Merge Forward and Backward Models.doc: 8087 Words

WHERE FORWARD-LOOKING AND BACKWARD-LOOKING MODELS MEET*

Peter J. Burke

Louis N. Gray

Washington State University

Draft Copy, Do Not Cite or Quote Without Permission.

* Prepared for the "Evolutionary Games and Networks" session of the Mathematical Sociology Section of the American Sociological Association 1997 meetings in Toronto, Ontario, Canada.

WHERE FORWARD-LOOKING AND BACKWARD-LOOKING MODELS MEET

Abstract

The present paper begins by deriving an instantaneous formulation for the backward-looking (reinforcement based learning) satisfaction balance model of Gray and Tallman (1984). This model is then used to generate interactional data from four simulated agents in a network interaction experiment. Because this initial model does not generate stable interaction structures in the network experiment, it is altered step by step in the direction of a forward-looking (agent with goals) model that has been shown to generate such stable interaction structures. The purpose of the modifications are to learn what aspects of the forward-looking model are needed to evolve a stable interaction structure, and to learn how these aspects may be incorporated into a model that remains essentially reinforcement based.

WHERE FORWARD-LOOKING AND BACKWARD-LOOKING MODELS MEET

Introduction

The issue is to develop a theory of choice behavior that will lead to stable interaction structures in complex non-negotiated exchange situations in which rewards and punishments may be used to influence the behavioral choices of others. We begin with two fundamentally different theories, each of which has important properties, but also important flaws. Our task is to develop a new model that incorporates aspects of each of these theories and overcomes their inherent problems.

We begin by distinguishing between the two theories or models of choice behaviors that we call backward-looking and forward-looking. But, because these terms have had various uses and understandings (cf. Macy 1990), we make explicit our meaning of these terms. For us, a pure backward-looking model of choice behavior is one in which choices are made completely on the basis of the reinforcements (and punishments) for past behavioral choices. Reinforcements are perceived stimuli that, for our purposes, result as a consequence of prior behavioral choices. These stimuli (perceptions) increase the probability of the behavioral choice. Similarly, punishments are defined as perceived stimuli that decrease the probability of the behavioral choice.[1]

[1] An alternative formulation is discussed below in which the reinforcements and punishments alter the cost/benefit ratios that are used to select to select behaviors.

For us, a pure forward-looking model of choice behavior is one in which the choices are made on the basis of the consequences of those choices bringing perceptions of the situation closer to (or further from) being in line with an internally held standard or goal (Burke 1997). A perception is an input of stimuli from the situation, including, but not limited to, those stimuli that, in the backward-looking model, we classify as rewards or punishment. The forward-looking model is a cybernetic model. If, for example, the internally held standard is one in which points or money should be accumulating at a certain rate through exchanges, then behavioral choices will be made that bring perceptions of points or money accumulating at that rate and not faster or slower. If it should turn out that achieving this match is impossible, the standard will slowly change to a point that a match between perception and standard is possible to attain.

These are the two types of models with which we shall deal in this paper. We here describe a third type of model, however, to clarify some points about the two in our focus. This third model might be described as an automaton. In this model, behavioral choices are made on the basis of an internal standard and there are no perceptions of the situation; in this sense, the model is "flying blind." Unlike the forward-looking and backward-looking models, the automaton does not adjust to the situation. The backward-looking model has no goal, but adjusts to the situation by "giving in" to the pattern of rewards and punishments.[2] The forward-looking model adjusts to the situation by making use of situ-

[2] It would be incorrect to say that the "goal" of the backward-looking model is to get rewards and avoid punishments, because rewards and punishments are defined by their consequences of increasing or decreasing the probability of the behavior that precipitated them. They are not rewards and punishment in the abstract, only in their consequences.

ational contingencies to bring about its own goal.[3] Thus, actions in the backward-looking model are based solely on perceptions. Actions in the automaton are based solely on the standard. Actions in the forward-looking model are based on the relationship between the perceptions and the standard.

In the backward-looking model, perception of a reward always acts as a reinforcer, thus increasing the probability of the behavior that led to the reward. In the forward-looking model, perception of a reward will act as a "reinforcer" only if the standard holds that perception as a goal or if receipt of the reward puts the agent closer to the goal that is held.[4]

Burke (1996) has used computer simulations to model the interaction of several agents in an exchange context in which each serves to reward and/or punish the others. His work has shown that when the model is backward-looking, stable, predictable interaction structures do not emerge. Forward-looking or goal-oriented models of "agents," however, when simulated can and do find stable, predictable interaction structures when such are possible. In this paper we identify several characteristics of backward-looking models that prevent the achievement of stable interaction structures. By altering these characteristics, we can move toward models that bridge the gap between forward- and backward-looking models. In this way we allow essentially backward-looking models to develop forward-looking features.

---

[3] The internal standard of the forward-looking model is a goal in the sense that it can be thought of as representing an abstraction of a "potential future state" that is realized when perceptions of the situation match that state.

[4] This latter would be the case, for example, if the agent held a standard indicating an average of 4 points per round need to be achieved and the agent had received an average of only 3.6 points per round thus far.

Background

Stable Interaction Structures

Before talking about the various models we need to be clear about our own criteria for evaluating them. We are looking for models of individuals or agents such that when several agents are put together in an exchange context, stable patterns of interaction/exchange emerge. There are several features of this. The first is *stability*; patterns of behavioral choice develop such that sequences of interaction can be identified that are stable and repetitive over time. The second is *conditionality*; the behavioral choices of each agent are conditioned by the choices of others.[5] The third is *situational sensitivity*; patterns of behavioral choice vary by the structural conditions of the experimental setup.

Conditionality is shown if behavioral choices of two actors are become synchronized. Such synchronization indicates the behaviors of one are contingent on the other (or, possibly, both are contingent on some external condition). In the present research we measure synchronization with an index of reciprocity, which is the proportion of groups in which one agent (A) establishes a mutual reward pattern with either of the two potential exchange partners (B or C), which patterns holds at least 95 percent of the time.[6] Stability is shown if the probability of particular behaviors remains constant over some period of time. If the probability of a behavior achieves 1.00, there is clear constancy. However, a stable sequence of behaviors would also manifest unchanging probabilities of each type

[5] In one sense the conditionality is definitionally built into the model because the probability of a choice, or the choice itself, is influenced by the behavior of others. However, at a more macro level, when averaged over a series of rounds, we may not see this conditionality manifest itself in correlations between the behavioral choices. It is the latter we are after as evidence of stable, structured interaction.

of behavior in the sequence.[7] In the present research we measure stability as the proportion of groups in which the average standard deviation of each of A's behavior probabilities is less than 0.01. Finally situational sensitivity is shown if the (experimental) context in which the of the interactions influences the patterns that emerge. In the present research we perform an analysis of variance on A's behavioral probabilities and examine the proportion of total variance that is accounted for by the eight experimental conditions (averaged across the four behaviors).

In addition to these three criteria for the emergence of stable interaction structures, we also examine the efficacy of the behavioral choices in terms of yielding desirable outcomes such as points or goal achievement.

Methods

We will be examining these several computer simulation models (theories) within the context of eight different experimental conditions initially studied Linda Molm (1989). These conditions (shown diagrammatically in Figure 1) involve four actors, each of whom can interact with two others in a square network configuration. Each actor can either reward or punish each of his or her potential partners, with the magnitude of the reward and punishment values given by the experimental condition. The eight conditions are divided across a 2x2x2 factorial design in which the reward power of A is high or

---

[6] Each of these measures is taken over the last 100 of the 500 rounds.

