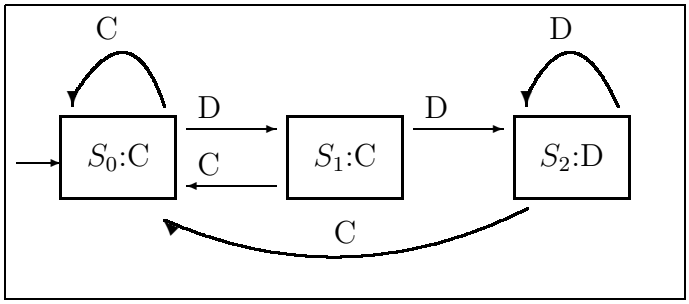


Figure 4: tit-for-two-tats

In this paper, we consider a total of 26 possible repeated game strategies. When the stage game is 2×2, this set corresponds to all unique strategies representable by one- or two-state automata.[7]

[7]There are two possible states for an automaton and two possible actions by the opponent. The transition function thus maps four different possibilities of (own state, the opponent's action) pairs into the set of two

3.2 Reinforcement Learning Among Strategies

Consider a population of players $\mathcal{N} = \{1, 2, 3, ..., N\}$ and a specified symmetric 2×2 stage game. During the first phase of learning, players drawn from the population are matched pairwise to play the stage game repeatedly. Players use the same strategy across several periods, but occasionally switch strategies as part of a process of experimentation. They are also randomly rematched with other partners from time to time. Let $\rho \in (0, 1)$ represent the probability that a player switches to a (possibly) new strategy at the start of any given period, and $\mu \in (0, \rho)$ the probability that a player is randomly re-matched against a (possibly) different opponent at the start of any given period.

Players have propensities or 'attractions' associated with each of their strategies, and these attractions determine the probabilities with which strategies are chosen when players experiment. At the start of the first phase, all strategies have equal attraction and hence equal probability of being chosen. Learning takes place through the evolution of attractions: prior to updating her strategy, a player evaluates the performance of the strategy she has been utilizing and updates her attractions accordingly (the precise manner in which this occurs is described below). Since the learning process is defined over strategies, players are required to play the stage game a number of times to evaluate their current strategy, that is, to obtain information on their strategy's payoff consequences. If ρ is not too large, meaningful evaluations of repeated game strategies are possible. Notice that unlike the action learning, players need not update their strategies simultaneously. When players are re-matched, they also update their strategies.[8] This process continues until the limiting distribution of attractions is approximated, at which point the second 'experimental' phase begins. This consists of a fixed number of periods without further re-matching. Learning occurs also in this phase, building on the attractions generated during the pre-experimental

states to be taken in the next period. This generates $2^4 = 16$ cases in total. As each of these can have one of two initial states, we have a total 32 possible automata. Among these, however, four always play the first action and another four always play the second. Elimination of non-unique automata yields a total of 26, of which two are one-state and the rest two-state automata. Appendix B contains a complete listing of these.

[8] When a player is re-matched, she does not know the previous action played by the new opponent. Rather than assume that the old strategy is retained but enters its initial state, we assume that a strategy revision occurs.

phase.

Let $A_s^i(t)$ denote player i's attraction to the strategy $s \in S$ at period t, where $S = \{1, 2, 3, \ldots, 26\}$ is player i's set of 26 strategies. For each player, attractions are updated when the player updates her strategy. Only the strategy that was chosen at the previous strategy update is reinforced, as follows. Consider a player who updates her strategy choice at the start of period t, and uses the same strategy $s \in S$ without further updates until the start of period $t + \tau$. Specifically, suppose that $s^i(t) = s^i(t+1) = ... = s^i(t+\tau-1)$, where $s^i(r)$ is the strategy used by player i in period r. Define the reinforcement value $R^i(t, t+\tau-1)$ of the strategy used over the periods $t, ..., t+\tau-1$ as the average payoff obtained by player i over this period:

$$R^i(t, t+\tau-1) = \frac{1}{\tau} \sum_{r=t}^{t+\tau-1} \pi^i(r),$$

where $\pi^i(r)$ is the payoff obtained by player i in period r. When strategy revision next occurs (at the start of period $t+\tau$), player i's attraction or propensity for playing strategy s evolves as a weighted average of its previous value and the reinforcement value:

$$A_s^i(t+\tau) = \begin{cases} (1-\omega) A_s^i(t) + \omega R^i(t, t+\tau-1) & \text{if } s = s^i(t) = ... = s^i(t+\tau-1), \\ A_s(t) & \text{otherwise.} \end{cases} \tag{1}$$

Here $\omega \in (0, 1)$ is a weight placed on the reinforcement value, $R(\cdot, \cdot)$, which is the average payoff the player has obtained from using strategy s since the last strategy update, i.e., between period t and $t+\tau-1$.[9]

The probability of a player i choosing strategy s, when she updates her strategy in the beginning of period t, depends on the attractions as follows:

$$p_s^i(t) = \frac{e^{\lambda A_s^i(t)}}{\sum_{k \in S} e^{\lambda A_k^i(t)}} \tag{2}$$

The parameter $\lambda \geq 0$ in the logistic transformation represents the extent to which strategies with higher attractions are favored in strategy choice. When $\lambda = 0$, all strategies are equally likely to be chosen, regardless of their attractions. As λ increases, strategies with higher

[9]Note that the attraction of each strategy approaches its historical average payoff as the number of updates becomes large.

attractions become disproportionately more likely to be chosen. In the limiting case $\lambda \to \infty$, the strategy with the highest attraction is chosen with probability one.

In the long horizon 'pre-experimental' phase of learning, the initial attraction, $A_s(0)$, for all strategies is set equal to the expected payoff given random choice of actions by both players. As an example, consider the Prisoner's Dilemma game described in Figure 2. Here a player's initial attraction for each of her 26 strategies is given by $A_1(0) = \ldots = A_{26}(0) = \frac{1}{4}(8+1+9+2) = 5$. In the second 'experimental' phase of learning, we assume that players bring with them to the laboratory the values of $A_s(\cdot)$ that they have reached at the conclusion of the first phase.

