

# Simple Reinforcement Learning Models and Reciprocation in the Prisoner's Dilemma Game

Ido Erev<sup>1</sup> and Alvin E. Roth<sup>2</sup>

<sup>1</sup>Faculty of Industrial Engineering and Management, Technion, Haifa 3200, Israel

<sup>2</sup>Dept. of Economics, and Harvard Business School, Harvard University, Cambridge, MA 02138, U.S.A.

## ABSTRACT

The observation that subjects can learn to cooperate in repeated prisoner's dilemma games suggests that human players are more sophisticated and/or less self-interested than the predictions of simple adaptive learning models proposed in recent research. The present chapter demonstrates that this phenomenon is consistent with simple reinforcement learning, when learning is over a strategy set that includes a repeated game strategy that allows reciprocation to be learned. A three-parameter model that quantifies this assumption was found to outperform alternative models in predicting the results of the three experiments conducted by Amnon Rapoport and Mowshowitz (1966).

## INTRODUCTION

Experimental study of the effect of experience on choice behavior typically reveals a slow adaptive adjustment process. If one of the possible alternatives yields higher return, decision makers slowly learn to prefer it. Recent research (e.g., Roth and Erev 1995; Erev and Roth 1998b; Rapoport et al. 1998; Cheung and Friedman 1999; Tang 1996; Camerer and Ho 1999; Mookerjee and Sopher 1997; Sarin and Vahid 1998; Daniel et al. 1998) demonstrates that this common

adjustment process can be approximated by surprisingly simple and general adaptive learning models. For example, Roth et al. (1999) and Erev et al. (1999) show that a two-parameter reinforcement learning model that looks only at immediate reinforcement of simple, one-period actions, and assumes a very low level of rationality, can be used to predict behavior in more than 60 binary choice tasks.

An important exception to this regularity occurs in strategic settings that facilitate reciprocity (situations in which players can coordinate and benefit from mutual cooperation). For example, Anatol Rapoport and Channah (1965) and subsequently Amnon Rapoport and Mowshowitz (1966) found that in certain two-person iterated prisoner's dilemma (PD) games (e.g., the game presented in Figure 12.2) experience reduces the frequency of selecting the alternative that yields higher payoffs (alternative D). Instead, some players learn to cooperate (the proportion of C choices increases with time).

In a recent paper (Erev and Roth 1998a) we speculated that the latter finding may not reflect a limitation of the reinforcement learning approach (and of the assumption of bounded rationality), rather that the players learn not just from immediate actions (stage-game strategies) but also from repeated game strategies. To capture learning in games, reciprocity strategies must be explicitly modeled. In this chapter, we take a preliminary step toward this goal by proposing and evaluating a reinforcement learning model in which, in addition to stage-game strategies, players can learn a reciprocity strategy. (A fuller treatment would consider more carefully how particular repeated game strategies arise. Here we simply investigate the extent to which reinforcement learning can capture the emergence of cooperation when learning is among repeated game strategies.)

We first summarize recent findings which demonstrate that behavior in simple games can be captured by models of reinforcement learning over actions. A two-parameter model that can account for behavior in matrix games in which players cannot reciprocate is presented. Next, we review the robust regularities observed in previous studies of repeated PD games, focusing on the four experiments conducted by Rapoport and Mowshowitz (1966) to test simple Markovian learning models.

In the third section, we show that a minimal generalization of the two-parameter basic reinforcement learning model, the addition of a "forgiving" reciprocity strategy, can be used to capture Rapoport and Mowshowitz's findings. We estimated the model's three parameters on the basis of the data of Experiment 1 in Rapoport and Mowshowitz (1966) and evaluated it using the results of Experiments 2, 3, and 4. We conclude with a discussion of potential extensions of the model and the main implications of the current results to the study of bounded rationality.

## REINFORCEMENT LEARNING

In an earlier paper, we demonstrated the potential of reinforcement learning models (Erev and Roth 1998b). Utilizing a wide set of experiments that study the effect of long experience (more than 100 trials) in bi-matrix games in which players could not reciprocate (Suppes and Atkinson 1960; Malcolm and Lieberman 1965; O'Neill 1987; Rapoport and Boebel 1992; Ochs 1995), we showed that observed behavior is reasonably approximated by simple adaptive learning models. The results obtained from five of these games are summarized in Figure 12.1. Experimental results and equilibrium predictions are presented in the left-hand column. The three right-hand columns present the predictions of three adaptive learning models (one- and three-parameter reinforcement learning models, and a four-parameter generalized fictitious play model). All models appear to capture the major trends in behavior both when it is consistent and inconsistent with equilibrium predictions.

To evaluate the predictive power of the adaptive learning models, we (in Erev and Roth 1998b) calculated the mean squared deviation (MSD) between the mean results and the different predictions. The MSD score of equilibrium was 0.035; however, all learning models considered had MSDs below 0.01. The best model, the three-parameter reinforcement learning, had an MSD of 0.006 when the three parameters were estimated to fit the data, and an MSD of 0.007 when the data of each experiment was predicted by the parameters that best fitted all other games.

More recently (Roth et al. 1999; Erev et al. 1999), we have shown that a two-parameter reinforcement learning model outperforms the three models that we studied earlier. With a single set of (two) parameters, this model captures the 11 games studied by Erev and Roth (1998b), 9 probability learning tasks (Erev et al. 1999), and 40 randomly selected constant sum games (Roth et al. 1999; a "representative sample" in Gigerenzer et al.'s [1991] terminology). This model can be summarized by the following three learning assumptions.

