An Adaptive Network That Constructs and Uses an Internal Model of Its World

Richard S. Sutton
Andrew G. Barto
University of Massachusetts at Amherst

Abstract

Many theorists have emphasized the role of an "internal model of the world" in directing intelligent adaptive behavior. An internal model can be used to internally simulate the consequences of possible actions in order to choose among them without the necessity of overtly performing them. Animal learning theorists have taken latent learning experiments as demonstrations that animals can learn and use such internal models. In this paper, we describe an adaptive network of neuronlike components that constructs and uses an internal model, and we demonstrate this ability by describing a computer simulation of its behavior in a simplified analog of a latent learning task. The task has been made as simple as possible while still retaining those features that make behavior in latent learning tasks difficult to account for by connectionistic models. The network illustrates a principle by which connectionistic-like learning rules can give rise to behavior apparently requiring the formation and use of internal models. As such, it may help form a bridge between brain theory and connectionistic models on the one hand, and cognitive and information processing models on the other.

INTERNAL MODELS FOR SEARCH AND SIMULATION

Many theorists of the mind have emphasized the role of an "internal model of the world" in directing intelligent adaptive behavior (e.g., Arbib, 1972; Craik, 1943; Gregory, 1969; Johnson-Laird, 1980; MacKay, 1963; Piaget, 1954). The use of this expression has not been entirely uniform. For some, an internal model is a general knowledge store capable of answering any sort of question about the world. For others, an internal model is much more limited in that it can answer only a single question: "What should be done next?" In the first case another part of the mind can ask the internal model many questions before taking action, whereas in the second the internal model generates a recommendation for action only in response to the immediate situation. The first could conceivably answer the question "How heavy is that brick?" by giving its weight in pounds, whereas the second could only provide the appropriate motor commands to compensate for its weight while lifting the brick. The kind of internal model

Requests for reprints should be sent to Richard S. Sutton, Graduate Research Center, COINS, University of Massachusetts, Amherst, Mass. 01003.

with which we are concerned in this paper is of a generality intermediate between these two extremes. By an internal model we mean any part of an adaptive system that can provide expectations or predictions about what would happen in particular actual or hypothetical situations. Further, this paper is concerned specifically with those cases in which the model is used to simulate mentally the consequences of various actions in order to choose among them without having to try them overtly.

Kenneth Craik (1943) was one of the first to state clearly the view of thought as an internal simulation of the world, allowing many courses of action to be hypothetically attempted and evaluated:

> If the organism carries a "small-scale model" of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies that face it [p. 61].

Aspects of this theory of thought, however, are much older than Craik's work. Donald Campbell (1959) traces a very similar theory of "creative thought" back to the writings of Alexander Bain (1855/1874), Ernst Mach (1896), and Poincaré (1908, 1913).

Campbell emphasizes that the interaction with both the world and the internal model can involve trial and error:

> At this level [the level of creative thought] there is a *substitute* exploration of a *substitute* representation of the environment, the "solution" being selected from the multifarious exploratory thought-trials according to a criterion *substituting* for an external state of affairs. In so far as the three substitutions are accurate, the solutions when put into overt locomotion are adaptive, leading to behavior which lacks blind floundering [pp. 212–213].

Unfortunately, the idea of "trial and error" in search has frequently been mistaken for that of random or blind search. A search by trial and error can be a highly structured and heuristically guided one. By trial and error search we mean any search undertaken under the guidance of a certain kind of feedback process in which options are tried and then evaluated and retracted or changed if in error. Any "hypothesize and test" search, or any search using backtracking, would qualify as a search using trial and error in this sense.

Internal trial and error as a model of thought and reasoning turns out to be a view that is held extremely widely among theorists of the mind. Such a modeling/simulation view plays an important role in the theories of Dennett (1978) in philosophy; Simon (1969) in artificial intelligence; Sommerhoff (1975) and Arbib (1972) in brain theory; Dawkins (1976) in biology; and Galanter and Gerstenhaber (1956) and Miller, Galanter, and Pribram (1960) in psychology.

Figure 1 summarizes the essential features of this view of thought as used in this paper: An organism constructs an internal model of the world that allows prediction of the observable behavior of the world as a function of possible actions by the organism. The internal model is used to select behavior in an interactive manner similar to the interaction of the entire organism with the external environment. Trial and error search for the action that achieves the best result from the external environment is replaced by covert, internal trial and error search for the hypothetical action that secures the best anticipated result from the internal model. The internal model must be either faster, easier, or safer to interact with than the external environment in order for it to be useful.

This paper takes a few first steps toward formalizing this model-based theory of thought. The animal learning theory literature has been found to be extremely useful in obtaining a more concrete idea of what it means to create and use an internal model of the world. Animal learning theorists have concentrated on devising experiments that we can view as revealing indirect effects of an internal model on behavior. They have called the phenomena their experiments revealed

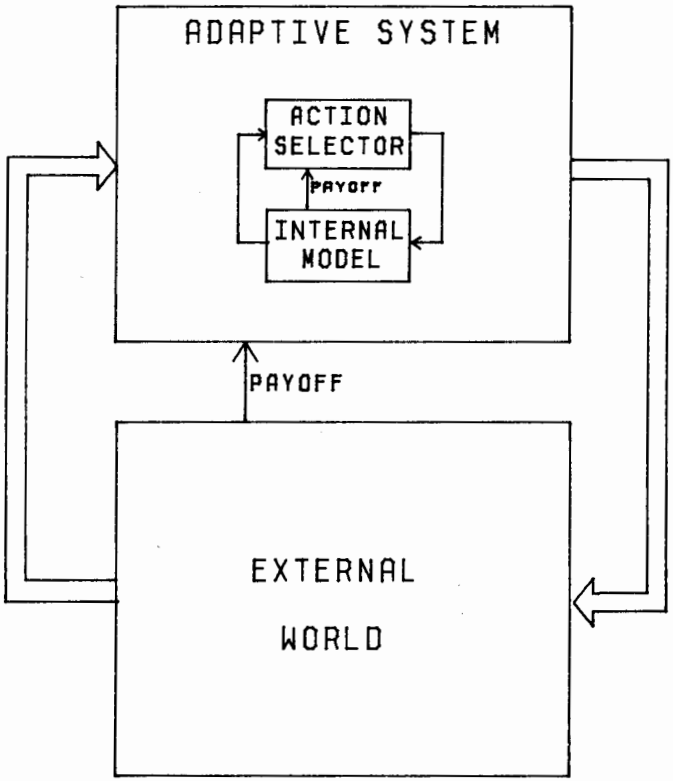

FIG. 1. An adaptive system based on the idea of internal simulation. The system interacts with its model in the same way that it interacts with the real world.

such things as reasoning, latent learning, and insight. The centerpiece of this paper is the presentation of an adaptive network that constructs and uses an internal model to solve a task similar to one in the animal learning theory literature. The intent was to find as simple a network and task as possible while still being able to demonstrate behavior that animal learning theorists would consider "reasoning," or model requiring. The network is constructed from adaptive neuronlike elements that have been extensively discussed elsewhere (Barto & Sutton, 1981a, 1981b, 1981c; Barto, Sutton, & Brouwer, 1981; Sutton & Barto, 1981), and whose design has been strongly influenced by the work of Klopf (1972, 1981). Both the adaptive network and the task environment were simulated by computer, and the results are presented here.

MAZE PROBLEMS AND INTERNAL MODELS

Figure 2 is a floor plan of an early form of a classic maze problem for rats (Tolman & Honzik, 1930). Its solution is considered to involve spatial reasoning capabilities. In short, the rats were familiar with all three paths to the goal and preferred them in order of increasing path length: A over B, and B over C. When

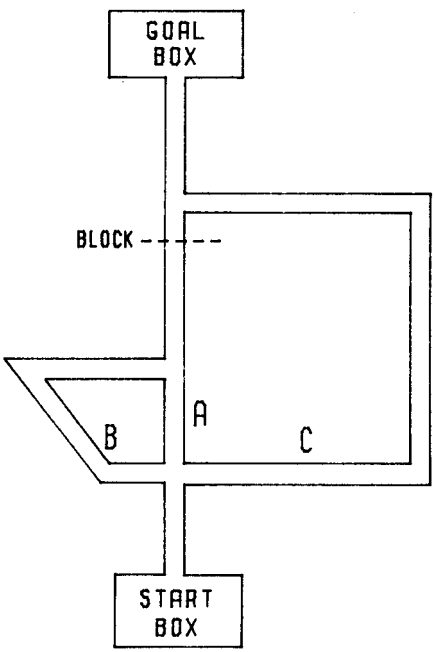

FIG. 2. A maze used to test insight in rats. The rats are familiar with all three paths to the goal box and prefer them in order of decreasing length: A over B, B over C. If they have "insight," then after taking A to discover the block, they next try path C rather than B.