There are total of four parameters in this model: the strategy updating rate ρ, the rematching probability μ, the weight ω on reinforcement values in attraction updates, and the sensitivity λ of the strategy choice to the attraction level in the logistic transformation. The number of players N needs to be large to ensure multiple interactions among various players in the pre-experimental phase.

4 Results

We have limited our attention to the four symmetric 2×2 games for which results are reported in the AMP, namely the 2×2 Stag Hunt, Prisoner's Dilemma, Chicken, and the Battle of the Sexes.[10] In the first phase of learning, attractions to strategies are 'initialized' in anticipation of the second, experimental phase. The final values of $A_s^i(\cdot)$ from the pre-experimental phase are the initial values $A_s^i(0)$ for the experimental phase. This endogenizes the initial attractions or initial probability weights that appear as parameters in standard learning models.

The pre-experimental phase continues until the limiting distribution of attractions or probability weights placed on strategies are approximated. Let $\overline{p_s}(\cdot)$ be the population

[10] In order to make the game symmetric, the actions for the Column players in the Battle of the Sexes game have been relabeled as shown in figures 8 and 9 below.

average probability weight on strategy s in a given period, and let

$$\overline{\overline{p_s}}(m) = \frac{1}{1000} \sum_{t=1000(m-1)+1}^{1000m} \overline{p_s}(t)$$

be the mean of the population average probability weight on strategy s over the m-th block of 1000 periods. The convergence criterion employed in the simulation was

$$\frac{1}{|S|} \sum_{s \in S} \mid \overline{\overline{p_s}}(m) - \overline{\overline{p_s}}(m-1) \mid \; < \varepsilon$$

for 20 consecutive m's. That is, the pre-experimental phase is terminated if the absolute difference between two consecutive means of the population average probability weights are, on average, less than ε for a long time.[11] In the experimental phase, players are randomly paired to play and learn over the course of one match with 50 periods (without re-matching). This corresponds to McKelvey and Palfrey's experimental conditions, as reported in AMP.[12]

We begin by describing results of the pre-experimental phase, i.e., the limiting distributions of probability weights across the 26 strategies, with a focus on strategies that obtain high limiting weights, for a particular set of parameter values. We then proceed to discuss the results in the experimental phase. The set of parameter values are as follows: the strategy update rate $\rho = 0.05$, the rematching probability $\mu = 0.02$, weights on the reinforcement values in attraction updates $\omega = 0.1$, and the sensitivity of the strategy choice to the attraction level $\lambda = 4$. Among the variety of parameter configurations with which we have experimented, this set of values provides the highest average performance for the four games considered here.[13] Our focus is on qualitative performance, namely the ability of the model to replicate the broad contours of the experimental data with respect to the attainment of

[11]The maximum length of the pre-experimental phase in each of the simulation runs has been set to 500,000 periods. In principle, it is possible to have a simulation run that does not satisfy the convergence criterion before the final period if ε is very small. However, we obtained convergence in all cases, using a value of $\varepsilon = 0.005$.

[12]Simulations were coded and run using Borland C++. The source code is available from the authors upon request.

[13]We have experimented with all possible combinations of the following set of parameter values: $\rho = \{0.2, 0.1, 0.05\}$, $\omega = \{0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.5, 0.7\}$, $\mu = \{0.02, 0.01, 0.005\}$, and $\lambda = \{2.5, 3.0, 3.5, 4.0\}$. For all simulations, the population size is kept constant at 1000.

fair and efficient outcomes. Sensitivity of results to changes in parameter values are discussed in Appendix C.

4.1 Pre-experimental Phase

What are the strategies that simulated players bring with them to the laboratory? In this section, we discuss the limiting distributions of probability weights in the pre-experimental phase to answer this question.

Figures 5 to 8 show the approximate limiting distributions of probability weights for the 26 strategies in each of the four games. (Appendix B contains the complete set of strategies in automata representation. The strategy indices referred to in the figures as well as in the text of this section correspond to those in this appendix.) To simplify the discussion, we focus on strategies with high limiting probability weight. A strategy is said to have a high limiting probability weight if its weight is at least one standard deviation above the average probability weight across all strategies. These strategies are, in a sense, the principal strategies that players "bring to the laboratory" for the experimental phase.

Results for the 2×2 Stag Hunt are shown in Figure 5. The strategies with high limiting probability weight are 'always play B', 'grim trigger', 'tit-for-tat', 'punish until the opponent retaliates', and 'punish once', respectively. The initial state of each of these strategies is B. The first three do not require further explanation, since they have already been discussed in Section 3.1 above. Strategy 22 'punishes until the opponent retaliates': it starts by playing B and stays in this state as long as the opponent also plays B. Once the opponent plays A, however, it switches to playing A. It returns to B only if the opponent plays A, otherwise it stays in state A. Strategy 25, which 'punishes once', also starts with action B. It stays in state B unless the opponent plays A. Once the opponent plays A, state A is reached for exactly one period, after which the strategy returns to B regardless of the opponent's action. Notice that if these five strategies are matched against each other, we will observe all players playing action B forever to achieve the efficient and fair outcome.