### L1 Initial Propensities

Players have an initial propensity to play their stage-game strategies (i.e., their simple one-period actions), and only these strategies. At time  $t = 1$  (before any experience has been acquired), the players are indifferent between their strategies. Specifically, let  $A_n(1)$  be the expected payoff to  $n$  if all players choose their strategies randomly, with equal likelihood. Player  $n$ 's initial propensity to select strategy  $j$  equals this expected payoff from random choices, i.e., for each player  $n$ ,

$$q_{nj}(1) = A_n(1) \quad (12.1)$$

for all pure strategies  $j$ .

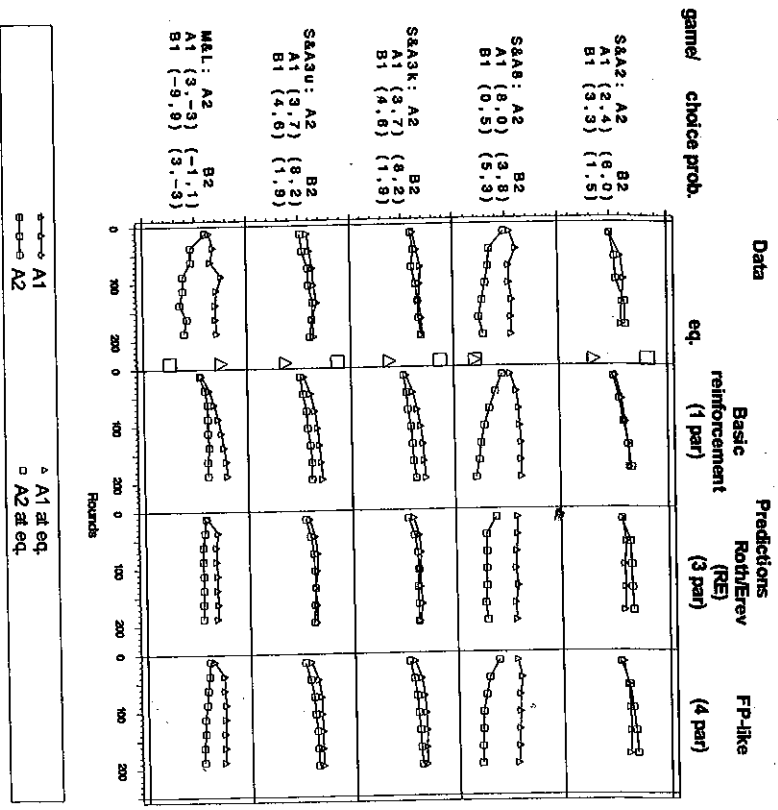


Figure 12.1 Repeated 2 x 2 games (S&A: Suppes and Atkinson 1960; M&L: Malcolm and Lieberman 1965). In the top four games, each payoff unit increases the probability of winning by 1/6 in S&A2, by 1/8 in S&A8, and by 1/10 in S&A3k and S&A3u. In the fifth game, payoffs were directly converted to money. Each cell in the left-hand column presents the experimental results: the proportion of A choices over subjects in each role (grouped in 5 to 8 blocks) as a function of time (200–210) trials in all cases. The three right-hand columns present the models' predictions in the same format. The equilibrium predictions are presented at the right-hand side of the data cells. Adapted from Erev and Roth (1998b).

**1.2 Average Updating**

The propensity of Player  $n$  to play strategy  $j$  at trial  $t + 1$  is a weighted average of the initial propensity ( $q_{nj}^0$ ) and the average payoff obtained by Player  $n$  from playing  $j$  in the first  $t$  trials ( $AVE_{nj}(t)$ ). The weight of the initial propensity is a function of an initial strength parameter  $N(1)$  and the number of time Player  $n$  chose strategy  $j$  ( $C_{nj}(t)$ ). Specifically,

$$q_{nj}(t+1) = q_{nj}^0 \frac{N(1)/m_n}{C_{nj}(t) + N(1)/m_n} + AVE_{nj}(t) \frac{C_{nj}(t)}{C_{nj}(t) + N(1)/m_n} \quad (12.2)$$

where  $m_n$  is the number of Player  $n$ 's pure strategies. Thus,  $N(1)$  can be naturally interpreted as the weight of the initial propensities, in units of number of experiences with the game (and  $N(1)/m_n$  is the number of "initial subjective experiences" with each strategy).

This averaging assumption implies that the updating from trial to trial can be described as follows: If Player  $n$  plays his  $j^{\text{th}}$  pure strategy at time  $t$  and receives a payoff  $x$ , then the propensity to play strategy  $j$  is updated to be:

$$q_{nj}(t+1) = q_{nj}(t)W(t) + x(1 - W(t)), \quad (12.2a)$$

where

$$W(t) = \frac{C_{nj}(t) + N(1)/m_n}{C_{nj}(t) + N(1)/m_n + 1} \quad (12.2b)$$

This trial-to-trial formulation reveals that the propensity to play strategy  $j$  increases with relatively high payoffs ( $x > q_{nj}(t)$ ) and decreases with low payoffs ( $x < q_{nj}(t)$ ). In addition, the updating speed decreases with experience, i.e.,  $W(t)$  increases with  $C_{nj}(t)$ .