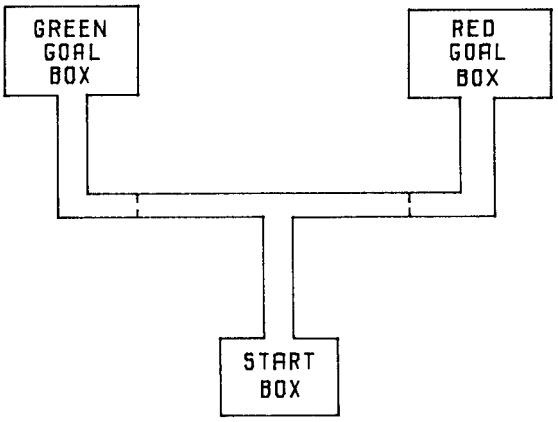

FIG. 3. A simple T-maze task with distinguishable detachable goal boxes used to test latent learning and reasoning in rats. The rat cannot see or backtrack through the one-way doors indicated by dashed lines. This task is conceptually very similar to the one posed to the simulated adaptive network presented in this paper.

a block was introduced in a corridor common to both A and B, as shown in Fig. 2, the rats tried A, discovered the block, and then predominantly chose path C, the longest of the paths, next. Inasmuch as their normal preference when A is blocked was B, the path of intermediate length, this result indicated that the rats used some sort of spatial map, or model of the maze that informed them that path B was also blocked. This experiment was seen as a positive test of insight or reasoning in the rat.

A much simpler experiment of the same intent uses a one-choice T-maze with detachable distinguishable goal boxes (Fig. 3). Because this problem is very similar to the one we have posed to the simulated adaptive network, we describe it in greater detail. There are three phases to the experiment: In *the exploration phase*, the subject is repeatedly placed at the entrance to the maze. When the subject reaches one of the goal boxes, it is removed from the apparatus. There is no food or other reinforcer anywhere in the maze. Backtracking is not allowed. In *the association phase*, the goal boxes are separated from the T-maze and carried to another room. There the subject is fed in the red goal box that was on its right and given a painful electric shock in the green goal box that was on its left. In the *testing phase* the goal boxes are replaced and the subject is then returned to the start of the T-maze.

The key question is: Which way will the subject turn on the first post-training trial? Most rats will turn right. Note that neither the action of turning right nor the action of turning left is ever temporally associated with reward or punishment in this experiment. In order to solve this task, the subject has to combine two separately learned facts about the world: (1) that turning right in the T-maze will bring it to the red goal box and turning left will bring it to the green goal box; and (2) that the red goal box is a place where it may be fed, and the green box

a place where it may be shocked. It is this combination that is thought of as the reasoning process, a sort of transitivity of prediction or primitive *modus ponens*.

Viewing the solution of this T-maze problem as an instance of the use of an internal model, in this case a spatial map, suggests two aspects of the idea of simulation by internal model that may account for the popularity and apparent promise of the idea. First, the sort of reasoning by predictive transitivity mentioned previously is precisely the sort of reasoning that is achieved by a simulation. To simulate a complex system by computer, we provide the step-by-step transition dynamics of the system, and the simulation scheme repeatedly applies these dynamics to update the state of the simulated model. In just this way a simulation can combine "right turn predicts (arrival at) red goal box" and "red goal box predicts food" to infer that food can be attained by turning right. Such a capability for propagating predictions is an important component of the ability to generate the consequences of proposed actions.

The second important aspect of the idea of simulation by internal model that appears in this simple T-maze example is that it provides a framework for learning about the environment even in the absence of rewarding or punishing events. From a simple reinforcement learning point of view, we are at a loss to explain the learning that the rats do about the maze in the first part of the experiment in which they receive no reward or punishment. The idea of constructing an internal model gives us another point of view from which this learning is much more understandable: They learn in order to form an accurate predictive model. Reinforcement, being a one-dimensional measure, can provide very little information compared to the torrent of sensory information. It has become generally recognized that intelligent artificial adaptive systems also must use this additional information to solve complex learning problems (see the discussion of the "credit-assignment problem" in Minsky, 1961). Much of the promise of the idea of an internal model may be that this concept explicitly encourages and provides a way of understanding learning in the absence of reward or punishment. This type of learning involves the construction of an accurate predictive model, a process that is normally independent of reinforcement. Once the model is formed, internal trial and error through simulation provides a framework for using the information from the environment.

THE SIMULATED TASK ENVIRONMENT

Because both the experimental subjects and their environment were simulated by computer, we were able to simplify the experimental design even further than was done in the T-maze experiment. The ground plan of the simulated environment is shown in Fig. 4. The lower area is used in a manner analogous to the T-maze, the two regions on the right and left being analogous to the red and green goal boxes at the ends of the T-maze. The two enclosed regions shown in the upper part of Fig. 4 are analogous to these same goal boxes when they

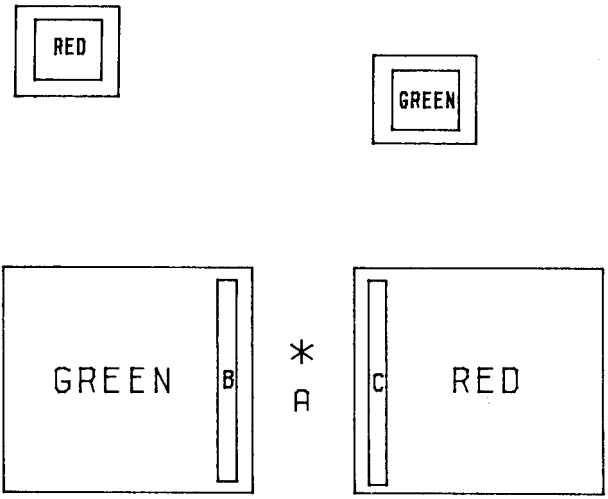

FIG. 4. Ground plan of the simulated environment. The lower area is used in a manner analogous to the T-maze in Figure 3, the two regions on the right and left being analogous to red and green goal boxes at the ends of the T-maze. The upper two enclosed regions are analogous to those same goal boxes when they have been moved to another room for association with food and shock in the absence of the T-maze. The initial position of the network is indicated by the asterisk at A.

have been moved to another room for association with food and shock in the absence of the T-maze. That these are actually separate regions is of no importance here: The adaptive network controlling the simulated subjects has only three sensory input lines, one for sensing being within a green region, one for sensing being within a red region, and one for sensing rewarding stimulation. In terms of this limited sensory vocabulary, all regions of the same color are indistinguishable. This is clearly an enormous simplification of the perceptual process.

The network has only two graded actions: moving to the right, and moving to the left. These are meant to be extreme simplifications of, and yet analogous to the right-turn and left-turn actions of the T-maze task. In the exploration phase of the simulation experiment, each network is placed at A, between the two large colored regions (Fig. 4), and allowed to wander back and forth randomly. (We use the terminology of actual placement and movement, though this is all simulated numerically in the computer run of the model.) The barriers at B and C obstruct its movement, thereby preventing it from moving too far away. This insures that it eventually gains experience moving to and from both regions. Thus, all trajectories are along a straight horizontal line between the two barriers. The two upper goal box areas shown in the upper part of Fig. 4 are used in the association phase. For the testing phase, the network is returned to location A between the lower two regions to see which region is entered first.

We next describe the adaptive network and then proceed through each phase of the simulation experiment, discussing the experimental manipulations and

network changes in detail. For reference, Appendix A contains a summary of the details of the three phases of the simulation experiment, and Appendix B contains a detailed specification of the simulated adaptive network model.

THE SIMULATED ADAPTIVE NETWORK

Figure 5 is a block diagram of the adaptive network design. The network is divided into two major components: an action-selecting mechanism and an internal model of the environment. The action-selection mechanism uses the actual environment and the model of the environment in exactly the same way—both provide feedback to evaluate actions attempted by the action-selecting mechanism. The evaluations by the model and by the environment of the most recently selected action are added to yield the evaluation input to the action-selecting component. Importantly, *the feedback loop through the internal model is much faster than the feedback loop through the environment*, and thus proposed actions can be evaluated by the model so quickly that the rejected alternatives have very little influence on the environment and the organism's overt behavior. This is accomplished in the simulated example system by letting the motion of the network depend not only on the instantaneous action selected, but also on past values, in an exponentially weighted manner. The result is that even though both the action selection and the overt action are changed and updated every time step of the simulation, the environment is slightly "viscous" relative to the network dynamics and has a kind of inertia that must be overcome before overt action aligns with the current action selection. A decisive overt movement occurs only when the system has converged onto a particular choice of action. Maintaining a particular action as the one selected for a significant period of time (about four time steps in the simulated system) causes that action to become expressed in overt movement.

The Action-Selecting Component

The division of the adaptive network into the action-selecting and internal model components makes its construction from adaptive elements relatively simple. All that is needed for the action selector is a bank of action elements that correlate their output, or action, with increases or decreases in the evaluation or reinforcement input to this subsystem (Fig. 6). We will need one element whose action represents the tendency to turn right and one whose action represents the tendency to turn left. The environment will then resolve any conflict between these two by responding to the difference in the tendencies (this is explained in detail in Appendix A). The action levels are originally chosen randomly, but if a correlation is found between action level and subsequent evaluation, the choice of action is biased to make positively correlated actions more likely to be selected and negatively correlated actions less likely. Mathematically, the momentary

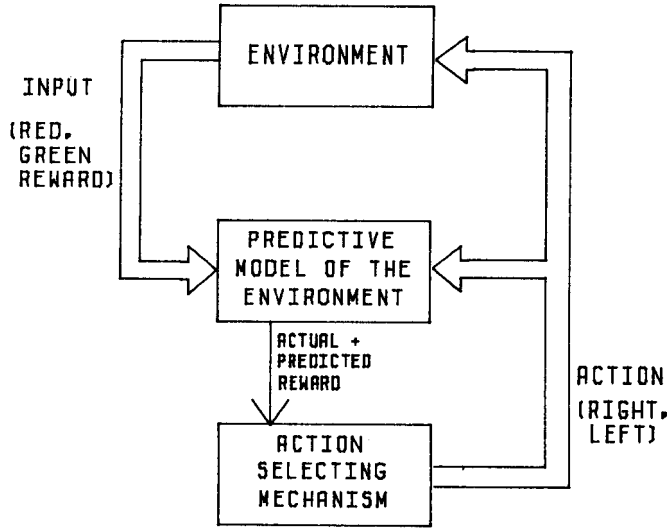

FIG. 5. Block diagram of the adaptive network and its connection to the environment. The action selecting mechanism has its choices evaluated via two feedback loops; one through the environment, and one through an internal model of the environment. If the model is faster than the environment, then the feedback loop through the model will control overt behavior.

action choice of each element is the sum of a random component and a bias or accumulated correlation component:

$$A_i(t) = F(v(t) + B_i(t)) \text{ for all actions } i$$

where

$$F(x) = \begin{cases} 1 & \text{if } x > 1 \\ x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x < 0 \end{cases} \tag{1}$$

and where $A_i(t)$ is the strength with which the ith action is selected at time t; $v(t)$ is a random variable, normally distributed with mean 0; $B_i(t)$ is the bias weight for the ith action at time t, an accumulated measure of the correlation observed between the ith action and reward changes (see following).