[7] The type of stability in which we are primarily interested in is one which is achieved as a balancing of the forces operating on the probabilities of the different types of behavior. Occasionally, the simulations show stability in an agent's behavior that is the result of *no* forces operating on that agent's probabilities for behavioral choice. The agent in question is the target of no action either rewarding or punishing. In the backward-looking models there is no mechanism to change these probabilities, so they remain at whatever level they had achieved earlier when the agent was a target of action.

low, the punishment power of A is high or low, and the direction of the reward and punishment power of A are the same or opposite. In all conditions, A can earn more from C than from B so would be better off exchanging rewards with C. B can earn more from A than from D so would be better off exchanging rewards with A. This creates an imbalance between the optimum strategies for agents A and B. How this gets resolved, of course, is the question. Finally, it is noted that agents A and C are in identical positions, as are agents B and D.

(Figure 1 About Here)

The interaction is arranged in a series of "rounds" during which each actor may either reward or punish either one or the other potential partner, but not both. Thus, each round involves choosing one act from the actor's repertoire. Reward involves giving points to another actor (from a pool maintained by the experimenter, not from the actor's profits). Punishment involves taking points away from another actor (though they do not go into the actor's own pool). Thus, each actor has four possible behaviors to choose from: 1) reward person, 2) reward other, 3) punish person, and 4) punish other.[8] Actors have no choice in the amount of reward or punishment they deliver. The amounts are fixed by the experimental design. Actors accumulate points based on the actions of their potential partners. There is no negotiation between the partners. Note that the consequences of an agent's actions on one round are not realized until the following round is completed, and cannot bear upon their choice until the round after that. Hence, there is always a delay in

[8] The terms person and other maintain consistency with Molm's (1989) design. Agents A and B are "persons" and agents C and D are "others," which, in the original experiments were computer simulations.

consequences. Two acts must be chosen before the consequences of the first act are known.

This arrangement provides a level of complexity that is important to understand. Each actor may be rewarded, punished, or ignored by each of two others on each round. All combinations are possible. If each actor chooses his or her behaviors randomly, then each actor is subject to random patterns of reward and punishment. Can such an arrangement yield stable structures? If an actor has a goal, how is that goal achieved through applying rewards and punishments to others?

In all our results, we run the simulation for five hundred rounds, and we calculate results averaged across the final 100 rounds.[9] Each of the eight "experimental" conditions are simulated 200 times (replications). For those models in which choice is based on probability functions, all of the agents begin with a 50/50 choice for both with whom they will interact and whether they will choose to reward or to punish. Generally, our focus will be on agent A as she interacts with agents B and C. We examine the proportions of each of the four acts that A might choose (reward B, punish B, reward C, or punish C). Also, we examine the structural outcomes of conditionality, stability, and situational sensitivity. Finally, we examine (across different models) the number of points on the average that A earns. Because of the different points awarded, it does not make sense to look at earnings across experimental conditions.

The measurement of conditionality, stability and situational sensitivity are measured over the last 100 of the 500 round exchange simulations. Conditionality is indicated by

---

[9] We found that the achievement of stability occurred only slowly in some structures. By selecting 500 rounds, most of the structures that would achieve stability did so.

the coordinated behavior. We examine the extent to which A and either B or C mutually and exclusively (over 95 percent of the time) reward each other; that is, the extent to which they form a stable positive exchange relationship. Stability is indicated by behavior probabilities that change very little over the last 100 rounds. Specifically we calculated the standard deviation of the behavior probabilities over these rounds and labeled as "stable" those with an average standard deviation less than 0.01 across the four behaviors. Finally, we assessed situational sensitivity using analysis of variance on the behavior probabilities. Our index is the average R-square across the four behaviors of person A.

Results

The matching law model

We begin with two baseline models: a very basic backward-looking reinforcement model such as that employed in the matching law (Estes 1957), and an identity theory based model (Burke 1997). We first consider a very simple model based on the matching law. In the matching law, the probability of choice A of a binary response is increased if the response is followed by a reward, and decreased if a punishment (or non-reward) follows it. The magnitude of the reward or punishment is not taken into account. This is expressed in equations 1 and 2

$$P_{n+1} = (1-\theta)\cdot P_n + \theta \qquad \text{if A is reinforced, and} \qquad \text{(Eq. 1)}$$

$$P_{n+1} = (1-\theta)\cdot P_n \qquad \text{if A is not reinforced,} \qquad \text{(Eq. 2)}$$

where P_n is the probability of the behavior A (rather than B) at time n, and θ is the instantaneous magnitude of impact (a constant, between 0 and 1, for a particular learning

situation and usually small, on the order of 0.1) that a reinforcement has on the probability of the behavior.[10] From this it can be seen that if choice A is reinforced with probability $p(R_A)$, in the long run choice A will be made with a probability, p(A), that matches the probability of reinforcement, $p(R_A)$. Note that the magnitudes of reward or punishments are not taken into account in this very simple model. We shall explore another backward-looking model later that does take magnitudes into account – the satisfaction-balance model (Gray and Tallman 1984).

Even with this simple model we must decide what external conditions will yield reinforcement. There are two ways to look at this question. From the experimenter's point of view, a decision is made about the conditions that will lead to the presentation of a predefined reward (i.e., a stimulus that will act as a reinforcer). From the agent's point of view, the question is what is perceived as a reward. For the basic reinforcement model as applied to the conditions of interest, a reward for the agent could be one of the following binary outcomes: (1) getting points as opposed to not getting points (positive reinforcement), (2) not having points taken away as opposed to having points taken away (negative reinforcement), (3) getting more points than on the previous round (an increase in positive reinforcement), (4) getting more points than another agent, or (5) not getting fewer points than another agent.[11] A reward could also be seen as one outcome of a three-category set of reward, neutral, and punishment. For the neutral outcome, the probability

---

[10] Technically, θ is the fraction of stimuli conditioned as a result of the reinforcement.

[11] This is not a question of what goal an agent will choose. In the context of the matching law, it is a question of which stimulus has the consequence of increasing the probability of the response that preceeded it. In any agent model this is a fixed condition. If we were observing agents, we would have to discover which stimulus has the reinforcing consequences. As modelers of agents we simply choose our condition and test its consequences.

of the response is neither increased (as when followed by a reward) nor decreased (as when followed by a punishment).

We constructed simulations based on each of these definitions of what is reinforcing. Analysis of each of these models produces similar results. We present a model based on number two above, defining a reinforcing event as one in which the agent does not loose points (which we call "avoiding punishment"). Table 1 shows results from an analysis of this baseline model. We see that the probabilities occasionally depart only slightly from their starting values of 0.25. The index of situational sensitivity is only 0.01. Indeed there is more variability in the probabilities within individual groups (across the last 100 rounds) than there is for the mean probabilities of the 200 replications. With respect to the remaining indicators of structure, there is almost no stability in the behavioral probabilities over the last 100 trials. There is also no conditionality as indicated by systematic linking of the behaviors of A and either B or C. Clearly, this model does ot lead to structured interaction as defined here. Figure 2 shows an example group over the complete 500 rounds.

(Table 1 and Figure 2 About Here)

The identity model

The identity based model has an internal standard which is initially set at earning points from the partner who can provide the most points. Each alternative behavior is tried in turn for several rounds to yield that outcome.[12] If, after trying all of the behaviors,

[12] The number of rounds tried for each behavior is a randomly set anew in each round at 1, 2, or 3. Thus, the agent may try one behavior for two rounds, then switch to the next for a single round, then switch to the next for three rounds, etc.

the outcome is not obtained, the standard switches to earning points from the other potential partner. Again, if that outcome is not obtained the agent switches back to the first standard. If the agent does succeed in obtaining its goal (perceptions match standard of earning points from the target agent), the agent acts to reward the target of her actions.