**1.3 Exponential Response Rule**

Probability,  $P_{nj}(t)$ , that Player  $n$  plays his  $j^{\text{th}}$  pure strategy at time  $t$  is given by:

$$P_{nj}(t) = \frac{EXP[q_{nj}(t)\lambda/S_n(t)]}{\sum_k EXP[q_{nk}(t)\lambda/S_n(t)]} \quad (12.3)$$

where the sum is over all of Player  $n$ 's pure strategies  $k$ ,  $\lambda$  is a free parameter that determines reinforcement sensitivity, and  $S_n(t)$  is a measure of the standard deviation of the payoffs Player  $n$  has experienced up to time  $t$ . Thus the probability of selecting a strategy increases with the propensity to select it (which increases with the average payoff from past selections). The division by the standard deviation measure implies that noisy reinforcements reduce reinforcement sensitivity (leading toward more uniform choice probabilities).

The standard deviation is estimated as the average absolute difference between the recent payoff ( $x$  at trial  $t$ ) and the accumulated average payoff in the first  $t$  trials ( $A_n(t)$ ). Following the logic of Equation 12.2a:

$$S_n(t+1) = S_n(t)W'(t) + |A_n(t) - x|(1 - W'(t)), \quad (12.4)$$

where

$$W'(t) = \frac{t + N(1)}{t + N(1) + 1} \quad (12.4a)$$

The initial value  $S_n(1)$  is the expected distance of the payoff from random choices from the expected payoff given random choice. (Note that the model is defined only for positive  $S_n(1)$ ) Average payoff,  $A_n(t)$ , is calculated in a similar manner:

As noted above,  $A_n(1)$  is the expected payoff from random choice.

$$A_n(t+1) = A_n(t)W'(t) + A_n(t)(1 - W'(t)) \tag{12.5}$$

### RECIPROCATION

In contrast to the success of simple reinforcement learning of stage-game strategies in games in which players cannot reciprocate, these models are clearly violated in games that allow for reciprocation. For example, stage-game strategy learning models predict a decrease in cooperation in PD games (cf. Figure 12.2). In contrast to these predictions, previous research on the effect of experience in iterated two-person PD games reveals that players can learn to reciprocate in some games.

An extensive examination of learning under conditions that facilitate reciprocation is provided by Rapoport and Chamah (1965) and by Rapoport and Mowshowitz (1966). Both studies explored behavior in 300 repetitions of PD games. Rapoport and Chamah explored seven games under distinct conditions, while Rapoport and Mowshowitz focused on one of these games and studied the interaction between human and preprogrammed strategies. Since Rapoport and Mowshowitz replicated Rapoport and Chamah's main results and added experiments that facilitate model development and comparison, we focus here on the four experiments conducted by Rapoport and Mowshowitz. Their main results can be summarized by the following five sets of robust behavioral regularities.

#### Increase in Mutual Cooperation with Time

In Figure 12.2, the left-hand column presents the percentage of trials in which both players cooperated (CC choices) in 6 blocks of 50 trials over the 19 pairs participated in Experiment 1 of the Rapoport and Mowshowitz study. The payoff matrix used in that study is presented on the top of the figure. The results reveal an increase in mutual cooperation over time. The second column shows the predictions of the basic RE model with only two strategies (C = cooperative; D = defect) and with the original parameters, and reveals that the model fails to capture the data. Moreover, the incorrect prediction of a decrease in joint cooperation is robust to the choice of parameters and the specific variant of the model: all the reinforcement learning models we considered (in Erev and Roth 1998b) predicted a drop in cooperation when the only strategies are "C" and "D." The third and fourth columns are simulations of the models developed and discussed below (see section on MODELING RECIPROICATION), which enlarge the set of strategies players may learn.

		Player 2	
		C	D
Player 1	C	(1, 1)	(-10, 10)
	D	(10, -10)	(-5, -5)

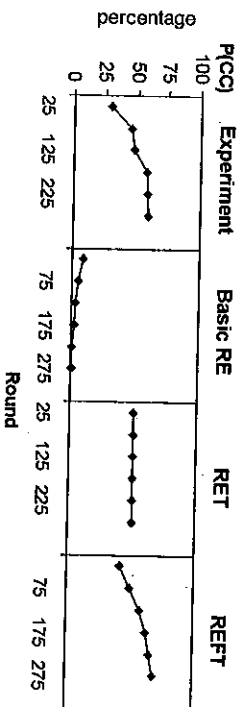


Figure 12.2 The prisoner's dilemma game studied by Rapoport and Mowshowitz (1966). The bottom panel displays the proportion of mutual cooperation (the CC [mutual] outcomes in 6 blocks of 50 trials) in Experiment 1 of Rapoport and Mowshowitz (1966) (left), and the relevant predictions of three models: the basic two-parameter model, the extension that assumes a Tit-for-Tat (TFT) strategy (RET), and the extension that assumes a forgiving TFT strategy (REFT).

#### Large Between-pair Variance

The first panel of Figure 12.3 presents the distribution of the proportion of cooperation (C) over the 300 trials summed across the 38 participants (Rapoport and Mowshowitz, Experiment 1). The results reveal extremely large variance. In fact, with the exception of a mode on 90–100%, the distribution is almost uniform.