To correlate actions with subsequent evaluation changes, each element maintains a short-term memory, known as its eligibility (following Klopf, 1972), of the extent to which it has been active. When an evaluation change occurs, element biases are modified according to the extent of their eligibility. Mathematically, the correlation bias weights are accumulated as follows:

$$B_i(t) = B_i(t-1) + C\cdot[E(t) - E(t-1)]\cdot\bar{A}_i(t) \qquad \text{for all actions } i$$

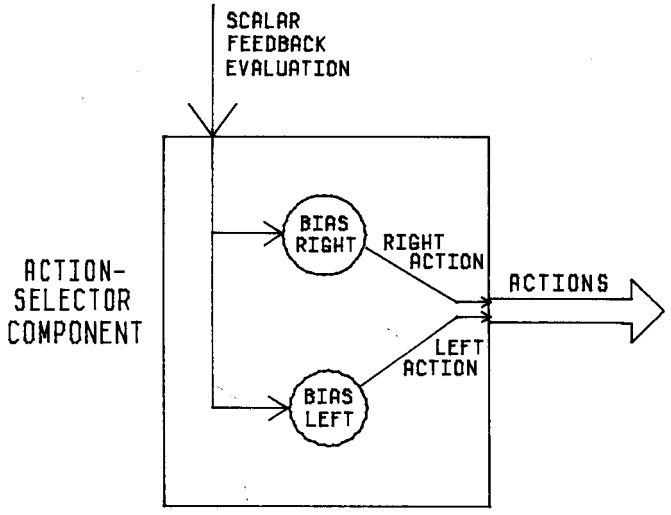

FIG. 6 Detail of the action-selector component of the simulated network blocked out in Figure 5. The elements correlate their output, or action, with increases or decreases in the evaluation input.

where $B_i(t)$ is the bias toward the ith action at time t; C is the learning rate parameter for the bias weights; $E(t)$ is the feedback evaluation or reinforcement at time t; $\bar{A}_i(t)$ is the eligibility of the ith action, an exponentially decreasing weighted trace of values of A_i before time t [in the simplest case $\bar{A}_i(t)$ is merely $A_i(t-1)$].

At the start of the simulation experiment, the bias for each action is zero, favoring neither right nor left actions. During the exploration phase, the simulated experimental subjects move back and forth randomly between the lower large red and green regions of the environment (Fig. 4) without reinforcement of any sort. Because reinforcement does not occur, nothing is predictive of reinforcement, and reinforcement is never predicted by the internal model component. Without reinforcement or its prediction, the action evaluation is always zero, and there can be no correlations between action and changes in evaluation during the exploration phase. Consequently, the bias weights remain zero during the exploration phase. Once the testing phase has been reached, an internal model will have been constructed such that a correlation will exist through the internal model even in the absence of any external stimulation. This will result in the action selector converging on a preference for one of the actions (this is discussed in more detail later).

The sort of trial and error learning system presented in the foregoing is well known from the work on Harth's ALOPEX system for mapping receptive fields (Harth, 1976; Harth & Tzanakou, 1974) and from the learning automata literature inspired by Tsetlin's work (Tsetlin, 1973). In a less simplified system than the network described here, it would be highly desirable to modify this action-selecting mechanism so that it is able to use sensory information in selecting

actions. This would allow it to learn to perform different actions in different situations without starting its search all over again each time the situation changed. Instead, it could remember for each situation what actions were most successful in previous experiences. Trial and error learning mechanisms can be made sensitive to context in a fairly straightforward manner (Barto, Sutton, & Brouwer, 1981; Klopf, 1972, 1981; Mendel & McLaren, 1970; Michie & Chambers, 1968). However, it is not clear whether the actual current input, or the predicted input, or some combination of the two should be used as the context for the action selector. This problem with the current design is closely related to several others that emerge when sequences of actions need to be internally simulated in order to evaluate possible next actions. An additional mechanism, such as a method for clearly separating actual from anticipated situations, is probably necessary to handle these cases. This example system is only a first step toward an adaptive network capable of creating and searching general internal models, and we do not consider these possibilities further. The artificial intelligence literature on planning (Sacerdoti, 1977) would be highly relevant to future extensions of this adaptive network mechanism.

The Internal Model Component

The construction of the internal model of the network's world is a system identification task, and the solution adopted here follows Kohonen's suggestion (Kohonen, 1977) for doing system identification using a trainable associative memory (Amari, 1977; Anderson, Silverstein, Ritz, & Jones, 1977; Cooper, 1974; Kohonen, 1977; Longuet-Higgins, Willshaw, & Buneman, 1970; Nakano, 1972; Palm, 1980; Wigstrom, 1973; Willshaw, Boneman, & Longuet-Higgins, 1969; Wood, 1978). Kohonen's general idea was to train the associative memory with sample input to the system to be identified as the recall key, and to use the

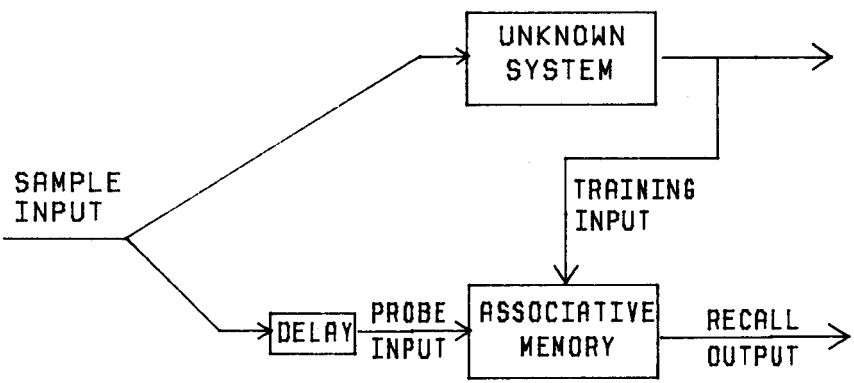

FIG. 7. Kohonen's (1977) suggestion for doing input-output system's identification with a standard learning associative memory. The associative memory is trained by presenting paired samples of the input and output of the unknown system.

resultant output of the unknown system as the training pattern to be recalled (Fig. 7). If the unknown system has no memory (that is, it simply implements a function from input to output), then the associative memory will form a best least squares linear approximation of the unknown function.

If the unknown system is the environment for an adaptive system, then this process will yield nearly the appropriate sort of model. Figure 8 is a slightly more detailed block diagram of the associative memory-based machinery in the simulated adaptive network for model construction and use. The associative memory here differs from the standard associative memories in being predictive: It produces as its recollection a prediction of what the next key will be. (In this sense it is similar to some of the early models of temporal associative memories; for example, Longuet-Higgins, 1968a, 1968b; Longuet-Higgins et al., 1970.)

Figure 9 shows a detailed wiring diagram of the model construction and readout machinery. This component consists of a bank of elements, each responsible for the prediction of a certain feature of the environmental stimulation, available in this case on the separate input lines for red, green, and reward

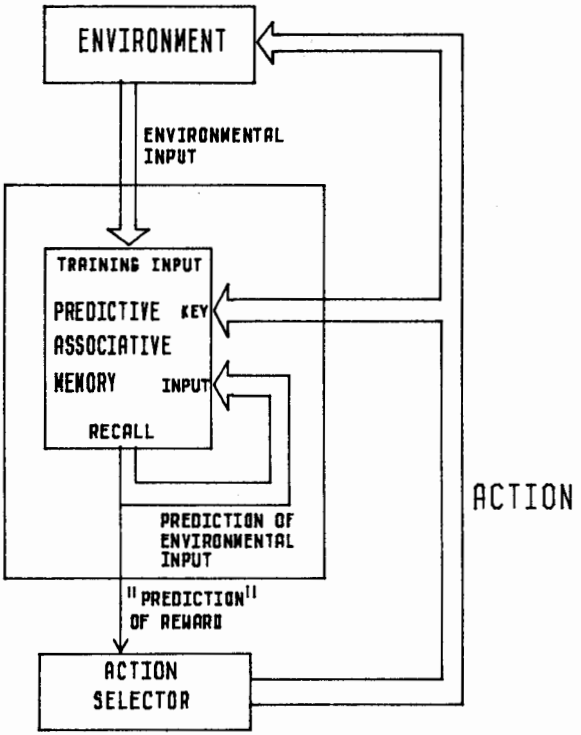

FIG. 8. A more detailed block diagram of the adaptive network (cf. Figure 5) showing the central role of a predictive associative memory.