Results from an analysis of this model, as shown in Table 2, contrast quite strongly with the results from the model based on the matching law. The probabilities for punishing either agent B or agent C are reduced to close to zero. The probabilities for rewarding B are either close to .20 or zero, while the probabilities for rewarding C are close to .80 or 1.00 depending upon the experimental condition. The sensitivity index for this model is 0.71. Stability of the behavioral probabilities across the last 100 rounds within a group is quite high. Overall, 59 percent of the groups are stable, and in four conditions 100 percent of the groups are stable. Finally, the index of reciprocity shows that A has formed a reciprocal exchange relationship with another in all groups in four of the conditions, but in no groups in the other four conditions. Thus, most of the groups developed some degree of structure using this model of the agent. Figure 3 shows an example group over the 500 rounds (though, recall that data for this group is based on the last 100 of these rounds).

(Table 2 and Figure 3 About Here)

Why is the matching law model as implemented here unable to settle into any stable pattern while the identity model settles very quickly into such a pattern? First, we suggest that the learning parameter, theta, which measures the impact of any reinforcement or punishment on the choice probabilities, is too small. Thus, the choice probabilities hover around their initial values in a kind of Brownian motion, unable to move toward any certainty upon which patterns can be built. Second, we suggest that the matching law model

provides no patterning in the reinforcements that any individual receives from others. Each agent's behavior is randomly rewarded and punished. Hence, each agent's behavior becomes a random variable influencing other's behavior randomly; a cycle that remains unbroken.

A third potential reason for the difference, and related to the second reason, is that too much is going on in the matching law model. In a sense, each agent is bombarded from all sides with rewards and punishments. Most experimental tests of the matching law are designed with fixed stimuli sources and an overall reinforcement schedule designed by a single experimenter. When taken into an open social context, such as employed in the present situation, this special character is lost. In the present situation, it is as if there are three different experimenters for each agent. In the identity model, the standard keeps the agent focussed on a particular target and seeking a particular outcome. All else is ignored as irrelevant.

Finally, in the matching law model, no agent has a chance to experiment with what has been learned in prior rounds before being hit by additional rewards or punishments. Again, this is related to the second and third points above. A behavior selected in one round has a consequence in the next round only after another behavior is chosen for that round. It is not until the third round that the learning can be put to use. In the identity model, each agent tries a behavior several times before deciding that it is or is not working.

Increase the learning parameter

Our first change, therefore, will be to increase the learning parameter, theta, from 0.1 to 0.5. As can be seen in the results of the analysis of this model presented in Table 3, quite a bit of structuring is now apparent. Compared to the earlier results in Table 1, the amount of punishment given has dropped to an average of 0.06. However, situational sensitivity is a modest 0.08, primarily because the magnitudes of rewards and punishments are not taken into account, and it is these that distinguish between conditions. Stability, on the other hand, has increased a lot, now showing an average of 90 percent of the groups as stable. Finally, the index of reciprocity shows that 43 percent of the groups had stable exchange relations between A and another agent.

(Table 3 and Figure 4 About Here)

From the example group illustrated in Figure 4, we see that the instability of the choice probabilities we saw for the avoid punishment model with a theta of .1 is much more pronounced with the larger theta. On the other hand, we see that after about 200 trials, the system seemed to "lock-in" to a stable pattern in the manner that the identity model locked in.

Clearly, our changing the learning parameter has had a significant impact on the Avoid Punishment model, allowing the choice probabilities to move away from their 50/50 mode toward a value of 1.00 where they could stabilize and provide a patterned reinforcement for others in the group. This moves us closer to the kind of results that were obtained with the identity model, though there is still a lot of randomness remaining.

Incorporating magnitudes of reward and punishment

One problem that we noted with the Avoid Punishment model was the fact that it was too simple. The differential distribution of reward and punishment magnitudes across the experimental conditions (that defines the conditions) was not recognized. The satisfaction-balance model is an improvement on the avoid punishment model in part because it takes into account the magnitudes of reward and punishment. In this way, each experimental condition, with its unique pattern of rewards and punishments should yield outcomes that are reflective of these patterns.

The satisfaction-balance model proposed by Gray and Tallman (1984) is a reinforcement based learning model that takes the magnitude of reward (and punishment) into account. It states (in its simplest form) that the probability of choice A (versus B, in two choice situations) is a function of magnitudes and probabilities of the rewards and punishments associated with those choices according to the formula

$$p(A) = \frac{V(A)^{1/2} \cdot C(B)^{1/2}}{V(A)^{1/2} \cdot C(B)^{1/2} + V(B)^{1/2} \cdot C(A)^{1/2}}, \qquad \text{(Eq. 3)}$$

where V(A) is the value of choice A, V(B) is the value of choice B, C(A) is the cost of choice A, and C(B) is the cost of choice B. The value of a choice, V(x), is a product of the likelihood of x being correct and the magnitude of the reward for choosing x, while the cost of a choice C(x) is a product of the likelihood of x being incorrect and the magnitude of the punishment for choosing x. Thus,

$$V(x) = pr(R_x) \cdot M(R_x), \text{ and} \qquad \text{(Eq. 4)}$$

$$C(x) = \left(1 - pr(R_x)\right) \cdot M(P_x) \qquad \text{(Eq. 5)}$$

This formulation only gives the long run expected outcome, rather than the instantaneous change in probability as the results of a prior reinforcement or punishment. The instantaneous formulation (for disjunctive outcomes) may be given as

$$P_{n+1} = (1-\theta)\cdot P_n + \theta \qquad \text{if A is reinforced, and} \qquad \text{(Eq. 6)}$$

$$P_{n+1} = (1-\eta)\cdot P_n \qquad \text{if A is not reinforced,} \qquad \text{(Eq. 7)}$$

where $\theta = k_1\cdot(M(R_A)\cdot M(P_B))^{1/2}$ is the scaled magnitude of the impact when the choice (A) is rewarded or correct (k_1 is the scaling constant), and $\eta = k_2\cdot(M(P_A)\cdot M(R_B))^{1/2}$ is the scaled magnitude of the impact when the choice is punished or incorrect.[13] As can be seen, if the magnitudes of the reward and punishment are equal, this simplifies to the matching law. Note that the impact coefficient in the satisfaction-balance model is a variable function of the magnitude of reward and/or punishment rather than a constant as in the matching law.

Results of the analysis of the satisfaction balance model are presented in Table 4. With the additional taking into account of the magnitudes of rewards and punishment in the satisfaction balance model, the proportion of punishment given has dropped virtually to zero. There is also now a tendency to reward agent C more than agent B. Situational sensitivity has almost doubled to 0.15, and stable reciprocal exchanges have jumped from 43 to 63 percent ($t = 11.6$, $p \leq .01$, though stability has dropped from 0.90 to 0.88 ($t = 2.14$, $p \leq .05$). In addition, the number of rounds until the groups achieve stability has dropped. This is illustrated in Figure 5, which presents the results for an example group.

---

[13] The scaling constants for any real organism in a given situation would have to be empirically determined. For the present simulation we have chosen a value that results in average values of $\theta = 0.5$ across the range of rewards and punishments.

(Table 4 and Figure 5 About Here)

Attention Focus

If part of the problem is that there are too many conflicting stimuli (rewards and punishments) impinging on the actor for any pattern to emerge, one way in which the number of stimuli may be reduced, and perhaps patterned, is for the actor to attend only to one other agent as a source of rewards of punishments. By this we mean that rewards and punishments coming from any agent other than the one being attended to are ignored; they do not change the choice probabilities. This addition moves us more in the direction of the forward-looking model since it relies on an internal structure, in this case knowledge of whom to attend. With this model, however, comes the question to which other should the actor attend? We have modeled several possibilities.

For example, the actor may attend only to the other from whom he or she has received the most rewards in the past (knowing which side your bread is buttered on). Or, the actor may attend only to the other toward whom the last act was directed (paying attention to the consequences of your own actions). A third possibility is that the actor may attend only to the other that has the highest reward potential in the given experimental setting (kissing up to the powerful). In all of these, the act is chosen according to the probabilities at the point of choice. The probabilities change as these acts are rewarded, punished, or ignored by the other to whom the actor attends. In the first and second, the other to whom the actor attends can change over time, in the third, the other attended to remains constant.