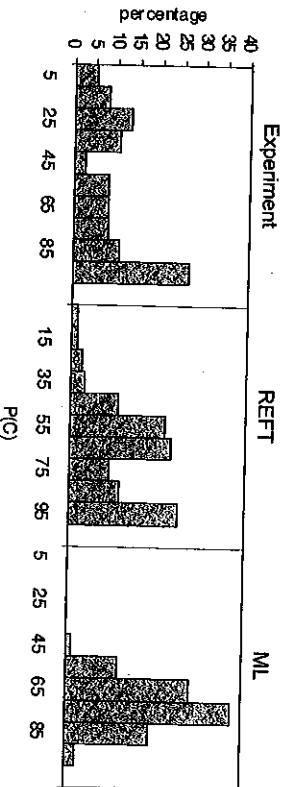


Figure 12.3 Distribution of C choices for all the subjects in Experiment 1 of Rapoport and Mowshowitz (1966) (left), and the predicted distribution by the REFT and the Markov/Learning (ML) models.

**Sequential Effects**

Rapoport and Mowshowitz (1966) modeled the probability of cooperation in trial  $r+1$  as a function of the decisions made in trial  $r$  by both players. Following Rapoport and Mowshowitz, we denote the conditional probability of cooperation by Player  $n$  following a decision  $X$  by  $n$  and  $Y$  by Player  $o$  as  $P(C|XY)$ . Summed across trials and players the observed probabilities are:  $P(C|CC) = 0.81$ ,  $P(C|CD) = 0.43$ ,  $P(C|DC) = 0.37$ , and  $P(C|DD) = 0.22$ . These probabilities are displayed graphically in Figure 12.4. Note that whereas they show some responsiveness to the other player's recent choice, the responsiveness is much weaker than the responsiveness predicted under the TFT (tit-for-tat) strategy (for further explanation, see section below, An RET Model). Under this strategy (to be discussed below) the expected probabilities are  $P(C|CC) = 1$ ,  $P(C|CD) = 0$ ,  $P(C|DC) = 1$ , and  $P(C|DD) = 0$ .

**Sensitivity to the Opponent's Choice Probabilities**

To facilitate a clean test of Markov chain models, Rapoport and Mowshowitz (1966) conducted three experiments in which human subjects played against preprogrammed opponents (using the payoff matrix of Experiment 1). In Experiment 2, the opponents were humans whose choices were determined by the following conditional probabilities:  $P(C|CC) = 0.76$ ,  $P(C|CD) = 0.25$ ,  $P(C|DC) = 0.46$ , and  $P(C|DD) = 0.22$ . In Experiment 3 the participants played against a computer whose choice probabilities were  $P(C|CC) = 0.72$ ,  $P(C|CD) = 0.26$ ,  $P(C|DC) = 0.42$ , and  $P(C|DD) = 0.22$ . Finally, in Experiment 4 the game was played against a "learning" computer program. This program used three fixed choice probabilities,  $P(C|CC) = 0.41$ ,  $P(C|DC) = 0.36$ , and  $P(C|DD) = 0.17$ , but increased the probability of repeated cooperation in trial

$r + 1$  ( $P_{r+1}(C|CC)$ ), following mutual cooperation. Specifically, the following linear operator model was assumed:

$$P_{r+1}(C|CC) = \begin{cases} \alpha P_r(C|CC) + (1-\alpha)\beta, & \text{if mutual cooperation was achieved at } r-1, \\ P_0(C|CC), & \text{otherwise;} \end{cases} \quad (12.6)$$

where  $P_0(C|CC) = 0.916$ ,  $\alpha = 0.7$ , and  $\beta = 0.985$ . (These values were set to fit the results of Experiment 1. We return to this point below.)

The first panel of Figure 12.5 shows the proportion of C choices by the human subjects in the four experiments. It shows that whereas the identity of the opponents (human or computer) had a small effect (cf. Experiment 1 to 4 and Experiment 2 to 3), the opponent's choice probabilities had a large effect.

**Failure of Simple Markov Chain Models**

Rapoport and Mowshowitz (1966) tried to account for their data using a simple Markov chain model that assumes fixed and independent switching probabilities from state to state. In Experiments 2 and 3, this model has four free parameters, the conditional probabilities ( $P(C|CC)$ ,  $P(C|CD)$ ,  $P(C|DC)$ , and  $P(C|DD)$ ) that determine the switching probabilities. Rapoport and Mowshowitz found that when the parameters are estimated by the observed switching probabilities (in the relevant experiment), the model under-predicts the frequency of mutual cooperation. In particular, it predicts 16% of the CC outcomes in Experiment 2

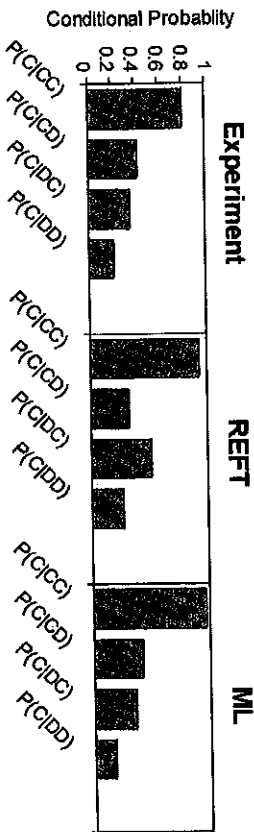


Figure 12.4 Conditional probabilities in Experiment 1 of Rapoport and Mowshowitz (1966) (left), and the predicted probabilities by the REFT model and the Markov/Learning (ML) model.