Figure 6 shows the outcome of the pre-experimental phase for the Prisoner's Dilemma. Strategies obtaining high limiting probability weights are 'punish once', 'tit-for-tat', 'punish until the opponent retaliates', and 'grim trigger', respectively. The initial state is A for these

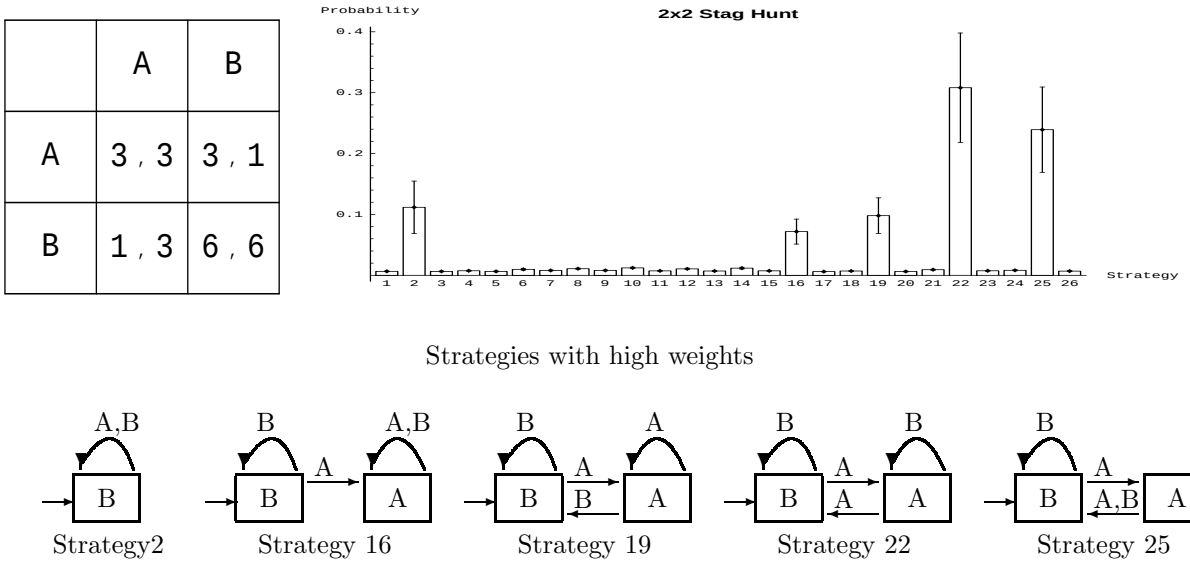


Figure 5: Approximated limiting distribution of probability weights across 26 strategies and strategies with high probability weights for 2×2 Stag Hunt game. Strategy indices correspond to those in the Appendix B. Probability weights are averaged over 100 realizations. Error bars in the histogram represent two standard deviations around the mean. Parameter values are $\rho = 0.05$, $\mu = 0.02$, $\omega = 0.1$, and $\lambda = 4.0$

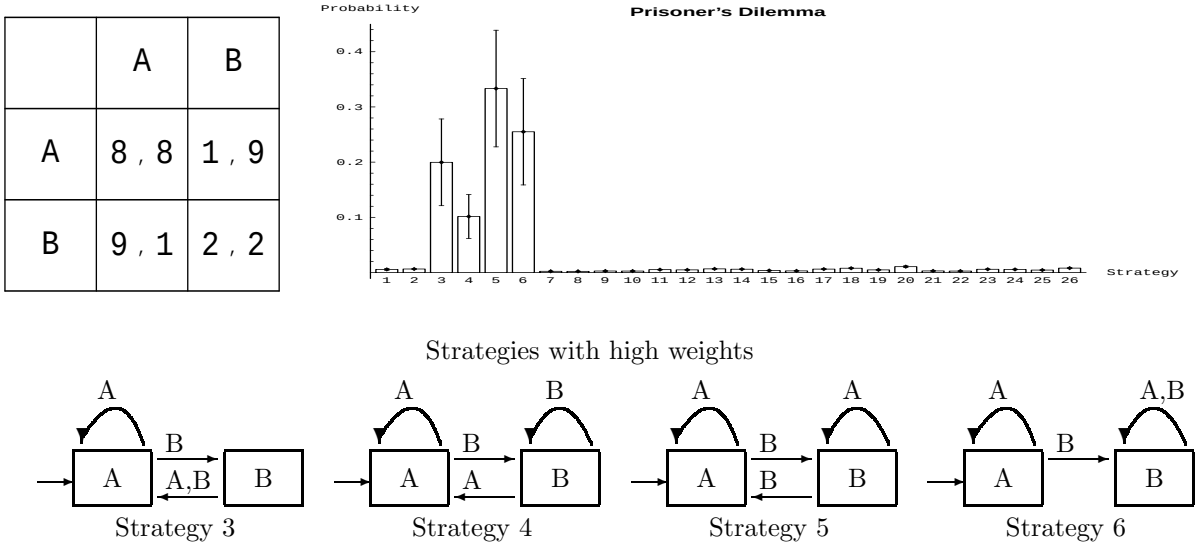


Figure 6: Approximated limiting distribution of probability weights across 26 strategies and strategies with high probability weights for Prisoner's Dilemma.

strategies. As in the case of the 2×2 Stag Hunt game, if these four strategies are played amongst themselves, the observed history of actions will involve mutual cooperation in all periods.

Since the Prisoner's Dilemma has received such widespread attention in economics, the strategies the emerge from the learning process in our model deserve further discussion. The 'tit-for-tat' strategy was the winner in two tournaments organized by Axelrod,[14] and has been a subject of extensive study, especially in the context of evolutionary game theory. Axelrod and Hamilton (1981) have shown that 'tit-for-tat' is a neutrally stable strategy in the infinitely repeated prisoners' dilemma with payoffs evaluated according to the limit-of-the-means criterion. This has been interpreted as providing theoretical support for the hypothesis that cooperation sustained through reciprocation is an inevitable outcome of evolutionary process. However, there are a large number of other repeated game strategies that are also neutrally stable, and some of them involve mutual defection in most periods. The set of stable strategies can be refined substantially by introducing complexity costs (as in Binmore and Samuelson, 1992) or the possibility of errors in the implementation of strategies (as in Fundenberg and Maskin, 1990). These refinements result in a prediction of mutual cooperation in the infinitely repeated prisoners' dilemma, although on the basis of strategies other than 'tit-for-tat'. We also find mutual cooperation to be the predicted outcome, although the model considered here is one of finite repetition and bounded complexity. The 'tit-for-tat' strategy survives but does not have the highest limiting probability weight: we find 'punish until the opponent retaliates' to be the most prolific strategy. The key difference between 'tit-for-tat' and 'punish until the opponent retaliates' is that the latter is able to exploit unconditional cooperation.