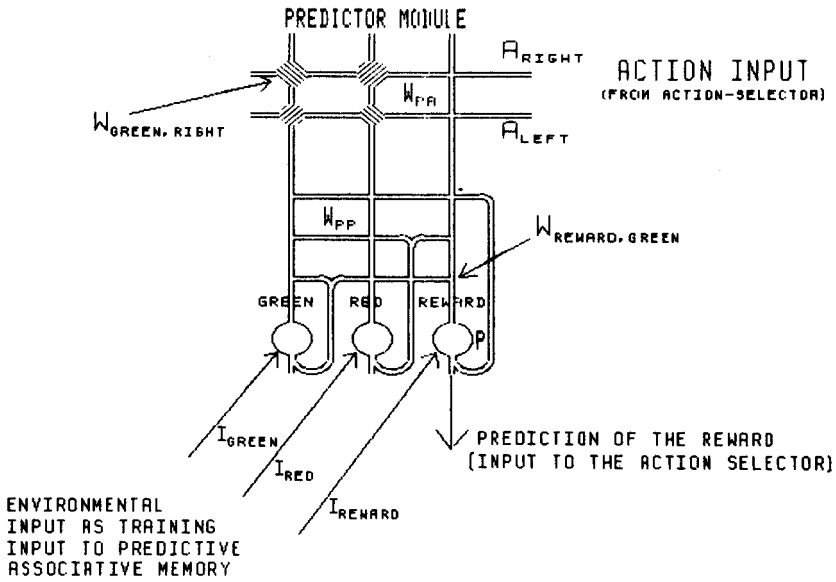

FIG. 9. A more detailed wiring diagram of the model construction and readout mechanism (see text for explanation). The size of the striped disks at the fiber intersections indicates the magnitudes of the weight associated with that connection. The sign of the weight is indicated by the slanting of the stripes (and, in later figures, by explicit plus or minus signs), where slanting up to the right indicates a positive weight and slanting up to the left a negative weight.

stimulation. As a basis for making these predictions, each element is provided with the current action selection from the action-selector component and the most recent predictions of stimulation from the other predictor elements in this component. The fact that predictions of stimulation are used to make further predictions results in the recurrent architecture of this network.

Each predictor element has a weight associated with each of its five inputs from the action selector and recurrently from the predictor module. For example, $W_{\text{green.right}}$ is the upper left-hand weight pointed out in Fig. 9 from the right-moving action A_i to the predictor for green. The size of the striped disk indicates the magnitude of the weight in this and later figures. Stripes slanting upward to the left indicate a negative weight and stripes slanting upward to the right indicate a positive weight; in Figs. 10–14 the sign of each weight is also indicated explicitly by a plus or minus sign next to the weight. The output $P(t)$ of these elements is the weighted sum of their inputs from the action selector and recurrently from the predictor elements, plus their special training input directly from the environment:

$$P_i(t) = F\left[\sum_j W_{ij}(t)A_j(t) + \sum_k W_{ik}(t)P_k(t-1) + I_k(t)\right]$$

for all predictor elements i, where the sums are over all actions j and all predictor elements k; the function $F()$ is given by Equation (1); and $I_k(t)$ is the value of the red, green, or reward input from the environment at time t.

A connection or synapse W_{ij} from action or predictor j to predictor i is said to be eligible at time t if its input fiber, the jth input to the predictor module, has been active in the recent past. A connection can have its weight changed only when it is eligible. If a predictor element experiences a change in its level of output activation, then the eligible synapses are changed in the same direction as that change in activation. Because the eligible connections are those that were active slightly earlier, the result will be such that if the same input situation occurred again, the change in activity would occur earlier. In other words, these adaptive elements signal their predictions by responding as they would if the predicted stimulation were already present. For example, if the input information indicates that an element is about to receive strong excitatory stimulation, then the element becomes highly active immediately. This property, combined with the recurrent network architecture, results in the ability to chain predictive associations of as great a depth as there are features to predict (e.g., A predicts B, B predicts C, etc.). For the present purposes, this is the only crucial property of these predictor elements. Appendix B contains a complete specification of their operation in this network and Sutton and Barto (1981) and Barto and Sutton (1981c) extensively discuss their behavior as individual elements.

Although sufficient for the simple example system presented here, these predictor elements may not be ideal for the purposes of constructing and using an internal model. Without going any further into the details of this element or the alternatives, we can note that these elements require the recurrent connections from each to itself to be made ineffective (zero and nonplastic) for effective operation in the architecture chosen for this network. This is easily arranged, as it is in the simulated example system, but it is not an elegant solution, suggesting that there may be other hidden difficulties. This problem may suggest directions to proceed in deriving elements better suited to this purpose.

SIMULATION RESULTS

Two hundred instantiations of this network, each with a different "seed" for the pseudorandom number generator, were run through the main experiment. The following subsections describe each phase of the experiment, and the general behavior of the networks during that phase. In addition, the state of a particular network is followed through each phase (Figs. 10–14) to illustrate why these networks exhibit each behavior.

The Exploration Phase

During the random wanderings of each simulated adaptive creature in the exploration phase, the internal model component is forming a model to predict the

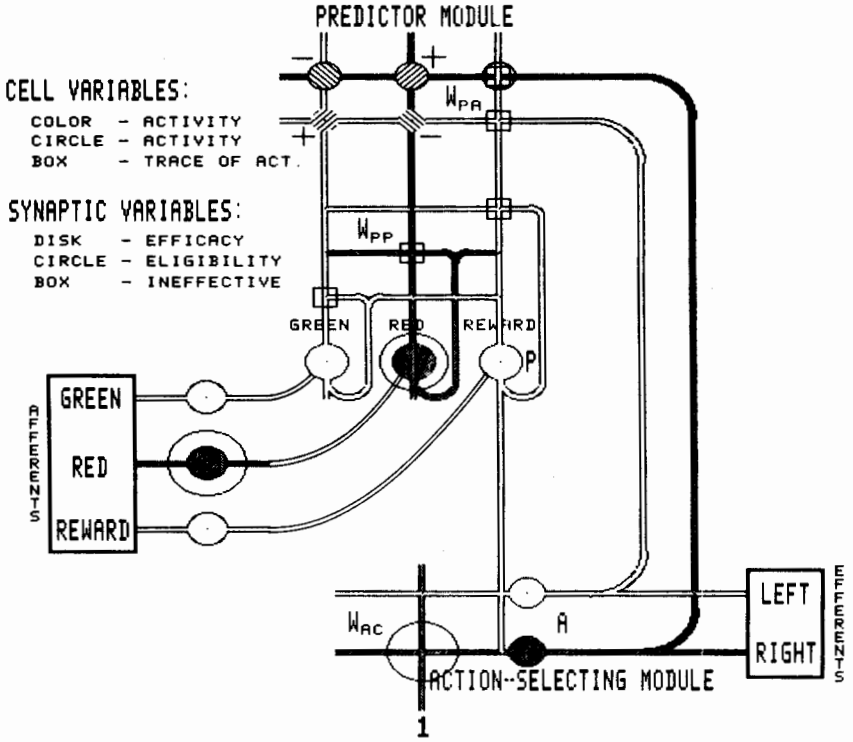

FIG. 10. The state of the network near the end of the exploration phase for one of the simulated experimental subjects. See text for explanation.

red and green stimulation changes it experiences when it occasionally wanders into or out of one of the two colored regions. Figure 10 shows the state of the network near the end of the exploration phase for one of the simulated experimental subjects. The four relatively large connections from the actions to the green and red predictors of the predictor module indicate that the net has learned that right-moving actions predict increases in red stimulation and decreases in green stimulation, whereas left-moving actions predict increases in green stimulation and decreases in red stimulation.

The particular snapshot of the network activity in Fig. 10 shows how this knowledge is accumulated. The right-moving action has been selected more than the left-moving action in the last few time steps, as indicated by the greater eligibility of the connections from this action to the predictor elements, and in fact the subject has moved right during the most recent time step. This rightward movement has just brought the subject into the red region (this is indicated by the high activity in the red input while its trace of activity—indicated by the size of the box—is still zero). The resultant sensory input stimulates the red predictor element, causing an increase in its activity (indicated by the circle centered on this element being larger than the square), and this causes an increase in the