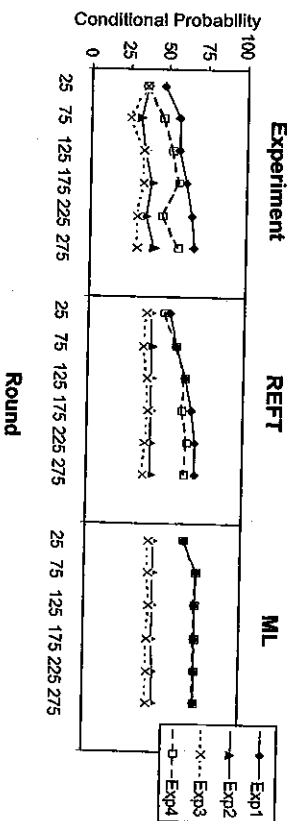


Figure 12.5 Cooperation rate (C choices by the human subjects) in the four experiments run by Rapoport and Mowshowitz (1966) (left), and the predictions of the REFT and the Markov/Learning (ML) models. Note that parameters were estimated based on Experiment 1.

and 11% of the CC outcomes in Experiment 3. The observed CC percentages were 21% and 15%, respectively.

## MODELLING RECIPROCATION

To distinguish among the possible explanations for the failure of the simple reinforcement learning models in games in which players can reciprocate, it is convenient to perform a "cognitive game theoretic" decomposition of these models. The model presented above and other reinforcement learning models can be decomposed into three basic submodels:

1. An abstraction of the incentive structure: all of the models we considered implicitly assume that subjects are sensitive to the objective payoff function.
2. An abstraction of the cognitive strategies considered by the subjects: in Erev and Roth (1998b), we modeled subjects as learning among simple actions (stage-game strategies).
3. An abstraction of the decision/learning rule among the relevant strategies: here we have focused on one particular reinforcement learning rule, whereas in Erev and Roth (1998b) we also studied expectation-based adaptive learning rules.

Previous explanations of reciprocation considered potential violations of all three submodels. Incentive-based explanations were provided by Kelly and Tibbaut (1978) and from recent work by Bolton and Ockenfels (1998) and Fehr and Schmidt (1999). According to these accounts, reciprocation is a result of subjective incentives that differ from the objective monetary incentives. Explanations based on cognitive strategies were provided by Axelrod (1980, 1984), Komorita and Parks (1995, 1998), and Messick and Liebrand (1995). In addition, Amnon Rapoport and Mowshowitz (1966) and Macy (1991) account for reciprocation by invoking particular learning rules.

While all three classes of explanations provide insightful accounts of reciprocation in many settings, it seems that a modification of the cognitive strategies' submodel is necessary to account for reciprocation in repeated games. One of the most robust findings reported by Anatol Rapoport et al. (1976; see also Bornstein et al. 1997) is that players were able to reciprocate by alternation in chicken-type games. Since alternation is a strategy that involves memory of the previous period, it could not be observed that subjects confined their learning to stage-game strategies. A mere modification of the other two submodels (the assumed learning rule and incentive structure) cannot explain this observation.

Our main goal in this chapter is to show how this apparently necessary modification can be accomplished in a reinforcement learning model. We hope to show that this necessary modification will turn out to be sufficient to account for the five regularities described above.

### An RET Model

Previous study of behavior in PD games reveals that the Tit-for-Tat (TFT) strategy can lead to efficient cooperation (e.g., Roth and Murnighan 1978; Axelrod 1980) in these settings. The TFT strategy can be thought of as the simplest quantification of the concept "reciprocation." Formally it states that

- R1 If Player  $n$  chooses to start reciprocating at trial  $t$ , he or she cooperates at that trial and cooperates at trial  $t + 1$  if and only if the other player (Player  $o$ ) has cooperated at  $t$

To apply the TFT strategy in the current adaptive learning framework, its length has to be defined. To allow learning, players should have multiple experiences with this and the alternative strategies. Moreover, it is natural to assume that the utilization length increases with experience and the potential benefit from reciprocation. These behavioral assumptions are quantified here as follows:

- R2 If Player  $n$  chooses a reciprocation strategy (TFT, denoted here as strategy  $k$ ) at trial  $t$ , the probability that he or she will continue to utilize it at trial  $t + 1$  is:

$$\text{CONT}(t) = \begin{cases} \frac{C_{nk}(t)}{C_{nk}(t) + N(0)} PR^\rho & \text{if } PR > 0 \\ 0 & \text{otherwise} \end{cases} \quad (12.7)$$

where  $C_{nk}(t)$  is the number of times that the reciprocation strategy was played,  $N(0) > 0$  is "convergence speed" parameter,  $\rho$  is a "playing length" parameter, and  $PR$  is the relative potential gain from reciprocation. The potential is estimated as:

$$PR = \frac{V_n(\text{rec}) - A_n(1)}{V_n(\max) - V_n(\min)} \quad (12.8)$$

where  $V_n(\text{rec})$  is the expected payoff for Player  $n$  given successful reciprocation,  $A_n(1)$  is the expected payoff from random play, and  $V_n(\max) > V_n(\min)$  are the best and worst possible payoffs for Player  $n$ . In the current game  $V_n(\text{rec}) = 1$ ,  $A_n(1) = -1$ ,  $V_n(\max) = 10$ , and  $V_n(\min) = -10$ . Thus, the potential from reciprocation,  $PR$ , equals 0.1.