Results from the game of Chicken are summarized in Figure 7. In this game, the strategies 'punish once', 'tit-for-tat', and 'punish until the opponent retaliates', are the ones with the high probability weights. Again, if these strategies are played only among themselves, we will observe only the efficient and fair outcome as both players continue to play action A. Unlike the two games discussed above, however, there are a few other strategies with non-negligible probability weights. The presence of these strategies will generate outcomes that

[14]See Axelrod and Hamilton (1981) for a brief summary of these tournaments.

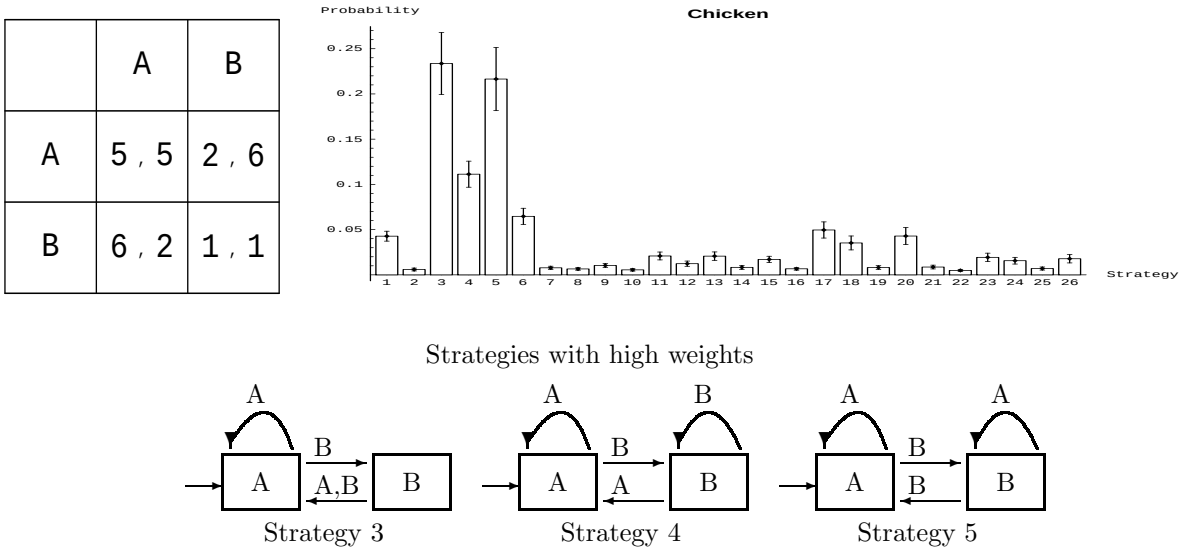


Figure 7: Approximated limiting distribution of probability weights across 26 strategies and strategies with high probability weights for Chicken.

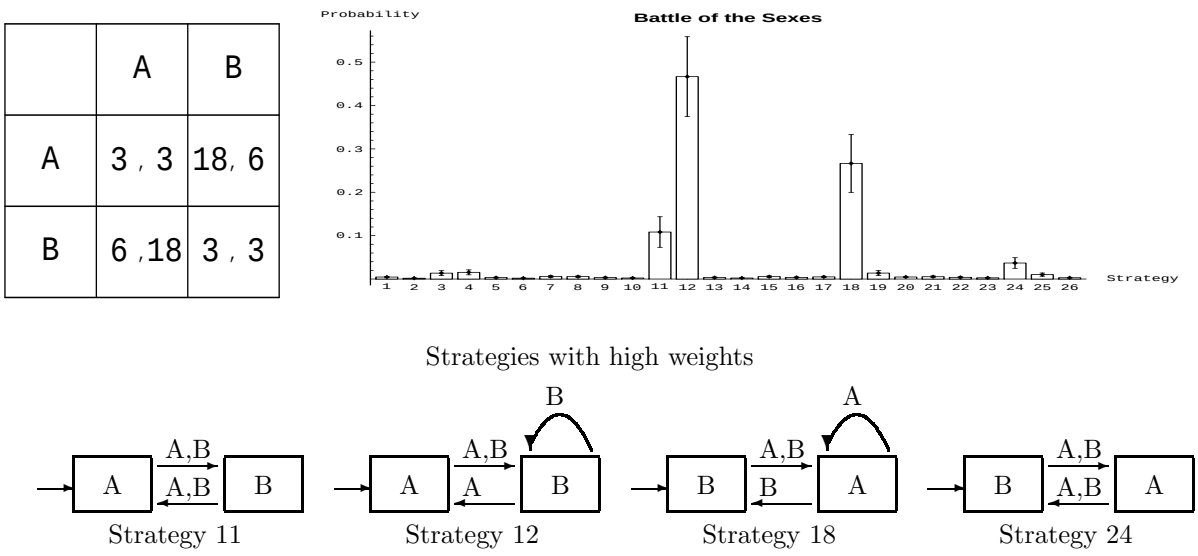


Figure 8: Approximated limiting distribution of probability weights across 26 strategies and strategies with high probability weights for Battle of the Sexes.

are not efficient in the experimental phase as we show below.

Figure 8 shows strategies with high weights in the Battle of the Sexes. Strategies 11 and 24 mechanically alternate between two states, but have different initial states. Strategy 12 starts by playing action A and, regardless of the opponent's action, switches to state B. It stays in this state if the opponent plays B, otherwise it returns to state A. Strategy 18 is qualitatively identical to Strategy 12, with the role of the two actions reversed. Note that strategies 12 and 18, when matched with any of the other strategies having significant weights, eventually achieve perfect alternation between the two pure strategy stage-game equilibria. Since these two 'flexible' strategies have the highest weights, players in the experimental stage eventually learn to use one of them when initial actions are not coordinated on the equitable and efficient outcome.

Just by examining the strategies with high limiting probability weights from the first phase of learning, we can expect that efficient and fair outcomes will be observed in the experimental phase. We now turn to discussion of our results for this phase.