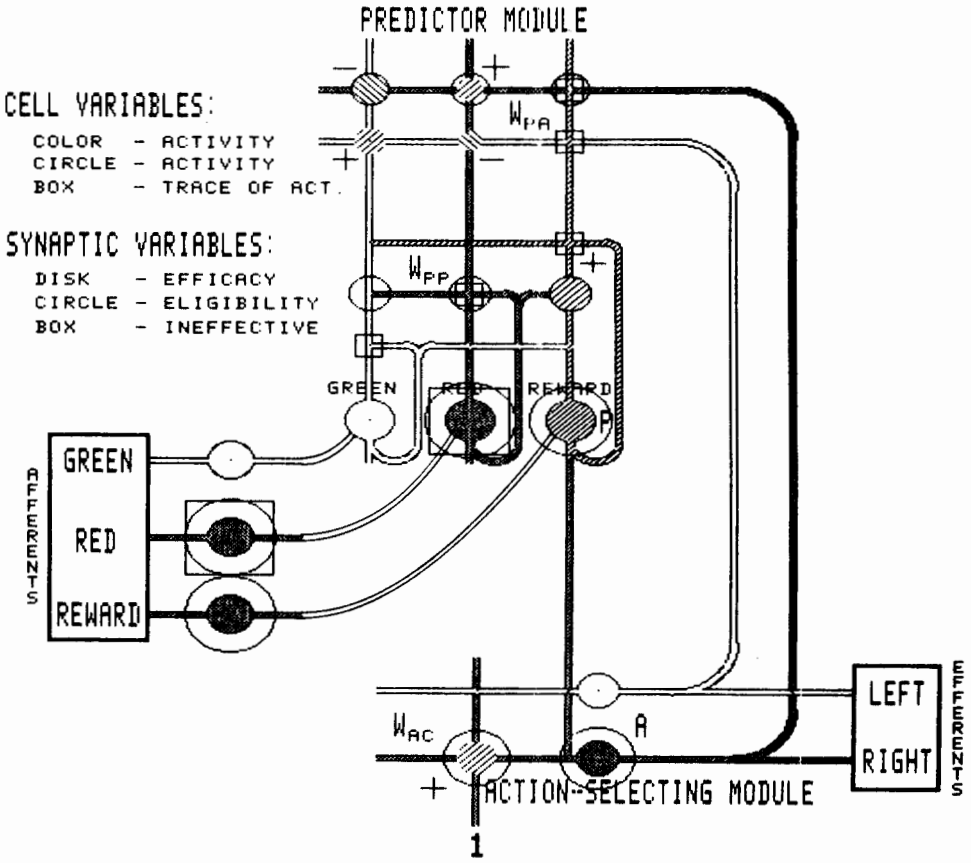

FIG. 11. The network state just after the reward is provided. See text for explanation.

eligible connections. The most eligible connections, as we have already seen, are those from the right-moving action. The net result is to further strengthen the pattern of learned associations that we already see present in these connections.

The Association Phase

After 1000 time steps of the exploration phase, each subject is moved to the red goal box, left there for two time steps, and then provided with full reward stimulation. Figure 11 shows the network state just after the reward is provided. Note the large positive connection from the red predictor to the reward predictor. The net has concluded that a prediction, or actual occurrence, of red predicts reward, as the red predictor element was highly active just prior to the increase

in reward stimulation. The eligibilities of the connections from the red predictor are indicated by the large circles at these connections. Next, each subject is moved to the green goal box, left there for two time steps, and then the reward stimulation is removed. By a completely analogous process, the green prediction becomes a predictor of loss of reward, and the corresponding connection becomes negative (Fig. 12).

The Testing Phase

In the testing phase, each subject is returned to the initial location between the lower large red and green regions. As soon as the subject enters one of these regions the trial is over and the simulation is stopped. Of 200 subjects, 141—over 70%—entered the red region first, statistically a highly significant result ($p < .005$). A second experiment was also performed in which the lower red and green regions and their barriers were removed during the training phase, but

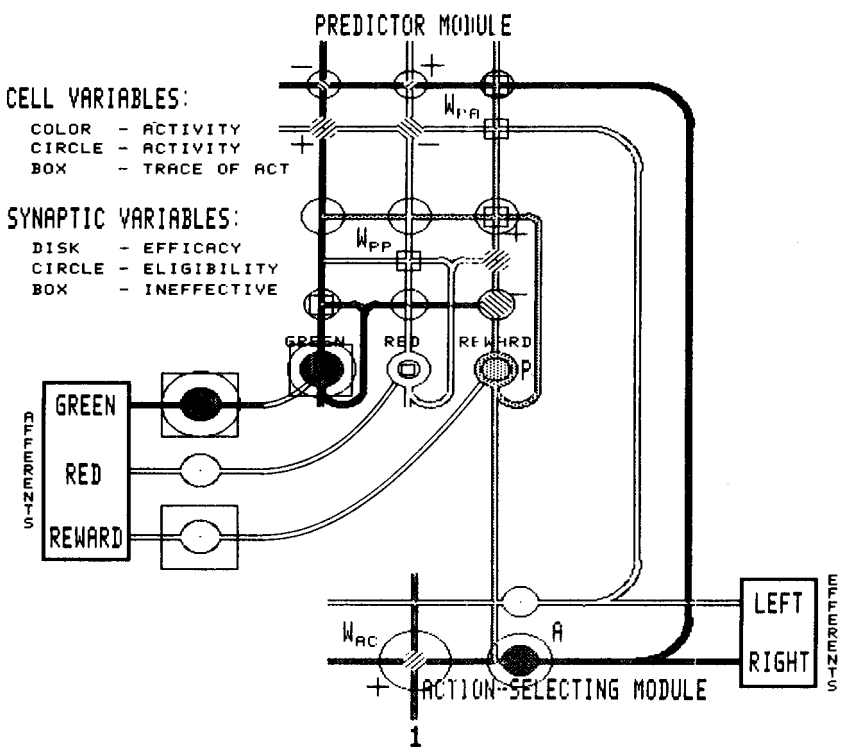

FIG. 12. The network state after the reward has been removed while the subject was in the green box. Prediction of green has become a predictor of loss of reward, as indicated by the negative connection between the predictor elements for green and reward.

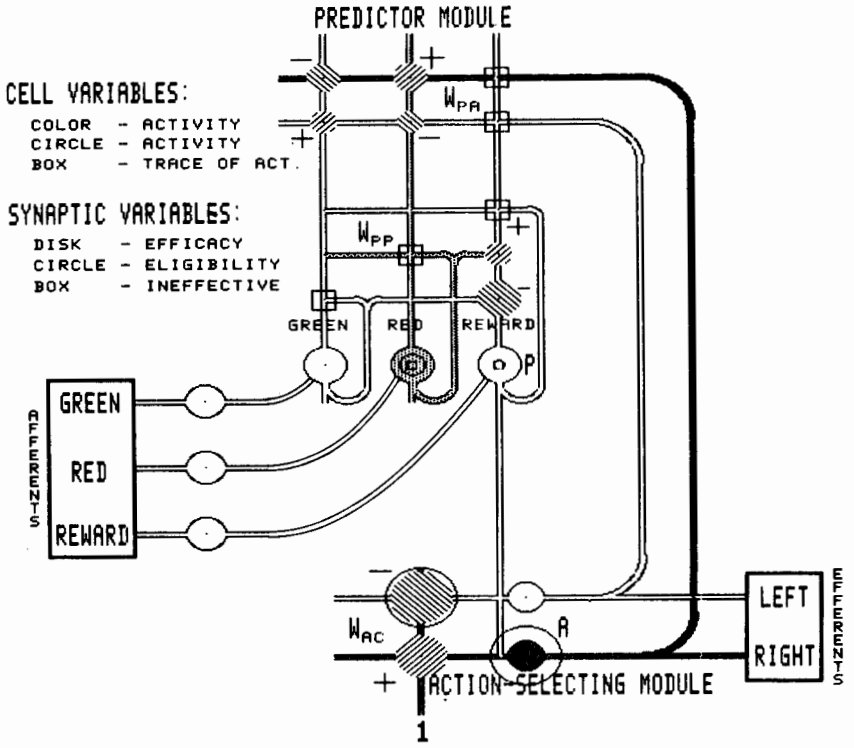

FIG. 13. The state of the network of one of the subjects of the primary experiment several time steps into the testing phase but before either region has been entered. See text for explanation.

which was otherwise identical to the first experiment. The testing phase was halted after 300 time steps and the position of the subject was recorded. Each of the 100 subjects had moved far to the right, many times beyond the original location of the red region, at the end of that time. This indicates that the statistical nature of the primary result is due to random movements bringing some of the subjects within the green region, and thus ending the trial, before they have had enough experience with their internal models to be directed to the right.

Figure 13 shows the state of the network of one of the subjects of the primary experiment several time steps into the testing phase but before either region has been entered. Notice that the bias weight for moving right (toward the red region) has become positive, whereas the bias weight for moving left (toward the green region) has become negative. The successive snapshots of network state in Figs. 13 and 14 provide an example of how this comes about during the testing phase. In Fig. 13 we see the action-selecting module happening to choose the instantaneous action causing movement right. This selection results in an increase in the activity of the red predictor element, because moving right was found to be

a predictor of red stimulation in the exploration phase. At the next time step (Fig. 14), we see the prediction of red stimulation cycling around to activate the reward-predicting element via the excitatory connection established during the association phase. Thus, as a consequence of the action selector's momentary choice of the right action, the predictor module, acting as an internal model, has generated the prediction of increased reward. Because the right action was selected at the previous time step, its bias weight is eligible (indicated by the large circle at its bias connection in Fig. 14) when the prediction of increased reward arrives. Thus, this bias weight is increased, and the network is further biased towards moving right. If on the other hand, the action selector had momentarily chosen the left action, green stimulation would have been predicted. This in turn would have predicted a decrease in reward stimulation, and the selection of the left action would have been punished.