Analysis shows that the proportion of groups that achieve stability differs considerably between attending to a powerful person (the person who has been most rewarding, or the person who has the most rewards to give) and attending to the target of one's last act.[14] In the models attending to others with power or who have rewarded most, only about 50% of the groups achieve stability by the 400th round, and stable reciprocal exchange relations developed in only about 35 percent of the groups. Further, situational sensitivity is between 0.03 and 0.06. In the model based on attending to the target of the last act, 80% of the groups achieved stability, and 82 percent have achieved stable exchange relationships. Additionally, situational sensitivity is about 0.12. Table 5 presents the results of the analysis of this last model, and Figure 6 presents the results for an example group over the 500 rounds of exchange. The proportion of punishments is generally lower than the avoid punishment ($\theta = 0.5$) model, but not as low as in the basic satisfaction balance model (with no special attention paid to any particular other). Overall stability has decreased somewhat, but stable exchange relationships have increased as a result of attending to the target of one's last act. And, as a result, there is much more rewarding of agent C than is the case in either the avoid punishment or satisfaction balance models.

(Table 5 and Figure 6 About Here)

An alternative formulation of the satisfaction-balance model.

Up to this point we have formulated the satisfaction balance model in such a way that it's relationship to the classical matching law was apparent. In this formulation reinforcements and punishments modify the probabilities of each action. An alternative formula-

[14] Data is available on request from the first author.

tion of the satisfaction-balance model suggests that behavior is chosen, not probabilistically, but on the basis of a cost/value assessment (Gray and Tallman 1996). In this model, a behavior is chosen if it has the smallest cost to value ratio, with the choice being random among those that have an equal cost to value ratio. Costs and values are measured as indicated earlier in terms of the number of points lost or gained, but a behavior is always chosen if it has the lowest cost to value ratio.[15] This results in much more stability over time in the choices of an agent as the agent continues to choose a behavior as long as it has the lowest relative cost. In the present implementation, we have kept the quasi-standard of attending to the target of the last act.

The result of this model is a large increase in the proportion of groups that form stable exchange relationships from 0.63 to 0.81 ($t = 11.49$, $p <= .01$). However, there is a small drop in the proportion of groups that have stable behaviors over the last 100 rounds from 0.88 to 0.83 ($t = 3.53$, $p <= .01$). Finally, situational sensitivity is only .02. These results are presented in Table 6, and an example group is given in Figure7.

(Table 6 and Figure 7 about here)

Adding reinforcement to the identity model

Thus far, we have been modifying the reinforcement-based models to bring their features and operational outcomes closer to those achieved in the identity model. This has been quite successful in achieving stable exchange relations in the four-person situations studied. Indeed, in the cost/value ratio based satisfaction balance model, we have achieved a

[15] The cost of a behavior is the number of points taken away, the value of a behavior is the number of points gained. Each behavior also has a constant cost of one point each time that it is selected. The cost/value ratio

greater and more stable structuring across the eight experimental conditions than was achieved in the forward-looking identity model. The identity model, as simulated here, has no learning ability. That is, it has no way to restructure itself based on outcomes it has experienced. Basically, the identity model tries each possible behavior in turn for several rounds, until it finds the behavior that accomplishes the goal of matching perceptions to standard, and then acts to maintain that perception. It does not learn from what was effective in the past. If perceptions change such that they no longer match the standard, the agent goes through all the behaviors again.

In a modification of the identity model, we create an agent that chooses a behavior according to a probability function when its perceptions do not match the standard. Behaviors that do not work to bring the perception into line with the standard have their probabilities reduced. When the perceptions match the standard, the agent no longer selects behaviors probabilistically, but simply continues to act to maintain the perception by rewarding the target.

In this model, the standard is to earn points from the potential exchange partner who can provide the largest reward. For any outcome that does not match the standard, the probability of the prior behavior is reduced as in the satisfaction balance model. Occasionally, the target does provide a reward (an apparent success), but because of punishment by the other potential exchange partner the points earned are less than *could* be earned from that other partner. This occurrence is simply taken as an outcome that does

---

is the total cumulated cost of the behavior divided by the total cumulated value of the behavior weighted by the number of times the behavior has been tried.

not match the standard, and the probability of the behavior is reduced accordingly and the next behavior is chosen on the basis of the modified probabilities.

Results from this model are given in Table 7, with an example group in Figures 8. We see that the proportion of stable cases increases dramatically from 0.59 to 0.97 compared with the prior version of this forward-looking model. We also see an increase in the degree to which actor A rewards actor C to virtually 100 percent. In the earlier identity model, four of the conditions had rates of rewarding that were close to 100%, and generally less than twenty percent in the other four conditions. In the conditions with the lower rates, a cycle developed in which agent A rewarded agent B about every eighth turn. This was in response to a punishment from B. This stable cycle remained because agent B, without any learning, had to cycle through her entire behavioral repertoire seeking to get points from A. This was successful only about every eighth round. No learning took place and the entire repertoire was gone through each time in approximately the same way (the number of time each behavior was tried would vary). This made it difficult for stable exchange relationships to emerge. In the identity learning model, this cycle was broken, and agents learned which behaviors tended to work and which not. Time was not wasted on behaviors that had a low chance of success.

(Table 7 and Figure 8 About Here)

Discussion

There are a number of differences between backward-looking models (such as in reinforcement based learning theory or exchange theory) and forward-looking cybernetic models (such as in identity theory), and yet, at some level there should be a convergence

between the two models. The purpose of the present paper has been to explore the direction of that convergence. In doing that, a number of important points have emerged which should help us better understand the nature of agency, learning, and interaction between agents.

The simplest of the backward-looking agent models failed to form stable interaction structures. While we had an initial expectation that the patterning of rewards and punishments would be carried into a patterning of behavior, such did not happen under the present experimental conditions.[16] It appeared that the patterning of rewards and punishments given to any agent was itself so random that each agent continued to randomly select behaviors that acted as conditioners for others' behaviors.

Three different conditions contributed to the perpetuated chaos. First, each reward or punishment had only a small impact that was cancelled out by the impacts of other rewards or punishments before any patterns could emerge. A kind of Brownian motion ensued. Second, each agent received independent and perhaps conflicting rewards and punishments from two other agents. Again, there is a tendency for these effects to cancel each other out and prevent any patterned interaction from developing.

Third, each agent selected a behavior probabilistically *on each round*. Hence, a new behavior was selected before the agent could learn the consequence of the last behavior, and behavior was selected randomly, even after the consequences could have been learned.

---

[16] Nor did it happen under a variety of additional experimental conditions that were tried with varying reward and punishment levels, including models that did not include punishment. None of the conditions, however, reflected stable circumstances for the agent to learn.

The forward-looking model had none of these deficits. The presence of an identity standard by which changes in perception could be assessed made the difference. Behaviors were not chosen probabilistically, thus the basis of a choice could change dramatically and with certainty at any time. The agent attended only to the perceptual signal relevant to the identity standard; thus the impact of other "rewards and punishments" had no effect. And, the agent repeated a particular behavior long enough to learn the effectiveness of that behavior for achieving its goal.

To understand the impact of each of these differences, a number of models were explored which allowed us to assess each of these differences. The results of these analyses are summarized in Table 8. In moving from the initial backward-looking model that provided little basis for the emergence of a stable interaction structure toward the forward-looking model that provided considerable basis, several features were added. First, the impact of reinforcement (the theta parameter) was increased to the extent that large changes in the probabilities of alternative behaviors could occur. Without the probabilities of some behaviors approaching 1.0 agents could not form stable exchange structures.