Note that the current quantification implies that when reciprocation is not expected to lead to a higher payoff than the payoff from random play ( $PR \leq 0$ ) it will not be played more than once per strategy choice. When reciprocation is expected to lead to higher payoff ( $PR > 0$ ), the probability of repeated selection converges with experience to  $PR^\rho$ , the parameter  $N(0)$  determines the speed of this convergence process. Thus, like  $N(1)$ ,  $N(0)$  can be thought of as a "conservatism" parameter. To reduce the number of free parameters we impose the constraint  $N(0) = N(1)$ .

The first model considered here, referred to as the RET model, is a minimal extension of the reinforcement model described above (see section on REINFORCEMENT LEARNING). The extension implies that players consider three strategies: the two stage-game strategies and the "probabilistic length" TFT strategy as defined by assumptions R1 and R2.

Consequently, this model has three parameters: The two learning parameters  $N(1)$  and  $\lambda$  and one reciprocation length parameter ( $\rho$ ). To evaluate the model we first derived its predictions of the evolution of the mutual cooperation rate in Experiment 1 of Rapoport and Mowshowitz (1966) and found the parameters that best fit the data (minimize the MSD between the observed and predicted aggregated 6-block learning curve).

To derive the predictions of this model for the current game, we conducted computer simulations in which 1000 virtual pairs of players played the game for 300 rounds (a replication of Experiment 1 in Rapoport and Mowshowitz). To describe the simulation (and explain the model) it is useful to distinguish between the computation conducted before the first trial and during each trial. Before the first trial the (uniform) initial propensities,  $A_n(1)$  and  $S_n(1)$  were calculated for each player. In the current game,

$$A_n(1) = \frac{1-10+10-5}{4} = -1,$$

$$S_n(1) = \frac{|1-(-1)|+|(-10)-(-1)|+|0-(-1)|+|-5-(-1)|}{4} = 7, \quad (12.9)$$

and  $q_{nj}(t) = A_n(1) = -1$

for all  $n$  and  $j$ . In addition, the players' state was set to equal "decide."

The following steps were then taken during the simulation (in trial  $t$  of the simulation of each dyad):

1. *Strategy selection.* Each player whose state was "decide" selected one of the three possible strategies (C, D, or TFT) with the probabilities given by assumption L3. The counter of the selected strategy ( $C_{nj}(t)$ ) was updated (by adding 1). If the selected strategy was TFT, the player's state was set to "TFT-continue."
2. *Alternative selection.* If the state was not "TFT-continue," the selected alternative was identical to the strategy (C or D). When the state was "TFT-continue," which implies that the strategy was TFT, Player  $n$ 's selected alternative was C if and only if at least one of the following conditions (from R1) were met:
  - a. Player  $n$ 's opponent (Player  $o$ ) chose C at trial  $t-1$ .
  - b. Player  $n$  selected the TFT at trial  $t$  (the State was "decide" at the beginning of the trial).
3. *Payoff calculation.* The payoffs were determined by the chosen alternatives (using Figure 12.2's payoff matrix).

4. *Propensities updating.* The propensities of the selected alternatives were updated using assumption L2.
5. *Deviation and average measures.*  $S(t+1)$  and  $A(t+1)$  were calculated using the rules specified in assumption L3 (note that  $S(t+1)$  is updated based on  $A(t)$ ).

6. *Updating the state.* If the state was TFT-continue, a virtual coin was flipped and with probability  $1-\text{CONT}(t)$  the state was change to "decide" ( $\text{CONT}(t)$  was calculated using assumption R2).

The predictions of the RET model with the parameters that best fit the data in a grid search ( $N(1) = 10$ ,  $\lambda = 100$ ,  $\rho = 0.03$ ) are presented on the third column of Figure 12.2. The results reveal that the model predicts that almost all learning will occur in the first 50 trials. Thus, it does not capture the slow increase in mutual cooperation observed by Rapoport and Mowshowitz (1966).

#### A More Forgiving Model (REFT)

Previous research (e.g., Axelrod 1984) suggests that the failure of the TFT strategy, as quantified above, is likely to be a result of not being "forgiving" enough to capture cooperation in a noisy environment. For example, a single mistake (deviation from TFT) by one of the players on trial  $t$  implies no mutual cooperation until the two players deviate from TFT together at the same trial. Since players are assumed to make independent choices, a single deviation is more likely to occur than a joint deviation.

To evaluate if making the reciprocation more forgiving can improve the fit of the model to the Rapoport and Mowshowitz (1966) results, the current model replaces assumption R1 with a more forgiving assumption, which implies that if both players choose to reciprocate, cooperation will be achieved within two periods. Specifically, it is assumed that:

- R1f If Player  $n$  chooses to start reciprocating at trial  $t$ , he cooperates at that trial and cooperates at trial  $t+1$  if and only if at least one of following conditions is met:
- a. Player  $o$  has cooperated at trial  $t$ ,
  - b. Player  $o$  has cooperated at trial  $t-1$ , and  $n$  has not cooperated at trial  $t$ .

The predictions of the REFT model were derived using the simulation presented above after replacing the R1 conditions with the R1f condition (in step 2). The model's predictions with the parameters that best fit the data ( $N(1) = 1$ ,  $\lambda = 3.3$ ,  $\rho = 0.05$ ) are presented in the fourth column of Figure 12.2. The plot shows that the forgiving model captures the increase in cooperation and outperforms the RET model.