4.2 Experimental Phase

Figure 9 shows the results of simulation runs for the four games we have considered in the experimental phase. Also presented in the figure are results from two learning models — smoothed fictitious play and action-reinforcement learning. (See Appendix A for the algorithm used for generating the data for these models.)

For three of the four games – 2×2 Stag Hunt, Prisoner's Dilemma, and Chicken – the efficient outcome is also fair. A majority of the simulated players who learn among repeated game strategies are successful in obtaining such an outcome.[15] They achieve this by repeatedly playing the action profiles {B,B} in the 2×2 Stag Hunt and {A,A} in the Prisoner's Dilemma and Chicken. In contrast, fictitious play generates a stage-game Nash equilibrium outcome as the theory predicts. Fictitious play thus generates an outcome that corresponds to the experimental data only in the coordination game (2×2 Stag Hunt), in which the efficient outcome is one of the two pure-strategy stage-game Nash equilibria.

[15]The results for Chicken are somewhat weaker than those for the other games.

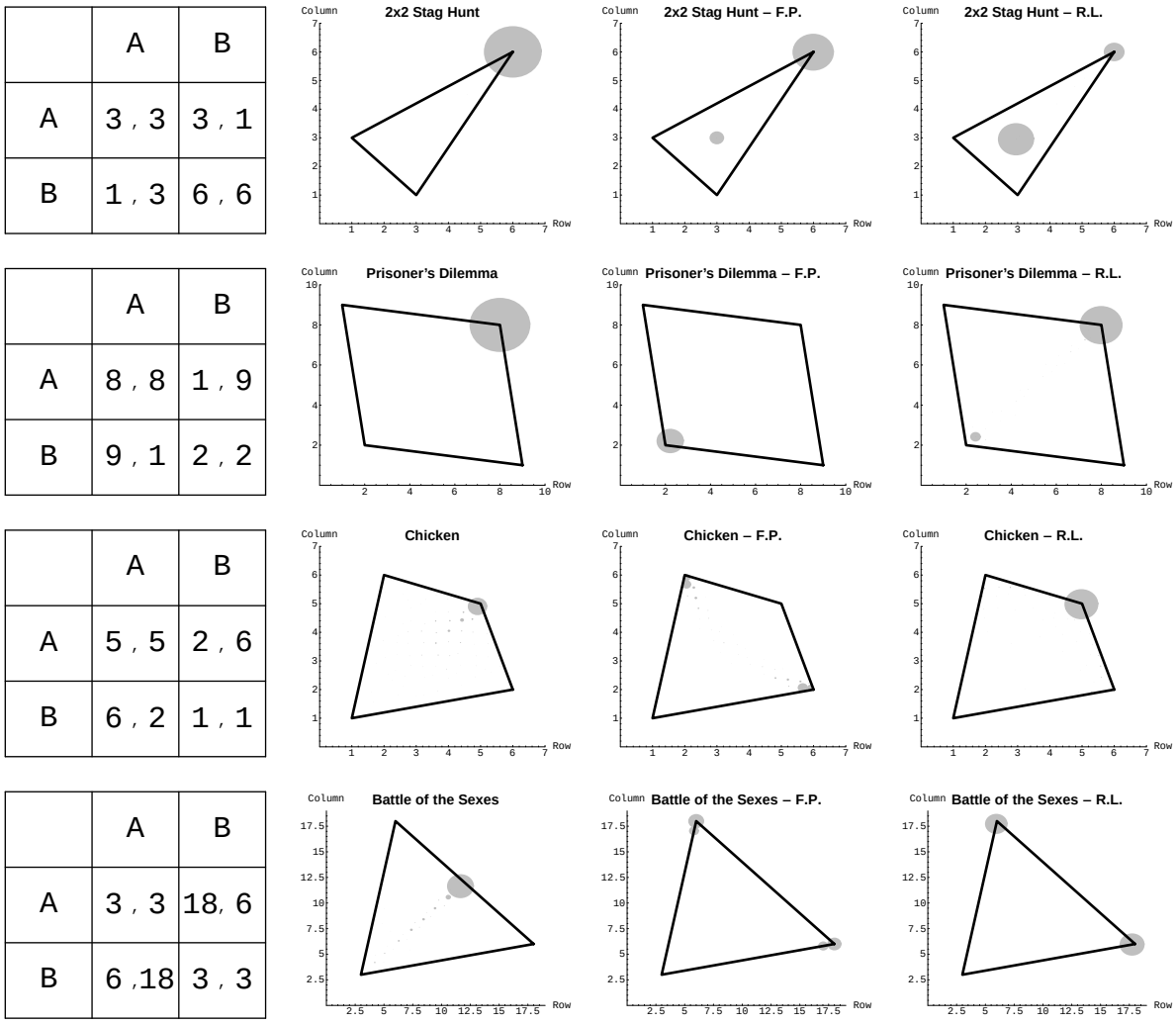


Figure 9: Comparison of the simulation outcomes among three models of learning: reinforcement learning among repeated game strategies (left) , smoothed fictitious play among actions (center), and reinforcement learning among actions (right). $\lambda = 4.0$ for all the models. Other parameters for the first model are set as ρ, μ, ω=$\{0.05, 0.02, 0.1\}$. The polygons in each plots represent the possible per stage average payoff space. Payoff for row (column) players are in the horizontal (vertical) axes. Each point corresponds to payoff profile for a pair, and the gray circles around the points represent the relative likelihood of observing that particular outcome. Each figure is based on a total of 1000 data points. (The figures for the first model is generated by running two simulations with 1000 players. Each simulation generates 500 data points.) Note that the size of circles are only comparable across the three models for the same game and not across the games for the same model.