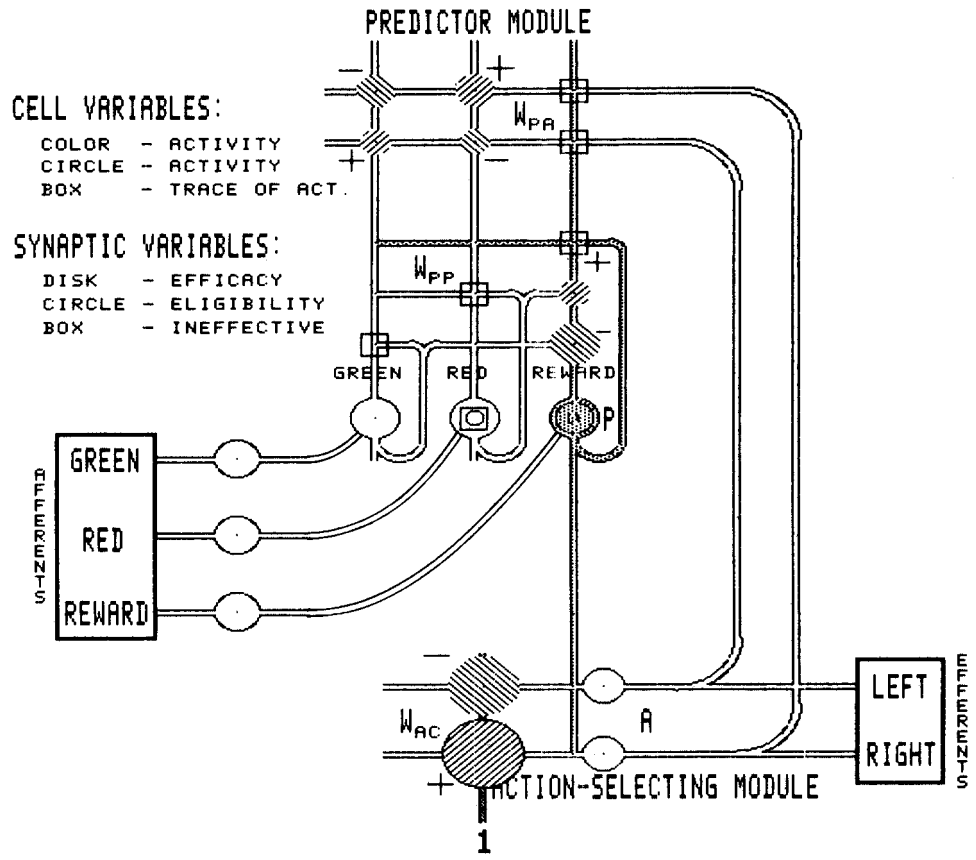

FIG. 14. The state of the network one time step after Figure 13. See text for explanation.

Recall that these momentary selections of right or left actions will not in general be accompanied by actual right or left movement. The environment responds to the actions selected in a relatively slow inertial manner: Several action selections in the same direction are usually necessary to cause actual motion in that direction. Both overt and covert actions are updated every time step of the simulation. The only difference is that the overt action, the "physical" movement, depends not only on the current covert action selection but also on past selections, weighted according to recency. Thus a consistently selected covert action becomes the overt action, while rapid fluctuations in covert action selection are averaged out. In this way the process of covert, internal trial and error via the predictive internal model can occur with relatively little overt action by the subject. No other delaying or decision-making machinery is necessary to make the transition from covert thought trials to overt movement. In fact, the action-selecting component (referring to Fig. 5) is completely oblivious to whether it is receiving feedback from the external environment or from its internal model. From the action-selecting component's point of view, the acquisition of an internal model merely means that the feedback for its action selections returns more rapidly, significantly easing its problem of controlling that feedback.

Superstitious Learning

During the association phase, reward is provided and then taken away from each subject while it is in the red and green goal boxes respectively. During this time, whatever action the network happened to select just before the reward stimulation is changed will be strongly reinforced or punished by that change. For example, the subject whose network state is shown in Fig. 12 happened to associate the left action with reinforcement, as shown by the larger bias weight for the left action than for the right action. One might expect that the effect of these reward changes would be dependent on whatever action was randomly chosen just before the reward changes and that the effect on later behavior would thus be, on the average, symmetrical with respect to right- and left-moving actions. To ensure that this was the case, a third experiment was performed that was identical to the first in all respects except that the action bias weights were set to zero just prior to the testing phase. This insured that there would be no initial bias either to the right or the left. Of the 100 subjects in this third experiment, just over 70% entered the red area first, confirming that the decision to move right can be made during the testing phase based completely on information stored in the internal model of the predictor module.

THE REPRESENTATION PROBLEM

This system was constructed to be the simplest possible complete system capable of constructing and using an internal model. As such a minimal example, it only

begins to address some of the critical issues involved. The simulated network was provided with a representation of the environment specially tailored to the task it was to solve. It had unique input lines for red and green stimulation, and the environment consisted only of areas that were entirely green, or entirely red, or neither. The relationships to be learned between actions and resultant stimulation, and between stimulations, were very simple ones in terms of the available action and stimulation representations. Mathematically, a network such as the one used here can only learn linear relationships between its representation of action, stimulation, and subsequent stimulation. To the extent that the actual relationships depart from linearity, such a network would be unable to form an accurate model.

One strategy for solving this difficulty is to retain the linear learning rules but to attempt to evolve continuously a representation compatible with that linearity. In general this is a difficult unsolved problem. Input features, output commands, and internal representations of environmental state (in the example system, environmental state was not necessary in forming a predictive model) all would need to be developed. This problem is closely related to the representation problem of artificial intelligence. Unfortunately, however, most of the work on this problem in artificial intelligence is unhelpful in that it merely attempts to find a good representation for a particular task rather than techniques for evolving representations in a more general setting. Genuinely relevant work includes the feature extraction work in pattern recognition (Bledsoe & Browning, 1959; Klopf & Gose, 1969; Selfridge, 1955; Uhr & Vossler, 1961), Samuel's checker player (Samuel, 1959), and the work on nonlinear associative memories (Poggio, 1975; and Barto, Anderson, & Sutton, in preparation). A fundamental heuristic central to much of this representation development work is to direct the search for better representations according to which representation elements have already proved most useful.

Although the network used here was given a sufficient representation *ab initio*, and has no capabilities for representation development, it does serve as a basis for considering what simultaneous environmental interaction and representation development may involve. In particular, we assume that there must be some property of the environmental interaction that indicates when and in what way the current representation needs to be changed. If this sort of example allows us to observe these properties in a simple case, then we will have made progress toward making an adaptive network appropriately sensitive to them.

DISCUSSION

The network presented here embodies a method for adaptive control based on systems identification (model construction) that is very general: (1) use repeated experiments with the input–output behavior of the system to be controlled to construct a model that yields similar behavior; and (2) to select each control

action, first *interact with the model*, in an input–output or black-box manner, to determine which action is optimal in terms of the model. The essential aspect of the model is that it is a behavioral model: Its successful use depends only on its input–output validity. The model can be interacted with to achieve an optimal response just as the external world is interacted with in the absence of a model. These perspectives on the nature and use of a model were summarized pictorially in Fig. 1. It should be noted that appropriate general techniques for this sort of interaction with an environment or model are not currently well understood. This is an area of current active investigation by our research group.

Although we emphasize an input–output view of the internal model, this is not a return to the pre-state-space ways of thinking characteristic of the work of the 1950s. Such an internal model will in general include states (even though the model in the simulated example system did not). However, we do wish to emphasize that for this use of the model only the input–output aspects are important.

Another important feature of the network presented here is the method used for coordinating interaction with the real environment and interaction with the model of that environment. Consider the approach taken in most artificial intelligence systems for problem solving or reasoning about actions, such as SRI's famous robot SHAKY (see Raphael, 1976). SHAKY operates in three identifiable modes. In one, he visually scans the environment and constructs an internal model of it, aided by a priori knowledge and assumptions. In another mode, he uses his model of the environment and his current goals to perform a sophisticated search through the space of possible paths and actions. This search takes the form of an internal simulation with backtracking of many of the possible action paths. Finally, in the third mode, SHAKY shuts off his internal model and visual apparatus and executes "ballistically" his precomputed next action or series of actions. When the action is complete, or some unusual event occurs, SHAKY returns to one of the other modes. A lot of work in artificial intelligence has concentrated on the model search step of the foregoing scenario, without going any further toward coordinating the model interaction and the interaction with the real environment. By contrast, the example adaptive network presented here performs all three functions—model acquisition, model interaction (search), and real time environmental interaction—simultaneously. If this example adaptive network is of interest, it is not because of its search capabilities, which are limited and primitive, but because it is a first step toward integrating the learning, search, and use of internal models of the world. The fact that this integration was possible with little specialized machinery—both adaptive elements used have been studied for more primitive purposes, and their interconnection pattern is not a highly restricted one—is a promising sign.

As mentioned earlier, the example network has been so constructed that the action-selecting component (Fig. 5) is completely oblivious to whether it is interacting with the real world or the internal model. The effect of acquiring an internal model is merely that whatever it is with which the action selector is interacting begins to respond more rapidly to contemplated actions and thus

becomes easier to control. When an internal model is viewed from this perspective, it becomes clear that there is an even simpler case of the use of an internal model. In the parlance of animal learning theory, a secondary reinforcer is an originally neutral event that has taken on reinforcing properties by virtue of being predictive of a primary reinforcer (Hilgard & Bower, 1975). To a network receiving this secondary as well as primary reinforcement, the development of the secondary reinforcer means that reinforcement for its actions arrives sooner following the actions than it did previously. The model consists of the rapid simulation of the tendency of the primary reinforcer to follow the secondary reinforcer. This results in an effective environment for internal action-selecting elements that is more amenable to learning techniques. In this way secondary reinforcement can be seen as a very simple case of the construction and use of an internal model.