Although the REFT model was fitted to the aggregate curve, it tends to predict large between-subject variability. Figure 12.3 compares the distribution of

C choices in the Rapoport and Mowshowitz (1966) experiment and in the simulation of the REFT model, and reveals a similar (but not identical) almost uniform distribution.

Figure 12.4 compares the observed and predicted conditional probabilities (in Experiment 1). It shows that the model captures the main experimental trends: high  $P(C|CC)$  and low  $P(C|DD)$  values. However, the model incorrectly predicts that  $P(C|CD)$  exceeds  $P(C|DC)$ . This bias suggests that the current "forgiving" model is not "forgiving enough" to capture exactly the sequential dependencies in the Rapoport and Mowshowitz study.

#### Predictive Power and Alternative Models

The second column in Figure 12.5 shows the model's predicted C rate in the four experiments run by Rapoport and Mowshowitz. Note that the parameters were estimated based on Experiment 1 results. Thus, the fit of Experiments 2, 3, and 4 allows evaluation of the model's predictive power. Figure 12.4 shows that the model captures the rank ordering and trends observed in the three experiments: increase in cooperation in Experiment 4, a flat curve in Experiment 2, and lowest level of cooperation in Experiment 3. In addition, the simulations reveal that the current "parameter-free" predictions of Experiments 2 and 3 outperform the prediction of the four-parameter Markov chain model discussed above. The average CC percentages predicted by the current model (in the last 250 trials) are 22% and 18%.

The failure of the basic Markov chain model led Rapoport and Mowshowitz to propose a variant of this model, which assumed that the likelihood of an exit from a CC state decreases with repeated CC outcomes. This Markov/Learning (ML) model was implemented as the experimental condition in Experiment 4 and described above. Note that it has six parameters (the parameters are estimated from the observed sequential dependencies). The third column in Figure 12.5 shows the predictions of the ML model with the parameters estimated by Rapoport and Mowshowitz based on the results of Experiment 1. Whereas this model captures the main experimental trends and clearly outperforms the Markov chain model, its quantitative predictions are less accurate than the predictions of the REFT model. In all four cases the REFT model is closer to the experimental curves.

A more obvious advantage of the REFT model over the ML model is displayed in Figure 12.3. Here the ML model predicts a normal shape distribution of individual subjects and fails to capture the observation that a significant portion of pairs appear to become stuck at a very low frequency of cooperation.

Finally, Figure 12.4 shows that the ML model outperforms the REFT model in describing the conditional probabilities. The advantage of the ML model in this case is not surprising as three of its parameters were estimated from these data.

### ADDITIONAL BEHAVIORAL REGULARITIES

Research in progress (Erev and Roth, in preparation) suggests that the predictive power of the current model is not limited to the payoff matrix used by Rapoport and Mowshowitz (1966). Similarly, good fits are found for all the PD games studied by Rapoport and Chamnah (1965). With the parameters estimated here (on Experiment 1 of Rapoport and Mowshowitz 1966), the model captures the effect of manipulations on the payoff matrix. Most importantly, it accurately predicts that cooperation will be obtained by most players in certain PD games, but not in others.

Erev and Roth (in preparation) also suggest that with a more general definition of the reciprocation strategy, the current model can account for behavior in all the repeated  $2 \times 2$  games studied by Rapoport et al. (1976). For example, it can capture the conditions under which players are able to reach efficient and fair alternation outcomes.

Future research is needed to extend the current model to address repeated supergames like the prisoner's dilemma studied by Selten and Stoeker (1986).

### CONCLUSIONS

Repeated games give players the opportunity to achieve cooperation through the use of repeated game strategies that make current actions contingent on the previous history of play. Learning models that do not allow players to make such contingent decisions cannot reproduce the observed behavior. In this chapter, we discussed how only a little sophistication must be added to the reinforcement learning model previously explored (a two-parameter model that captures behavior in binary decision tasks and games in which players cannot reciprocate) in order to model behavior in the repeated prisoner's dilemma. Results reveal that the addition of a single reciprocation strategy, together with the two stage-game strategies, is sufficient to permit simple reinforcement learning to capture many aspects of the observed data. A three-parameter model that quantifies this assumption was found to outperform alternative models in predicting the results of the three experiments conducted by Rapoport and Mowshowitz (1966) to compare alternative descriptive models of behavior in the PD game.

Although the current model is only an example of a strategy set sufficient to capture the effect of experience on cooperation, it has two important advantages over some previous explanations. First, most recent research focuses on the qualitative demonstrations that particular assumptions can lead to reciprocation. Current research (like the pioneering research by Anatol Rapoport and Chamnah [1965] and Amnon Rapoport and Mowshowitz [1966]) shows that useful quantitative predictions can be made.

Second, the reinforcement learning "engine" is potentially general. The present reinforcement learning model is a generalization of a model that captures

behavior in more than 60 games in which players cannot reciprocate and, as we hope to show in the future (Erev and Roth, in prep.), can be extended to any  $2 \times 2$  game.

In general, the approach that we have outlined here involves separating players' learning rules from their cognitive model of the available strategies. The present example is intended to make plausible the speculation that, with an appropriate model of strategies, even very simple models of reinforcement learning may be sufficient to predict how behavior will evolve over time in repeated play of simple games.

#### ACKNOWLEDGMENT

This research was supported by the Henry Gutwirth Promotion of Research Fund, Israel-U.S.A. Binational Science Foundation, and the National Science Foundation (U.S.A.).