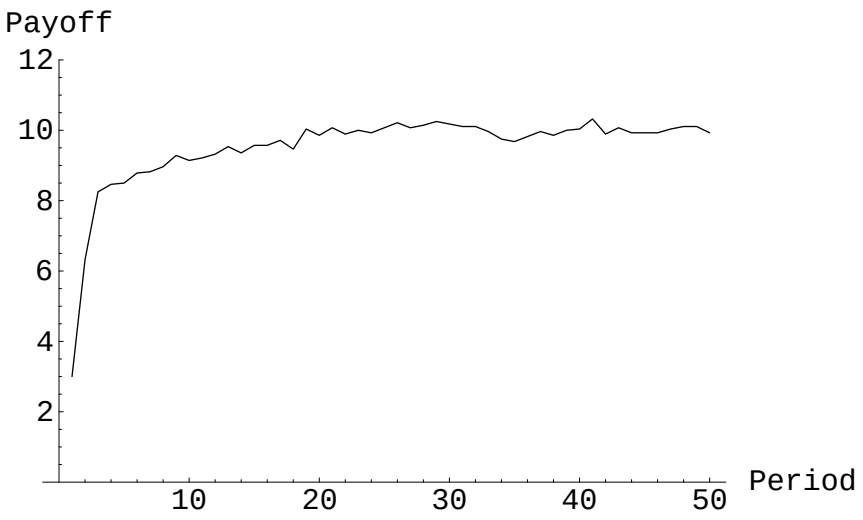


Figure 10: Within experimental phase dynamics of the average payoffs for the players who did not obtain efficient outcomes in the initial periods.

Also note that the Pareto superior equilibrium is more likely to be observed than the Pareto inferior one under fictitious play. Reinforcement learning among actions generates an efficient outcome that is not the stage-game Nash equilibrium in Prisoner's Dilemma and Chicken, but it is more likely to result in the Pareto inferior equilibrium in the coordination game. This is an interesting contrast between the two models and deserves further investigation.

In the Battle of the Sexes game, the efficient and fair outcome requires coordinated alternation between two pure-strategy stage-game Nash equilibria. The results for this game are particularly striking. The simulation outcome shows that many of the players successfully learned, out of 26 possible repeated game strategies, to play the strategies that enable them to achieve the efficient and fair outcome. As one can clearly see in the figure, neither of the two action-learning models generates such an outcome.

In the strategy-learning model presented in this paper, there are players who did not obtain the fully efficient outcome.[16] This is a result of an initial mismatch between the strategies chosen by the players. When there is such a mismatch, however, players successfully learn to switch to other strategies. Figure 10 plots the experimental phase dynamics of

[16]The center of the largest gray circle is located inside the triangle. This means that per period average payoff was a slightly below the efficient one.

average payoffs for those players who did not obtain the efficient outcome initially. The figure shows that average payoffs increase steadily over time. The initial rapid increase is primarily due to the use of strategies that may be initially mismatched but which nevertheless achieve efficient alternation within three periods. This would be the case, for instance, if Strategy 12 were matched against Strategy 11 (see Figure 8). The rise in payoffs in subsequent periods results from the process of learning. For instance, if players both initially adopt Strategy 11 they will both be perpetually mismatched, eventually inducing one of them to experiment with one of the other strategies. After the switch, convergence to efficient alternation occurs within three periods. The probabilistic nature of strategy choice, however, can cause the players to mismatch even after several periods of successful alternation. This results in a fluctuation of payoffs in later periods. The battle of the sexes is a striking example of a game in which our approach predicts outcomes that are both consistent with experimental observation and virtually impossible to replicate with action-learning models.

5 Conclusion

We have demonstrated that a simple reinforcement model of learning applied to a restricted set of repeated game strategies can account for the behavior of human subjects in environments where fairness and reciprocity seem to play a significant role. We have done so without assuming that fairness and reciprocity are primitive concerns. Our results may also be of some interest from the perspective of the problem of equilibrium selection in games. In pure coordination games, where fairness and efficiency are not in conflict, our findings predict that learning will converge to the efficient action profile. In the Battle of the Sexes, where efficient stage-game equilibria are unfair, the model predicts alternation over time to achieve a profile of average payoffs that is both efficient and fair. In the Prisoner's Dilemma, where fairness and efficiency are not in conflict but cannot be attained in equilibrium, the model predicts convergence to nonequilibrium strategy profiles.

One important direction for further research would be to study the feasibility of our approach in settings of greater complexity, with a larger set of players and stage game actions. A potential empirical extension is an analysis of the goodness-of-fit of the model

to the large and varied experimental data that is available. This would require estimation of the model parameters ρ, μ, ω, and λ and out-of-sample comparisons with other learning models.

Finally, it would be well worth developing a deeper analytical understanding of the process by which learning on the basis of material payoffs can result in behavior that appears to be motivated by fairness and efficiency concerns. A characterization of the class of games for which the learning dynamics converge to fair and efficient payoff profiles would be of considerable interest.

A Fictitious Play and Action-Reinforcement Learning

We provide here a brief discussion of two standard learning models. As shown in Camerer and Ho (1999), both fictitious play and reinforcement learning model can be considered as special cases of the experience weighted attraction (EWA) learning model. The following formulation is a simplified version of the EWA model.

Let $A_a^i(t)$ be player i's attractions to the action $a \in S$ at period t, where S is player i's action set. For each player, attractions evolve over time as weighted averages of their previous values and current reinforcement values. Let $a_{-i}(t)$ be the actions chosen by a player's opponents, denoted by $-i$, at period t. The player's attraction to action a evolves as follows:

$$A_a^i(t+1) = (1 - \omega_a^i(t+1))A_a^i(t) + \omega_a^i(t+1)\ \pi^i(a, a_{-i}(t)) \tag{3}$$

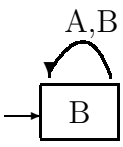
It is easy to verify that fictitious play is equivalent to having $\omega_a^i(t+1) = 1/(t+1)$ for all a. A reinforcement learning model can be obtained by setting

$$\omega_a^i(t+1) = \begin{cases} \frac{1}{n_a(t+1)} & \text{if action } a \text{ is chosen in period } t \\ 0 & \text{otherwise} \end{cases}$$

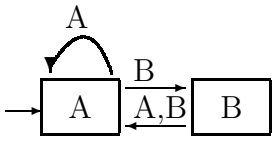
where $n_a(\cdot)$ is the total number of times the action a has been chosen since the beginning of play plus its initial value $n_a(0)$. In the simulation in section 4.2, we assume no 'pre-experimental' learning for these models, as in the previous literature. We set $n_a(0)$ equal to 1, and the initial attraction for all actions, $A_a^i(0)$ is set to be the expected payoff given a random choice by both players. The probability with which each action is chosen is given by equation 2 in Section 3.2.