(Table 8 About Here)

The second feature that was added was an internal standard or goal in the form of a target person to whom the agent attends for defining rewards and punishments. In the present models this was done by fiat in the sense that this was simply programmed in and did not emerge from experience. This latter step (having the program learn the standard and how to behave with respect to it) is the direction of future research. Note however, that the selection of the target to whom the agent should attend was not fixed (for the models presented), but was simply the person to whom the agent directed her last act.

The third feature that increased the ability of the agents to form stable exchange relationships was to move the basis of selection of behaviors away from probabilistic choice patterns to value based choice patterns. While this did not significantly increase either the measured stability of the behaviors or the proportion of stable exchange relationships formed by agent A, there was another change that was dramatic. This can be seen in Table 9, which presents additional summary results that include agent B as well as A.

(Table 9 About Here)

We have been considering the stability of agent A's behavior, and the degree to which agent A formed stable exchange structures with either B or C. If we also look at what is happening with agent B, another picture emerges. The stability of the forward-looking identity model was achieved by only looking at agent A. Agent B never formed stable exchange relations with either A or D. When we moved to the satisfaction balance model (probability based) B began to form some stable exchange relationships, though these were diminished when the internal standard of attending to the target of the last act was included in the model. This occurred because it increased the chances that A would form a stable exchange relationship with C. Moving to the value based decision form of the satisfaction balance model not only increased the consistency of A's behavior (as well as the behavior of the other agents), but allowed B to form stable exchange relationships. Overall, therefore, this model provides the most stable social interaction structure of all those studied.

Also shown in Table 9 are the earnings of agents A and B over the last 100 rounds of the exchange. The greatest earnings for A occur in the identity model with learning, because A forms a stable exchange relationship with C almost all the time, and this is the

most profitable relationship that can occur. On the other hand, B earns almost nothing. In the value based satisfaction balance model, because she does form stable exchange relationships, B can earn almost as much as A on the average.

In summary, we have found three features of the forward-looking identity model that, when incorporated into a backward-looking (reinforcement based) model of an agent allow it to form stable exchange relationships with other agents in a network exchange situation. The impact of the reinforcement must be large, the agent must have an internal structure that is used as a standard with which to compare current perceptions of outcomes (in the present case, the models attended to the target of their last act, all other outcomes were ignored), and the agent must continue to choose the behavior that works best (as defined by prior experience).

The first and last of these features are easily made part of a backward-looking model. The second, the presence of a standard, is what distinguishes a forward-looking from backward-looking model. The question to which we will address ourselves in the future is how the backward-looking model can “learn” to have a standard on the basis of its experience in interaction. The problem is to create a model that has the capacity to turn a pattern of perceptions into a standard with which to compare new perceptions. Furthermore, for the standard function as the one use in the present simulations (in attending to the target of one’s last act), some rudimentary notion of “self” must be included so that the agent can distinguish perceptions of its own acts from acts produced by others.

References

Burke, Peter J. 1996. "Agency and Interaction." Social Interaction Processes Session of the American Sociological Association Meetings. New York, August, 1996.

_____. 1997. "An Identity Model for Network Exchange." *American Sociological Review* 62:134-50.

Estes, William K. 1957. "Of Models and Men." *American Psychologist* 12:609-17.

Gray, Louis N. and Irving Tallman. 1984. "A Satisfaction Balance Model of Decision Making and Choice Behavior." *Social Psychology Quarterly* 47:146-59.

_____. 1996. "Cost Equalization as a Determinant of Behavioral Allocation: The Case of Binary Choice." Social Psychology Quarterly 59:154-61.

Macy, Michael W. 1990. "Learning Theory and the Logic of Critical Mass." *American Sociological Review* 55:809-26.

Molm, Linda D. 1989. "Punishment Power: A Balancing Process in Power-Dependence Relations." *American Journal of Sociology* 94:1392-418.

Table 1. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Avoid Punishment Model, Theta = 0.1.

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.25	(0.07)+	0.25	(0.06)	0.25	(0.06)	0.25	(0.06)	0.00	0.00
RH.PL.op	0.25	(0.05)	0.25	(0.05)	0.25	(0.05)	0.25	(0.05)	0.00	0.00
RL.PH.op	0.25	(0.07)	0.25	(0.05)	0.25	(0.05)	0.26	(0.05)	0.00	0.00
RL.PL.op	0.26	(0.06)	0.25	(0.06)	0.25	(0.06)	0.24	(0.05)	0.00	0.00
RH.PH.sa	0.25	(0.05)	0.26	(0.05)	0.24	(0.05)	0.25	(0.05)	0.00	0.00
RH.PL.sa	0.25	(0.05)	0.25	(0.05)	0.25	(0.05)	0.25	(0.05)	0.00	0.00
RL.PH.sa	0.26	(0.05)	0.25	(0.05)	0.25	(0.05)	0.24	(0.05)	0.00	0.00
RL.PL.sa	0.25	(0.04)	0.25	(0.04)	0.25	(0.05)	0.25	(0.04)	0.00	0.00
Total	0.25	(0.06)	0.25	(0.05)	0.25	(0.05)	0.25	(0.05)	0.00	0.00

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is about 0.08.

Table 2. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Identity Model.

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.17	(0.03)+	0.00	(0.01)	0.82	(0.04)	0.02	(0.01)	0.21	0.00
RH.PL.op	0.17	(0.03)	0.01	(0.01)	0.80	(0.04)	0.02	(0.01)	0.09	0.00
RL.PH.op	0.17	(0.03)	0.01	(0.01)	0.81	(0.04)	0.02	(0.01)	0.09	0.00
RL.PL.op	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
RH.PH.sa	0.22	(0.03)	0.00	(0.01)	0.77	(0.03)	0.01	(0.01)	0.35	0.00
RH.PL.sa	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
RL.PH.sa	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
RL.PL.sa	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
Total	0.09	(0.10)	0.00	(0.01)	0.90	(0.11)	0.01	(0.01)	0.59	0.50

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is 0.00

Table 3. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Avoid Punishment Model, Theta = 0.5.

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.42	(0.49)+	0.12	(0.31)	0.27	(0.44)	0.19	(0.38)	0.94	0.32
RH.PL.op	0.31	(0.46)	0.16	(0.36)	0.46	(0.50)	0.08	(0.26)	0.96	0.35
RL.PH.op	0.50	(0.49)	0.01	(0.04)	0.36	(0.47)	0.12	(0.32)	0.94	0.38
RL.PL.op	0.47	(0.49)	0.10	(0.30)	0.42	(0.49)	0.01	(0.05)	0.95	0.50
RH.PH.sa	0.40	(0.46)	0.03	(0.09)	0.54	(0.47)	0.03	(0.08)	0.85	0.49
RH.PL.sa	0.47	(0.48)	0.02	(0.07)	0.49	(0.48)	0.02	(0.07)	0.87	0.49
RL.PH.sa	0.44	(0.47)	0.03	(0.08)	0.50	(0.48)	0.03	(0.07)	0.85	0.50
RL.PL.sa	0.46	(0.47)	0.03	(0.08)	0.48	(0.48)	0.03	(0.08)	0.84	0.42
Total	0.43	(0.48)	0.06	(0.21)	0.44	(0.48)	0.06	(0.21)	0.90	0.43

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is 0.02.

Table 4. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Satisfaction Balance Model, Theta = 0.5.