#### REFERENCES

- Axelrod, R. 1980. Effective choice in the prisoner's dilemma. *J. Confl. Res.* 24:3-26.
- Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic.
- Bolton, G., and A. Ockenfels. 1998. Strategy and equity: An ERC-analysis of the Guth-van Damme game. *J. Math. Psych.* 42:215-226.
- Bornstein, G., D.V. Budescu, and S. Zamir. 1997. Cooperation in intergroup, N-person and two-person games of chicken. *J. Confl. Res.* 41:384-406.
- Camerer, C.F., and T. Ho. 1999. Experience-weighted attraction in games. *Econometrica* 67:827-874.
- Cheung, Y.W., and D. Friedman. 1999. A comparison of learning and replicator dynamics using experimental data. *J. Econ. Behav. Org.*, in press.
- Daniel, T.E., D.A. Seale, and A. Rapoport. 1998. Strategic play and adaptive learning in sealed bid bargaining mechanism. *J. Math. Psych.* 42:133-166.
- Erev, I., Y. Bercby-Meyer, and A. Roth. 1999. The effect of adding constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. *J. Econ. Behav. Org.*, in press.
- Erev, I., and A.E. Roth. 1998a. On 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. D. Budescu, I. Erev, and R. Zwick, pp. 53-78. Mahwah, NJ: Erlbaum.
- Erev, I., and A.E. Roth. 1998b. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* 88(4):848-881.
- Fehr, E., and K. Schmidt. 1999. How to account for fair and unfair outcomes—A model of biased inequality aversion. *Q. J. Econ.*, in press.
- Gigerenzer, G., U. Hoffrage, and H. Kleinbolting. 1991. Probabilistic mental models: A Brunswikian theory of confidence. *Psych. Rev.* 98:506-528.
- Kelly, H.H., and J. Tibbaut. 1978. *Interpersonal Relations: A Theory of Independence*. New York: Wiley.
- Komoria, S.S., and C.D. Parks. 1995. Interpersonal relations: Mixed-motive interaction. *Am. Rev. Psych.* 46:183-207.
- Komoria, S.S., and C.D. Parks. 1998. Reciprocity and cooperation in social dilemmas: Review and future directions. In: *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, ed. D. Budescu, I. Erev, and R. Zwick, pp. 315-330. Mahwah, NJ: Erlbaum.
- Macy, M.W. 1991. Learning to cooperate: Stochastic and tacit collusion in social exchange. *Am. J. Soc.* 97:808-843.
- Malcolm, D., and B. Lieberman. 1965. The behavior of responsive individuals playing a two-person, zero-sum game requiring the use of mixed strategies. *Psychonomic Sci.* 12:373-374.
- Messick, D.M., and W.B.G. Liebrand. 1995. Individual heuristics and the dynamics of cooperation in large groups. *Psych. Rev.* 102:131-145.
- Mookherjee, D., and B. Sopher. 1997. Learning and decision costs in experimental constant sum games. *Games Econ. Behav.* 19:62-91.
- Ochs, J. 1995. Simple games with unique mixed strategy equilibrium: An experimental study. *Games Econ. Behav.* 10(1):202-217.
- O'Neill, B. 1987. Nonmetric test of the minimax theory of two-person zerosum games. *Proc. Natl. Acad. Sci. USA* 84:2106-2109.
- Rapoport, Amnon, and R.B. Boebel. 1992. Mixed strategies in strictly competitive games: A further test of the minimax hypothesis. *Games Econ. Behav.* 4:261-283.
- Rapoport, Amnon, and A. Mowshowitz. 1966. Experimental studies of stochastic models for the prisoner dilemma. *Behav. Sci.* 11:444-458.
- Rapoport, Amnon, D.A. Seale, I. Erev, and J.A. Sundali. 1998. Coordination success in market entry games: Tests of equilibrium and adaptive learning models. *Manag. Sci.* 44:129-141.
- Rapoport, Anatol, and A.M. Channanah. 1965. Prisoner's Dilemma: A Study in Conflict and Cooperation. Ann Arbor: Univ. of Michigan Press.
- Rapoport, Anatol, M.J. Guyer, and D.G. Gordon. 1976. The  $2 \times 2$  Game. Ann Arbor: Univ. of Michigan Press.
- Roth, A.E., and I. Erev. 1995. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games Econ. Behav. (Spec. Iss.: Nobel Symp.)* 8:164-212.
- Roth, A.E., I. Erev, R.L. Slovic, and G. Barron. 1999. Learning and equilibrium as useful approximations: Accuracy of predictions in randomly selected constant sum games. Harfa: Technion. Mimeo.
- Roth, A.E., and J.K. Murnighan. 1978. Equilibrium behavior and repeated play of the prisoners' dilemma. *J. Math. Psych.* 17:189-198.
- Sarin, R., and F. Vahid. 1998. Predicting how people play games: A procedurally rational model of choice. Texas A&M Univ. Mimeo.
- Selten, R., and R. Stoeker. 1986. End behavior in sequences of finite prisoner's dilemma supergames: A learning theory approach. *J. Econ. Behav. Org.* 7:47-70.
- Suppes, P., and R.C. Atkinson. 1960. *Markov Learning Models for Multiperson Interactions*. Palo Alto: Stanford Univ. Press.
- Tang, F. 1996. Anticipatory learning in two-person games: An experimental study. II. Learning. Discussion Paper B-363. Bonn: Univ. of Bonn.