CONCLUSION

Although the task and network presented here are very simple ones, and there are admittedly some questions as to the direct extensibility of some of the learning algorithms, this little system remains an important demonstration for two reasons. First, what has been sacrificed in the complexity of this demonstration has been to some extent made up in "horizontal" completeness: The network not only has a simple internal model, it also acquires, interrogates, and acts on the basis of that model and performs all three of these functions simultaneously and in an integrated fashion. Second, the system shows how the important ideas of an internal model and internal simulation can be realized in a network or connectionistic form. As such, it can help to form a bridge between cognitive and information-processing models on the one hand, and brain theory and connectionistic models on the other.

ACKNOWLEDGMENTS

The authors would like to thank Michael Arbib for reading and providing valuable detailed criticisms on an earlier draft of this paper. This research was supported by the Air Force Office of Scientific Research and the Avionics Laboratory (Air Force Wright Aeronautical Laboratories) through Contract No. F33615–77–C–1191.

REFERENCES

Amari, S. A mathematical approach to neural systems. In: *Systems neuroscience*. Metzler, J. (Ed.). New York: Academic Press, 1977.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 1977, *85*, 413–451.

Arbib, M. A. *The metaphorical brain*. New York: Wiley-Interscience, 1972.

Bain, A. *The senses and the intellect*, 3rd edition (1st edition 1855). New York: Appleton, 1874.

Barto, A. G., Anderson, C. W., & Sutton, R. S. *Synthesis of non-linear control surfaces by a layered associative search network*. In preparation.

Barto, A. G., & Sutton, R. S. *Goal-seeking components for adaptive intelligence: An initial assessment*. Air Force Wright Aeronautical Laboratories Technical Report AFWAL-TR-81-1070. Avionics Laboratory, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio, 1981a.

Barto, A. G., & Sutton, R. S. Landmark learning: An illustration of associative search. To appear in *Biolulical Cybernetics*, 1981b.

Barto, A. G., & Sutton, R. S. Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. To appear in *Behavioral Brain Research*, 1981c.

Barto, A. G., Sutton, R. S., & Brouwer, P. Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, 1981, *40*, 201–211.

Bledsoe, W. W., & Browning, I. Pattern recognition and reading by machine. *Proceedings Eastern Joint Computer Conference*, 1959, 225–232.

Campbell, D. T. Blind variation and selective survival as a general strategy in knowledge-processes. In: *Self-organizing systems*, 205–231. Yovits, M. C., Cameron, S (Eds.). New York: Pergamon, 1959.

Craik, K. J. W. *The nature of explanation*. Cambridge: Cambridge University Press, 1943.

Cooper, L. N. A possible organization of animal memory and learning. In: *Proceedings of the Nobel symposium on collective properties of physical systems*. Lundquist, B., Lundquist, S. (Eds.). New York: Academic Press, 1974.

Dawkins, R. *The selfish gene*. New York: Oxford, 1976.

Dennett, D. C. Why the law of effect will not go away. In: *Brainstorms*. Montgomery, Vermont: Bradford, 1978.

Galanter, E., & Gerstenhaber, M. On thought: The extrinsic theory. *Psychological Review*, 1956, *63*, 218–227.

Gregory, R. L. On how so little information controls so much behavior. In: *Towards a theoretical biology, 2 sketches*. Waddington, C. H. (Ed.). Edinburgh: Edinburgh University Press, 1969.

Harth, E. Visual perception: A dynamic theory. *Biological Cybernetics*, 1976, *22*, 169–180.

Harth, E., & Tzanakou, E. ALOPEX: A stochastic method for determining visual receptive fields. *Vision Research*, 1974, *14*, 1475–1482.

Hilgard, E. R., & Bower, G. H. *Theories of learning* (fourth edition). Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Johnson-Laird, P. N. Mental models in cognitive science. *Cognitive Science*, 1980.

Klopf, A. H. *Brain function and adaptive systems—A heterostatic theory*. Air Force Cambridge Research Laboratories Research Report AFCRL-72-0164, Bedford, MA, 1972. (A summary appears in: *Proceedings of the International Conference on Systems, Man, and Cybernetics*, IEEE Systems, Man, and Cybernetics Society, Dallas, Texas, 1974).

Klopf, A. H. *The hedonistic neuron: A theory of memory, learning and intelligence*. Washington, D.C.: Hemisphere Publishing Corp., 1981 (to be published).

Klopf, A. H., & Gose, E. An evolutionary pattern recognition network. *IEEE Transactions on Systems Science and Cybernetics*, 1969, *SSC-5*, 3, 247–250.

Kohonen, T. *Associative memory: A system theoretic approach*. Berlin: Springer, 1977.

Longuet-Higgins, H. C. Holographic model of temporal recall. *Nature*, 1968a, *217*, 104.

Longuet-Higgins, H. C. The non-local storage of temporal information. *Proceedings Royal Society (London) B*, 1968b, *171*, 327.

Longuet-Higgins, H. C., Willishaw, D. J., & Buneman, O. P. Theories of associative recall. *Review Biophys.*, 1970, *3*, 223–244.

Mach, E. On the part played by accident in invention and discovery. *Monist*, 1896, *6*, 161–175.

MacKay, D. M. Internal representation of the external world. AGARD symposium on natural and artificial logic processors, Athens, mimeographed, 14 pages, 1963.

Mendel, J. M., & McLaren, R. W. Reinforcement-learning control and pattern recognition systems. In: *Adaptive, learning, and pattern recognition systems: Theory and applications*, 287–317. Mendel, J. M., Fu, K. S. (Eds.). New York: Academic Press, 1970.

Michie, D., & Chambers, R. A. BOXES: An experiment in adaptive control. *Machine Intelligence 2*, 137–152. Dale E., Michie, D. (Eds.). Edinburgh: Oliver and Boyd, 1968.

Miller, G. A., Galanter, E., & Pribram, K. H. *Plans and the structure of behavior*. New York: Henry Holt and Co., 1960.

Minsky, M. L. Steps toward artificial intelligence. *Proceedings IRE*, 1961, *49*, 8–30.

Nakano, K. Associatron—a model of associative memory. *IEEE Transactions on Systems, Man, Cybernetics*, 1972, *SMC-2*, 3, 380–388.

Palm, G. On associative memory. *Biological Cybernetics*, 1980, *36*, 19–31.

Piaget, J. *The construction of reality in the child*. New York: Basic Books, 1954.

Poggio, T. On optimal nonlinear associative recall. *Biological Cybernetics*, 1975, *19*, 201–209.

Poincaré, H. L'invention mathematique. *Bull. Inst. Gen. Psychol.*, 1908, *8*, 175–187.

Poincare, H. Mathematical creation. In: *The foundations of science*, Poincare, H., ed. New York: Science Press, 1913.

Raphael, B: The thinking computer: Mind inside matter. San Francisco: Freeman, 1976.

Sacerdoti, E. D. *A structure for plans and behavior*. New York: Elsevier, 1977.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal Research and Development*, 1959, *3*, 210–229.

Selfridge, O. G. Pattern recognition and modern computers. In: *Proceedings of the 1955 Western Joint Computer Conference*, Session on Learning Machines, W. H. Ware Chairman, 91–93, 1955.

Simon, H. A. *The sciences of the artificial*. Cambridge, Mass.: The MIT Press, 1969.

Sommerhoff, G. *Logic of the living brain*. New York: Wiley, 1975.

Sutton, R. S., & Barto, A. G. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 1981, *88*, 2, 135–170.

Tolman, E. C., & Honzik, C. H. "Insight" in rats. *Univ. Calif. Publ. Psychol.*, 1930, *4*, 215–232.

Tsetlin, M. L. *Automaton theory and modeling of biological systems*. New York: Academic Press, 1973.

Uhr, L., & Vossler, C. A pattern recognition program that generates, evaluates and adjusts its own operators. *Proceedings Western Joint Computer Conference*, 1961, 555–569.

Wigstrom, H. A neuron model with learning capability and its relation to mechanisms of association. *Kybernetic*, 1973, *12*, 204–215.

Willshaw, D. J., & Buneman, O. P., Longuet-Higgins, H. S. Non-holographic associative memory. *Nature*, 1969, *222*, 960–962.

Wood, C. C. Variations on a theme by Lashley: Lesion experiments on the neural model of Anderson, Silverstein, Ritz, and Jones. *Psychological Review*, 1978, *85*, 582–591.

APPENDIX A
DETAILS OF THE SIMULATION EXPERIMENTS

A.1 Computation of Movement

At each time step the simulated adaptive network provides an instantaneous action vector $[A_R(t), A_L(t)]$, the first component of which gives the network a tendency to turn right, the second to turn left. The computation of this vector is detailed in Appendix B. A record $[S_R(t), S_L(t)]$ is kept of the extent to which each action has been instantaneously selected recently:

$$S_R(t) = \alpha \cdot S_R(t - 1) + (1 - \alpha) \cdot A_R(t)$$
$$S_L(t) = \alpha \cdot S_L(t - 1) + (1 - \alpha) \cdot A_L(t)$$

Movement is determined by which of these traces is largest:

$$\text{Motion } (t) = \beta \cdot [S_R(t) - S_L(t)],$$

where positive motion means motion to the right, and negative motion means motion to the left. In all the simulation experiments α was set at 0.8 and β was set at 50.0 where the distance between the barriers at B and C is about 125.

If the motion computed in the foregoing causes the network to run into a barrier, the actual motion is halted at the point of contact. In addition, barrier collision neutralizes the inertial tendency to continue motion in that direction. Specifically, the inertial traces S_R and S_L are set to their average upon collision with a barrier. The inertia was also neutralized by setting both of these traces to zero each time a subject was "picked up" and moved as part of an experiment.