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.56	(0.42)+	0.05	(0.09)	0.35	(0.40)	0.04	(0.09)	0.67	0.35
RH.PL.op	0.34	(0.42)	0.00	(0.02)	0.65	(0.42)	0.00	(0.02)	0.80	0.56
RL.PH.op	0.69	(0.44)	0.00	(0.02)	0.30	(0.43)	0.00	(0.02)	0.89	0.61
RL.PL.op	0.51	(0.48)	0.00	(0.00)	0.49	(0.48)	0.00	(0.00)	0.95	0.74
RH.PH.sa	0.30	(0.44)	0.00	(0.02)	0.69	(0.45)	0.00	(0.02)	0.90	0.73
RH.PL.sa	0.32	(0.45)	0.00	(0.00)	0.68	(0.45)	0.00	(0.00)	0.93	0.71
RL.PH.sa	0.24	(0.41)	0.00	(0.01)	0.76	(0.42)	0.00	(0.01)	0.92	0.68
RL.PL.sa	0.34	(0.47)	0.00	(0.01)	0.66	(0.47)	0.00	(0.01)	0.96	0.68
Total	0.41	(0.47)	0.01	(0.04)	0.57	(0.47)	0.01	(0.04)	0.88	0.63

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is less than 0.01.

Table 5. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Attend Target SB Model, Theta = 0.5.

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.00	(0.02)+	0.00	(0.02)	0.92	(0.22)	0.07	(0.20)	0.88	0.87
RH.PL.op	0.03	(0.10)	0.02	(0.05)	0.80	(0.29)	0.15	(0.23)	0.71	0.65
RL.PH.op	0.09	(0.28)	0.01	(0.03)	0.88	(0.31)	0.02	(0.10)	0.93	0.93
RL.PL.op	0.34	(0.47)	0.01	(0.04)	0.64	(0.47)	0.01	(0.06)	0.94	0.95
RH.PH.sa	0.05	(0.08)	0.06	(0.10)	0.79	(0.32)	0.10	(0.17)	0.71	0.65
RH.PL.sa	0.14	(0.13)	0.17	(0.16)	0.54	(0.39)	0.15	(0.15)	0.47	0.40
RL.PH.sa	0.23	(0.39)	0.04	(0.10)	0.68	(0.43)	0.04	(0.11)	0.81	0.81
RL.PL.sa	0.59	(0.46)	0.06	(0.15)	0.33	(0.45)	0.02	(0.06)	0.82	0.83
Total	0.24	(0.40)	0.04	(0.11)	0.68	(0.44)	0.05	(0.13)	0.83	0.82

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is about 0.02.

Table 6. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Satisfaction Balance Model Based on Cost/Value Ratio with Attention to Target of Last Act, Theta = 1.0.

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.51	(0.49)+	0.05	(0.15)	0.34	(0.45)	0.09	(0.22)	0.79	0.75
RH.PL.op	0.53	(0.48)	0.05	(0.17)	0.38	(0.47)	0.04	(0.13)	0.84	0.82
RL.PH.op	0.31	(0.45)	0.01	(0.05)	0.57	(0.48)	0.11	(0.27)	0.83	0.80
RL.PL.op	0.30	(0.45)	0.02	(0.11)	0.60	(0.48)	0.08	(0.22)	0.87	0.84
RH.PH.sa	0.45	(0.47)	0.09	(0.21)	0.40	(0.48)	0.05	(0.15)	0.79	0.78
RH.PL.sa	0.51	(0.48)	0.06	(0.19)	0.38	(0.48)	0.04	(0.15)	0.82	0.82
RL.PH.sa	0.26	(0.43)	0.02	(0.05)	0.58	(0.48)	0.14	(0.29)	0.83	0.80
RL.PL.sa	0.30	(0.45)	0.02	(0.10)	0.59	(0.49)	0.09	(0.25)	0.90	0.87
Total	0.40	(0.47)	0.04	(0.14)	0.48	(0.48)	0.08	(0.22)	0.83	0.81

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is less than 0.01

Table 7. Proportions (and sd) of Each Type of Behavior By Experimental Condition, Identity Model with Learning

Condition*	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
RH.PH.op	0.01	(0.03)+	0.00	(0.00)	0.98	(0.04)	0.01	(0.02)	0.93	0.83
RH.PL.op	0.01	(0.03)	0.00	(0.00)	0.99	(0.03)	0.00	(0.01)	0.95	0.94
RL.PH.op	0.01	(0.03)	0.00	(0.00)	0.99	(0.03)	0.00	(0.00)	0.95	0.92
RL.PL.op	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
RH.PH.sa	0.01	(0.04)	0.00	(0.01)	0.99	(0.05)	0.00	(0.01)	0.94	0.88
RH.PL.sa	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
RL.PH.sa	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
RL.PL.sa	0.00	(0.00)	0.00	(0.00)	1.00	(0.00)	0.00	(0.00)	1.00	1.00
Total	0.00	(0.02)	0.00	(0.00)	0.99	(0.03)	0.00	(0.01)	0.97	0.94

* RH/L = High/Low Reward Power, PH/L = High/Low Punishment Power, op/sa = Direction of Reward and Punishment Power is opposite/same

+ Average sd for individual groups is less than 0.00

Table 8. Summary of Results

Model	Behavior								Stable	Recip.
	Reward B		Punish B		Reward C		Punish C			
Avoid Pun.	0.25	(0.06)	0.25	(0.05)	0.25	(0.05)	0.25	(0.05)	0.00	0.00
Identity	0.09	(0.10)	0.00	(0.01)	0.90	(0.11)	0.01	(0.01)	0.59	0.50
Avoid Pun. 2	0.43	(0.48)	0.06	(0.21)	0.44	(0.48)	0.06	(0.21)	0.90	0.43
SB	0.41	(0.47)	0.01	(0.04)	0.57	(0.47)	0.01	(0.04)	0.88	0.63
SB attend	0.24	(0.40)	0.04	(0.11)	0.68	(0.44)	0.05	(0.13)	0.83	0.82
SB CV attend	0.40	(0.47)	0.04	(0.14)	0.48	(0.48)	0.08	(0.22)	0.83	0.81
Identity Learn	0.00	(0.02)	0.00	(0.00)	0.99	(0.03)	0.00	(0.01)	0.97	0.94

Table 9. Additional Results

Model	Proportion of Stable Exchange Relationships			Earnings Over Last 100 Rounds	
	A-C Exchange	A-B Exchange	B-D Exchange	A earnings	B earnings
Avoid Pun.	0.00	0.00	0.00	2.88	2.64
Identity	0.50	0.00	0.00	712.85	116.87
Avoid Pun. 2	0.25	0.18	0.13	386.53	360.01
SB	0.40	0.23	0.15	545.30	394.66
SB attend	0.62	0.14	0.01	444.71	100.96
SB CV attend	0.44	0.37	0.44	443.23	420.79
Identity Learn	0.94	0.00	0.00	945.22	13.04

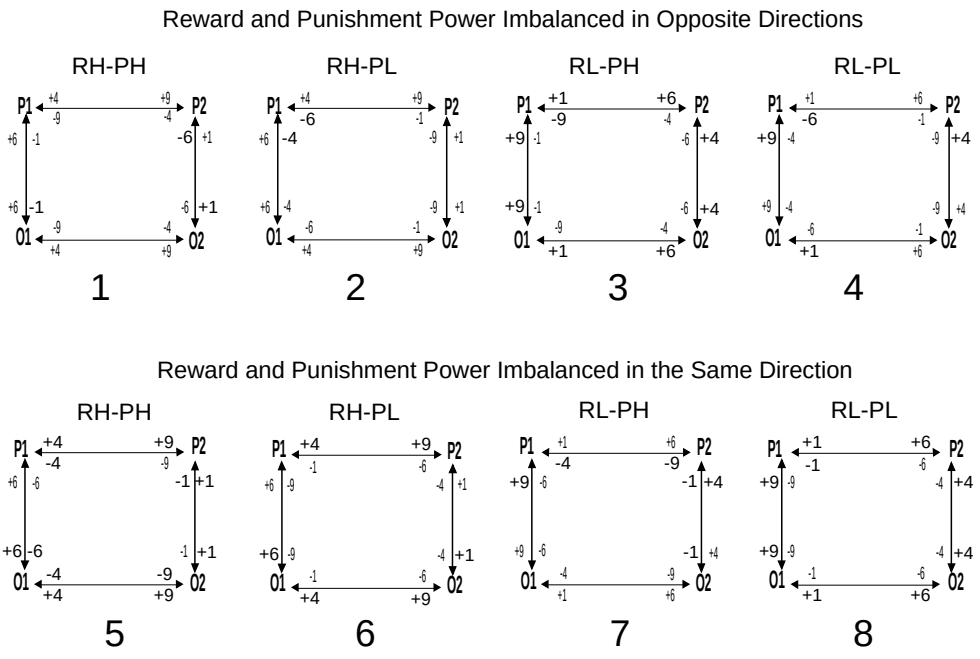

Figure 1. The eight power structure conditions.