B The Complete Set of Strategies Considered

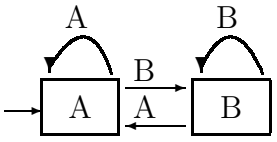
Strategy1



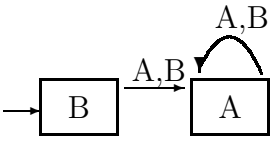
Strategy2



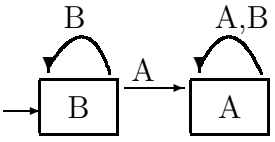
Strategy 3



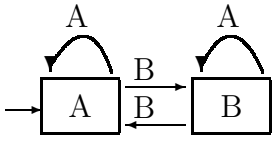
Strategy 4



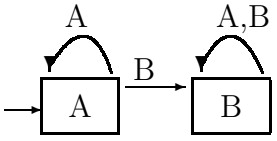
Strategy 15



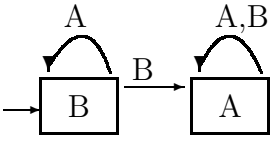
Strategy 16



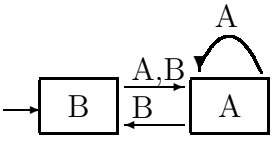
Strategy 5



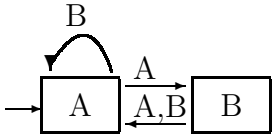
Strategy 6



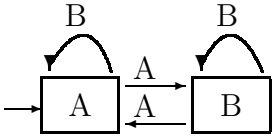
Strategy 17



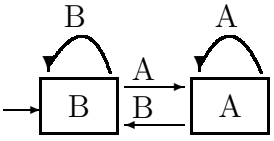
Strategy 18



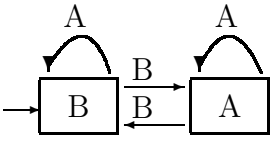
Strategy 7



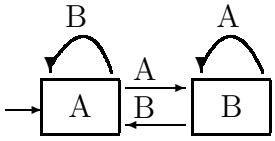
Strategy 8



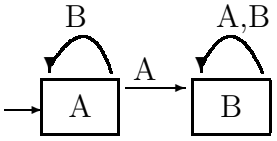
Strategy 19



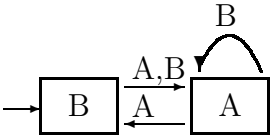
Strategy 20



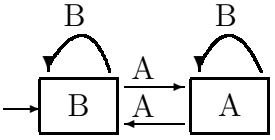
Strategy 9



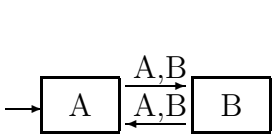
Strategy 10



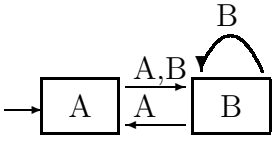
Strategy 21



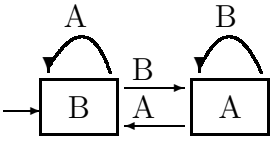
Strategy 22



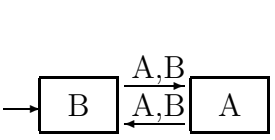
Strategy 11



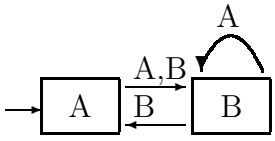
Strategy 12



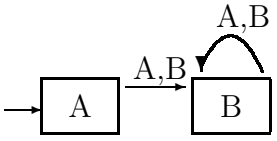
Strategy 23



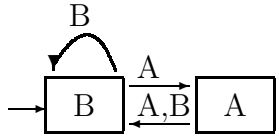
Strategy 24



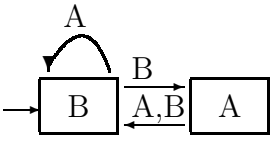
Strategy 13



Strategy 14



Strategy 25



Strategy 26

C Sensitivity of Results to Parameter Changes

The model has four parameters: the strategy update rate ρ, the re-matching rate, μ, the weight ω on reinforcement values in the updating of attractions, and the sensitivity λ of strategy choice to attractions. (The population size N, as long as it is sufficiently large, has no effect on the results.) We have experimented with all possible combinations of the following set of parameter values: $\rho = \{0.2, 0.1, 0.05\}$, $\omega = \{0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.5, 0.7\}$, $\mu = \{0.02, 0.01, 0.005\}$, and $\lambda = \{2.5, 3.0, 3.5, 4.0\}$. For all simulations, the population size is kept constant at 1000. The two parameters of greatest economic interest are the rematching rate μ and the frequency of experimentation ρ. We shall discuss here the sensitivity of our results to changes in these two parameters, holding constant $\omega = 0.1$ and $\lambda = 4.0$.

Recall that our measure of performance is qualitative, namely the extent to which the model's predictions fit the broad contours of the experimental data. For the games considered here, experimental subjects appear to coordinate frequently on patterns of play that produce efficient and fair average payoffs over time. One way to track the sensitivity of the model to parameter changes is therefore to look at deviations of average payoffs from the efficient level. This is done in Table 1, which summarizes the performance of the model for various pairs of ρ and μ. The entries in the table refer to the ratio of the average payoff to the efficient payoff. The parameter pair for which results are reported in the text is in bold.