A.2 Experiment I

Experiment 1 used 200 subjects, each run individually through the following three phases.

A.2.1 Exploration Phase

Each subject was released between the lower large colored regions (point A in Fig. 4). If the center of its body passed into a colored region, the corresponding sensory input line was set to a value of approximately 0.5. All motion was computed as described previously. After 1000 time steps the association phase began.

A.2.2 Association Phase

Each subject was moved to the enclosed red region of Fig. 4. The red input line was activated in the same way it was activated during the exploration phase when the subject was within the lower red region. After two time steps the reward input line was also set to 0.5. After one time step of this stimulation pattern, each subject was transferred to the enclosed green box shown in Fig. 4. The input pattern there was $I_{\text{red}} = 0.0$, $I_{\text{green}} = 0.5$, and $I_{\text{reward}} = 0.5$. After two time steps of this, the reward input line was set to zero again for one time step, and then the testing phase began. The following table summarizes the stimulation regime during the association phase.

Absolute Time	Duration	I_{red}	I_{green}	I_{reward}
1000–1001	2	0.5	0.0	0.0
1002	1	0.5	0.0	0.5
1003–1004	2	0.0	0.5	0.5
1005	1	0.0	0.5	0.0

A.2.3 The Testing Phase

In the testing phase of the primary experiment each subject was returned to location A of Fig. 4 and released, just as in the exploration phase. The testing phase ended when either of the two colored regions was entered. Of the 200 subjects, 141 entered the red region first and 59 entered the green region first. This result is statistically significant at at least the $p = .005$ level.

A.3 Experiment II

The second experiment was identical to the first during the exploration and association phases. Its testing phase differed in that the lower red and green regions and the barriers inside them were removed. After 300 time steps the testing phase ended and the position of the subject was recorded. All 100 subjects had moved very far to the right after the 300 time steps, the nearest being about twice as far off the page as the distance from A to the right edge of the page in Fig. 4.

A.4 Experiment III

The third experiment was identical to the first, except that the bias weights WAC_R and WAC_L (also called B_i in the text) were set to zero at the beginning of the testing phase. This ensured the networks had no initial tendency to move either right or left at the beginning of the testing phase. Of the 100 subjects, 71 entered the red area first, a result statistically significant at at least the $p = .005$ level.

APPENDIX B
DETAILS OF THE SIMULATED ADAPTIVE NETWORK

STIMULI is the set {RED, GREEN, REWARD}
ACTIONS is the set {RIGHT, LEFT}
This is a discrete time model (i.e., $t = 0, 1, 2, \ldots$).

B.1 Components

A PREDICTOR-MODULE, consisting of 3 PREDICTOR-ELEMENTS (corresponding to the 3 stimuli), a 3×3 matrix of PREDICTOR-TO-PREDICTOR-CONNECTIONS, and a 3×2 matrix of ACTOR-TO-PREDICTOR-CONNECTIONS.
An ACTION-SELECTING-MODULE, consisting of 2 ACTOR-ELEMENTS and a 2-element vector of CONSTANT-TO-ACTOR-CONNECTIONS.
A vector of 3 INPUT-LINES, corresponding to the three STIMULI.
A vector of two OUTPUT-LINES, corresponding to the 2 ACTIONS.

B.2 Descriptive Variables

B.2.1 Input Variables

$I_s(t) \in [0, 1]$, for all s $\in$ STIMULI, is the input to the network at time t. These indicate the color of the region that the subject is in (if any) and the presence or absence of reward.

B.2.2 Output Variables

$A_a(t) \in [0, 1]$, for all a $a \in$ ACTIONS, is the activity level at time t of the ACTOR-ELEMENT for action a, indicating the instantaneous selection of movement to the right or left.

B.2.3 State Variables

$P_s(t) \in [0, 1]$, for all $s \in$ STIMULI, is the activity level at time t of the PREDICTOR-ELEMENT for stimulus s. This indicates a combination of prediction of stimulation and actual stimulation.

$WPP_{s1,s2}(t) \in \mathbf{R}$, for all $s1$, $s2 \in$ STIMULI, is the efficacy of the PREDICTOR-TO-PREDICTOR-CONNECTION to the PREDICTOR-ELEMENT for stimulus $s1$ from the PREDICTOR-ELEMENT for stimulus $s2$.

$WPA_{s,a}(t) \in \mathbf{R}$, for all $s \in$ STIMULI, $a \in$ ACTIONS, is the efficacy at time t of the ACTOR-TO-PREDICTOR-CONNECTION to the PREDICTOR-ELEMENT for stimulus s from the ACTOR-ELEMENT for action a.

$WAC_a(t) \in \mathbf{R}$, for all $a \in$ ACTIONS, is the efficacy at time t of the CONSTANT-TO-ACTOR-CONNECTION to the ACTOR-ELEMENT for action a. These weights were called the bias weights and denoted B rather than WAC in the text.

$\bar{A}_a(t) \in [0, 1]$, for all $a \in$ ACTIONS, is the trace at time t of $A_a(t)$, the activity of the ACTOR-ELEMENT for action a.

$\bar{A}'_a(t) \in [0, 1]$, for all $a \in$ ACTIONS, is another trace at time t of $A_a(t)$.

$\bar{P}_s(t) \in [0, 1]$, for all $s \in$ STIMULI, is the trace at time t of $P_s(t)$, the activity of the PREDICTOR-ELEMENT for stimulus s.

B.2.4 Parameters

$CPP_{s1,s2} \in \mathbf{R}+$, for all $s1$, $s2 \in$ STIMULI, is the learning rate parameter for the PREDICTOR-TO PREDICTOR-CONNECTION from the PREDICTOR-ELEMENT for stimulus $s2$ to the PREDICTOR-ELEMENT for stimulus $s1$.

$CPA_{s,a}(t) \in \mathbf{R}+$, for all $s \in$ STIMULI, $a \in$ ACTIONS, is the learning rate parameter for the ACTOR-TO-PREDICTOR-CONNECTION from the ACTOR-ELEMENT for action a to the PREDICTOR-ELEMENT for stimulus s.

$C \in \mathbf{R}+$, is the learning rate parameter for the CONSTANT-TO-ACTOR-CONNECTION to the ACTOR-ELEMENT for action a.

$\alpha_P \in [0, 1]$ is the trace decay parameter for the trace of activity in the PREDICTOR-ELEMENTS.

$\alpha_A \in [0, 1]$ is the trace decay parameter for the trace of activity in the ACTOR-ELEMENTS that is used to change the ACTOR-TO-PREDICTOR-CONNECTION efficacies.

$\alpha_A' \in [0, 1]$ is the trace decay parameter for the trace of activity in the ACTOR-ELEMENTS that is used to change the CONSTANT-TO-ACTOR-CONNECTION efficacies.

B.3 Equations of Interaction

B.3.1 Equations of Primary Network Operation

$$A_R(t) = F\{A'_R(t) - A'_R(t)\}$$
$$A_L(t) = F\{A'_L(t) - A'_L(t)\}$$

where

$$A'_a(t) = \begin{cases} WAC_a(t) + \text{NOISE} & \text{if } WAC_a + \text{NOISE} > 0 \\ 0 & \text{else} \end{cases}$$

for all $a \in$ ACTIONS; $F()$ given by Equation 1; and NOISE a normally distributed random variable with mean 0.2 and standard deviation 0.4.

$$P(t) = F[I(t) + WPA(t) \cdot A(t) + WPP(t) \cdot P(t - 1)]$$
(using vector and matrix notation)

B.3.2 Equations for Change of Connection Efficacies

$$WPP_{s1,s2}(t + 1) = WPP_{s1,s2}(t) + CPP_{s1,s2} \cdot \{P_{s1}(t) - \bar{P}_{s1}(t)\} \cdot \bar{P}_{s2}(t - 1)$$

$$WPA_{s,a}(t + 1) = WPA_{s,a}(t) + CPA_{s,a} \cdot \{P_s(t) - \bar{P}_s(t)\} \cdot \bar{A}'_a(t)$$

$$WAC_a(t + 1) = WAC_a(t) + C \cdot \{P_{\text{reward}}(t) - \bar{P}_{\text{reward}}(t)\} \cdot \bar{A}_a(t)$$

where

$$\bar{P}_s(t + 1) = \alpha_P \bar{P}_s(t) + (1.0 - \alpha_P) \cdot P_s(t)$$
$$\bar{A}_a(t + 1) = \alpha_A \cdot \bar{A}_a(t) + (1.0 - \alpha_A) A_a(t)$$
$$\bar{A}_a(t + 1) = \alpha_A \cdot \bar{A}_a(t) + (1.0 - \alpha_A) A_a(t)$$

For all $a \in$ ACTIONS and s, $s1$, $s2 \in$ STIMULI.

B.4 Parameter Settings

CPP_{S_1,S_2}:

$S_2 \backslash S_1$	RED	GREEN	REWARD
REWARD	0.5	0.5	0.0
GREEN	0.5	0.0	1.5
RED	0.0	0.5	1.5

$CPA_{s,a}$:

a \ s	RED	GREEN	REWARD
RIGHT	0.2	0.2	0.0
LEFT	0.2	0.2	0.0

$C = 0.5$
$\alpha_P = 0.0$
$\alpha_A = 0.0$
$\alpha'_A = 0.8$