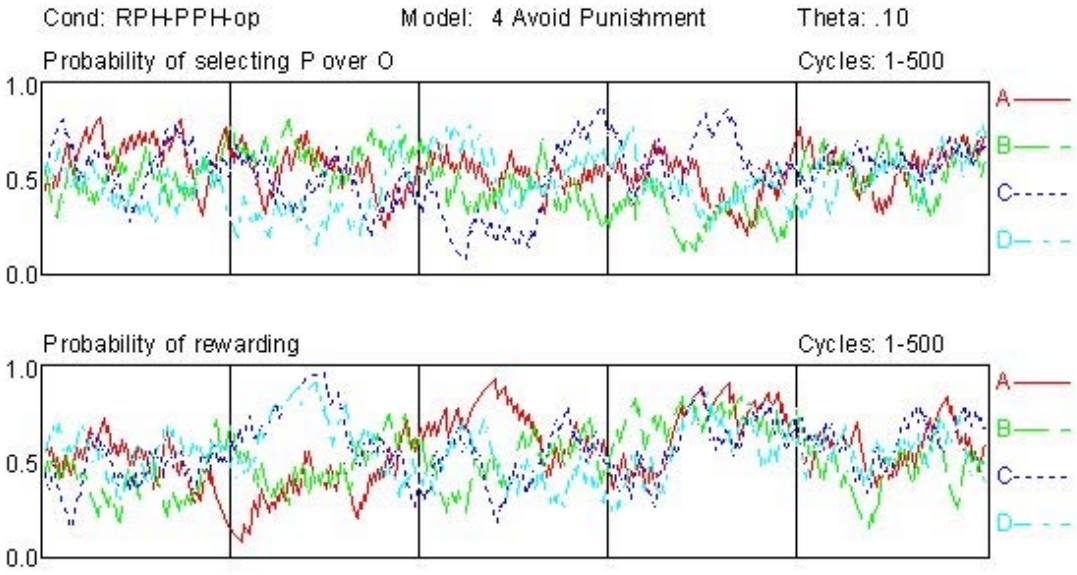

Figure 2. Example Group From the Avoid Punishment Model with theta = 0.10

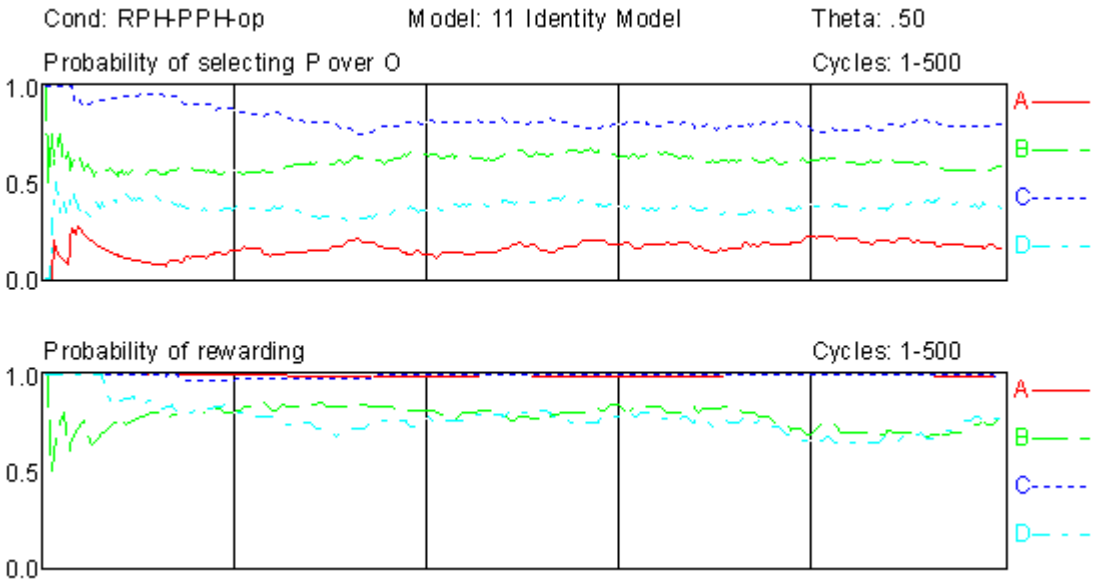

Figure 3. Example Group From the Identity Model

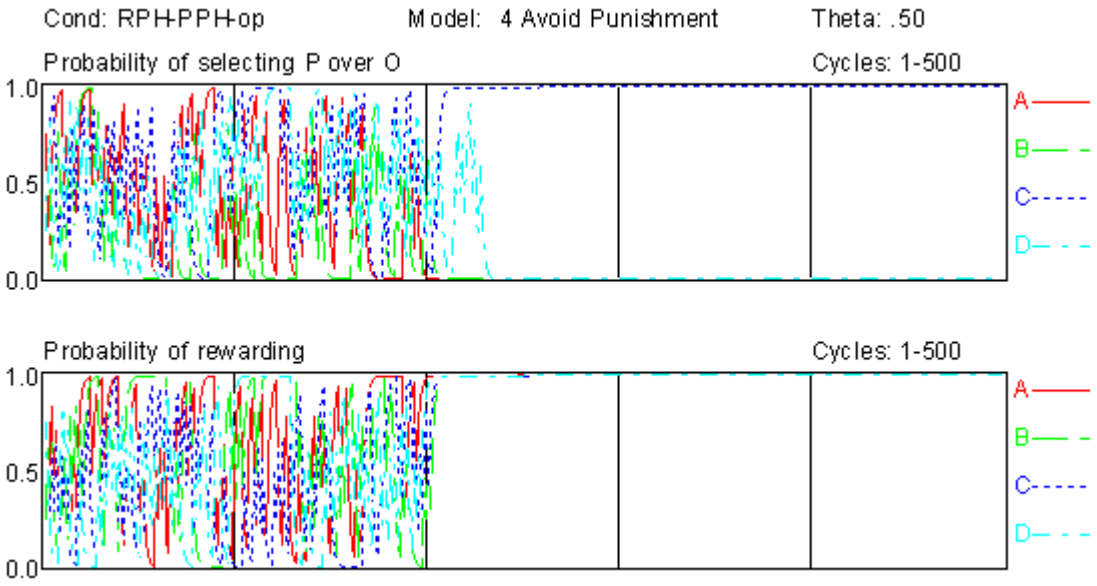

Figure 4. Example Group From Avoid Punishment Model with Theta = 0.5.