The results for the 2×2 Stag Hunt are relatively insensitive to changes in strategy updating rates and the re-matching probability. We suspect that this is because convergence occurs to a Nash equilibrium of the stage game, as is predicted also by most action-learning models. The other games show different degrees of sensitivity. Results for the Prisoner's Dilemma are quite sensitive to the strategy update rate ρ. If the strategy update rate is high, i.e., when $\rho = 0.2$, the mean payoff becomes much lower relative to the efficient outcome. This is because if players update their strategies too frequently, reciprocation strategies such as 'tit-for-tat' lose the opportunity to punish defectors for many periods in order to discourage others from defecting in the future. Therefore, a high ρ leads to the situation in which players learn to use strategies that involve more defections. The re-matching probability,

2x2 Stag Hunt

ρ / μ	0.02	0.01	0.005
0.2	0.999	1.000	0.995
0.1	0.999	1.000	0.999
0.05	**0.999**	1.000	0.999

Prisoner's Dilemma

ρ / μ	0.02	0.01	0.005
0.2	0.575	0.585	0.602
0.1	1.000	1.000	0.981
0.05	**1.000**	1.000	0.999

Chicken

ρ / μ	0.02	0.01	0.005
0.2	0.739	0.739	0.742
0.1	0.765	0.769	0.776
0.05	**0.868**	0.854	0.830

Battle of the Sexes

ρ / μ	0.02	0.01	0.005
0.2	0.671	0.686	0.704
0.1	0.715	0.677	0.694
0.05	**0.837**	0.836	0.630

Table 1: Population mean per-period average payoff (mean payoff) relative to the payoff in the efficient and fair outcome for each of the game. λ and ω are set equal to 4.0 and 0.1, respectively.

on the other hand, does not have a strong impact on the result (provided that it remains much lower than the experimentation rate). In the game of Chicken, as in the Prisoner's Dilemma, a higher strategy update rate prevents players from achieving the efficient outcome. The re-matching probability exercises a modest influence in this game, with higher rates of rematching tending to result in lower payoffs.

The environment in which results are most sensitive to parameter values is the Battle of the Sexes. Results are affected both by the strategy updating rate and the re-matching probability to a greater extent than in the other three cases, and these effects are rather complex. When updating is infrequent ($\rho = 0.05$), frequent rematching is helpful in generating alternation. On the other hand, when updating is itself more frequent, then more frequent rematching can lead to declines in efficiency. Hence the learning of successful alternation strategies in this setting depends quite critically on the conditions under which pre-experimental learning occurs. This sensitivity we attribute to the fact that the attainment of fair and efficient outcomes require a pattern of alternation over time that is harder to learn than the repetition of a single action profile.

References

Arifovic, J., R. D. McKelvey, and S. Pevnitskaya (2002): "An Initial Implementation of the Turing Tournament to Learning in Two Person Games," *Mimeo*, California Institute of Technology.

Axelrod, R., and W. D. Hamilton (1981): "The Evolution of Cooperation," *Science*, 211(4489), 1390–1396.

Binmore, K. G., and L. Samuelson (1992): "Evolutionary stability in repeated games played by finite automata," *Journal of Economic Theory*, 57(2), 278–305.

Bolton, G., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

Camerer, C., and T.-H. Ho (1999): "Experience-Weighted Attraction Learning in Normal Form Games," *Econometrica*, 7(4), 827–874.

Charness, G., and M. Rabin (2002): "Understanding social preferences with simple tests," *Quarterly Journal Of Economics*, 117(3), 817–869.

Cheung, Y.-W., and D. Friedman (1997): "Individual Learning in Normal Form Games: Some Laboratory Results," *Games and Economic Behavior*, 19, 46–76.

Crawford, V. P. (1995): "Adaptive Dynamics in Coordination Games," *Econometrica*, 63(1), 103–143.

Erev, I., and A. E. Roth (1998): "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review*, 88(4), 848–881.

——— (1999): "One the Role of Reinforcement Learning in Experimental Games: The Cognitive Game-Theoretic Approach," in *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, ed. by D. V. Budescu, I. Erev, and R. Zwick, chap. 4, pp. 53–77. Lawrence Erlbaum Associates, Inc.

——— (2001): "Simple Reinforcement Learning Models and Reciprocation in the Prisoner's Dilemma Game," in *Bounded Rationality: The Adaptive Toolbox*, ed. by G. Gigerenzer, and R. Selten, chap. 12, pp. 215–231. MIT Press, Cambridge, MA.

Fehr, E., and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114(3), 817–868.

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

Fundenberg, D., and E. S. Maskin (1990): "Evolution and cooperation in noisy repeated games," *American Economic Review Papers and Proceedings*, 80(2), 274–279.

Gintis, H. (2000): "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology*, 206(2), 169–179.

Güth, W., and M. Yaari (1992): "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in *Explaining process and change: Approaches to Evolutionary Economics*, ed. by Witt, pp. 23–34. University of Michigan Press, Ann Arbor, MI.

Huck, S., and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior*, 28(1), 13–24.

Levine, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1(3), 593–622.

McKelvey, R. D., and T. R. Palfrey (2001): "Playing in the Dark: Information, Learning, and Coordination in Repeated Games," *Mimeo*, California Institute of Technology.

Miller, J. H. (1996): "The Coevolution of Automata in the Repeated Prisoner's Dilemma," *Journal of Economics Behavior and Organization*, 29(1), 87–112.

Mookherjee, D., and B. Sopher (1997): "Learning and Decision Costs in Experimental Constant Sum Games," *Games and Economic Behavior*, 19, 97–132.

Osborne, M. J., and A. Rubinstein (1994): *A Course in Game Theory*. The MIT Press, Cambridge, MA.

Sethi, R., and E. Somanathan (2001): "Preference evolution and reciprocity," *Journal of Economic Theory*, 97(2), 273–297.

Stahl, D. O. (1999): "Evidence based rules and learning in symmetric normal-form games," *International Journal of Game Theory*, 28(1), 111–130.

——— (2000): "Rule Learning in Symmetric Normal-Form Games: Theory and Evidence," *Games and Economics Behavior*, 32(1), 105–138.

Stahl, D. O., and E. Haruvy (2002): "Aspiration-Based and Reciprocity-Based Rules in Learning Dynamics for Symmetric Normal-Form Games," *Journal of Mathematical Psychology*, 46(5), 531–553.