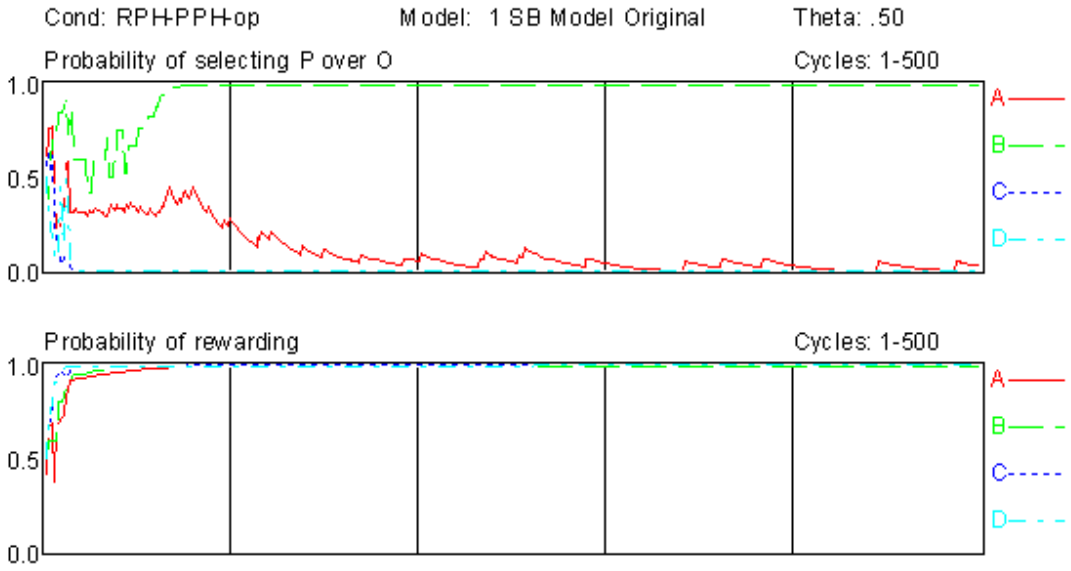

Figure 5. Example Group From Satisfaction Balance Model with theta = 0.50

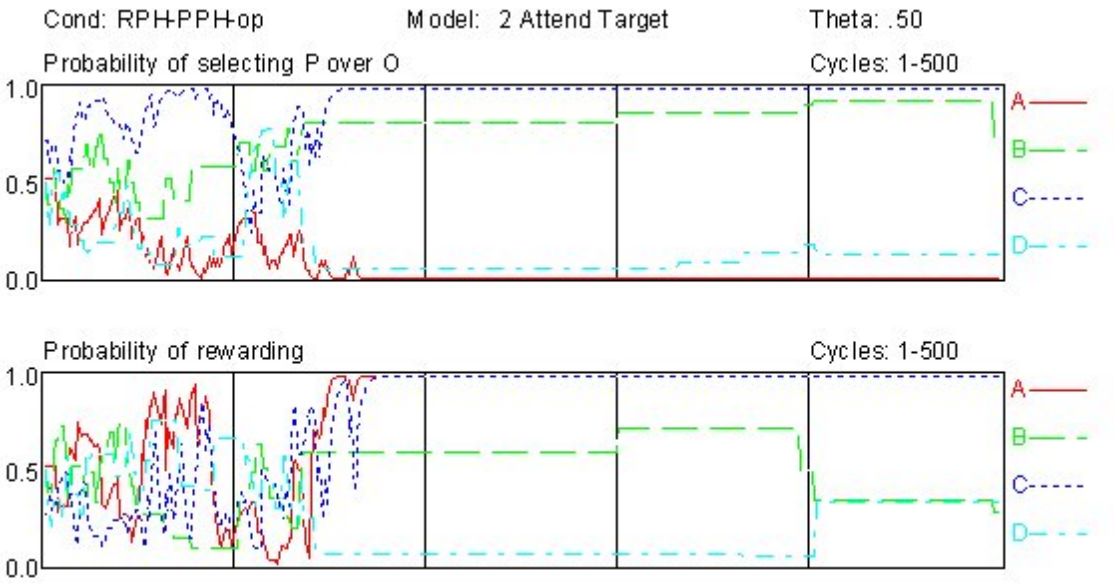

Figure 6. Example Group From Satisfaction Balance Model
with Attention to Target of Last Act, theta = 0.50.

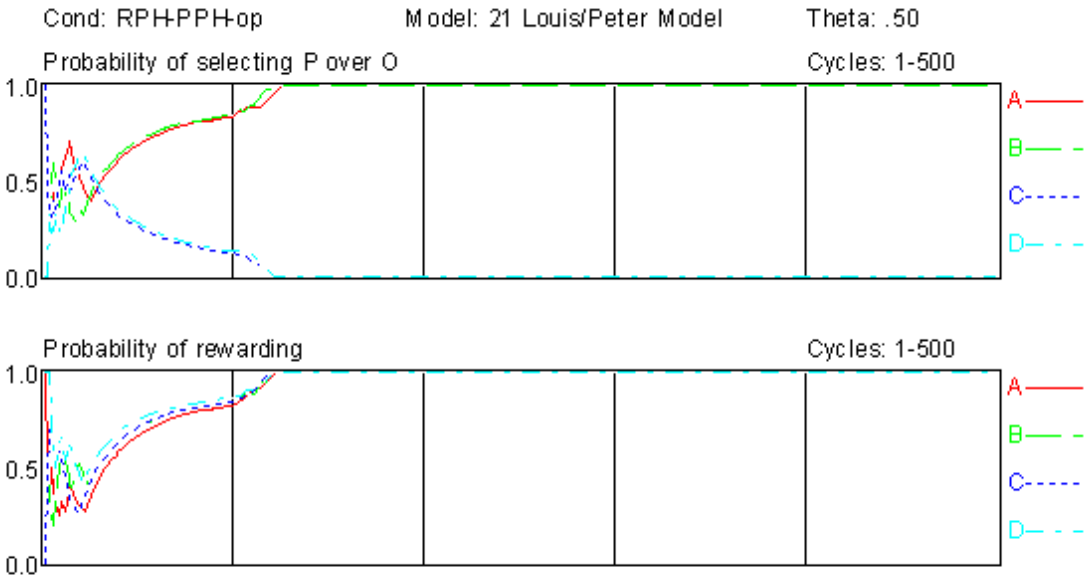

Figure 7.Example Group From Satisfaction Balance Model Based on Cost/Value Ratio with Attention to Target of Last Act.

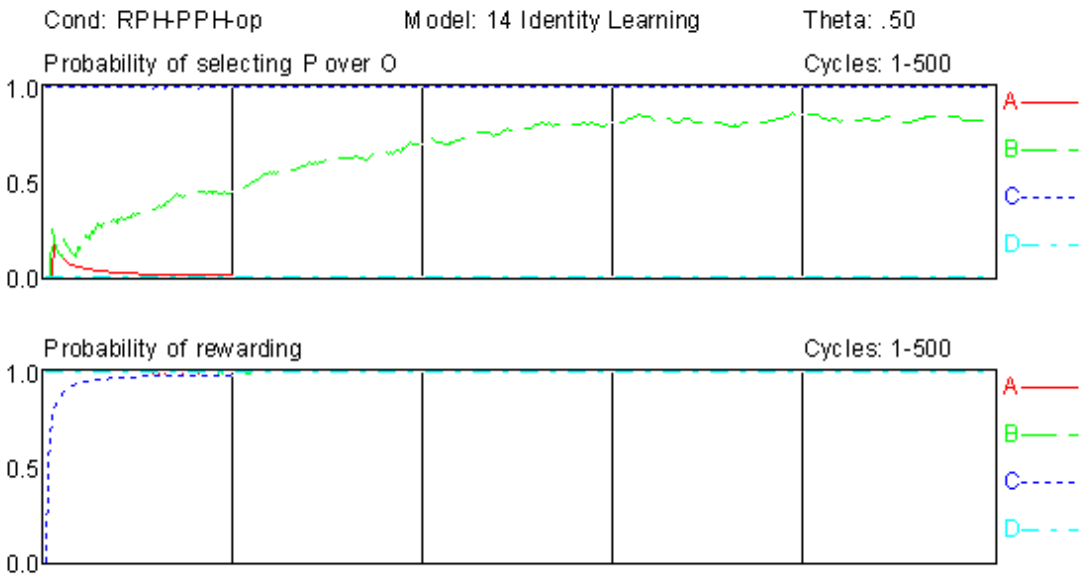

Figure 8. Example Group From Identity Model with Learning.