7 Neural Problem Solving

Andrew G. Barto
Richard S. Sutton
Department of Computer and Information Science
University of Massachusetts

Neural models, and indeed models in any domain, can differ widely in terms of the intentions with which they are constructed and the levels of empirical support on which they depend. A neural model might be based on detailed observations of a particular experimental preparation, or it may be less directly related to anatomical and physiological data, relying instead on behavioral parallels. As neural models become farther removed from anatomy and physiology and closer to "adaptive networks" or "self-organizing systems" of quasi-neural elements, they become less interesting to the neuroscientist, and the term "neural model" becomes more misleading. With this decreasing relevance to neuroscience, however, one might hope for increasing relevance to psychology and perhaps to artificial intelligence. Yet many such models have not been influential among psychologists despite the rich history in psychology of purely descriptive behavioral models, and they have not been influential among artificial intelligence researchers despite the fact that these researchers have explicitly excluded concern with neural mechanisms. One reason for this may be the fact that many abstract neural models neither make significant contact with behavioral data nor suggest algorithms that would be useful to the artificial intelligence researcher for solving nontrivial problems.

In this article, we present an overview of a research program that is intended explicitly to be a study of "adaptive networks" of quasi-neural elements. However, we have tried to maintain careful and significant contact with behavioral data from animal learning studies, with descriptive behavioral models in that field, and with problem-solving methods of artificial intelligence. Although the mechanisms we discuss can be given neural in-

terpretations, we feel that it is premature to propose an extensive and detailed neural model to bridge the gap between anatomical and physiological data and the behavioral level in which we are interested. We have instead concentrated on behavioral models that exhibit aspects of animal behavior that we consider to be adaptively significant, and on the relationship between these aspects of behavior and the computational requirements for solving nontrivial problems. We are considering problems that animals are capable of solving routinely, whose solutions provide obvious adaptive advantages, and that are genuinely difficult to solve irrespective of the methods used.

Our approach is to consider the general problem of *control.* Arbib (1972) emphasizes that: "the animal perceives its environment to the extent that it is *prepared* to *interact* with that environment in some reasonably structured fashion." This stress on what Arbib calls "action oriented perception" implies that modeling approaches are misleading insofar as they consider just sensory processing (e.g., pattern recognition), while neglecting highly structured action generation processes and the closed-loop interaction, mediated by the organism's environment, between action and sensory patterns. From an engineering point of view, we can say that animals are engaged in the problem of controlling their environments in a closed-loop fashion to achieve certain goals. Consequently, our strategy has been to consider entire control systems facing control problems posed by environmental interaction, and we have paid as much attention to the environments and the resulting control problems as we have to the controlling mechanisms themselves.

In addition to our emphasis on complete control problems, we have found it useful to endow each network component with relatively sophisticated computational power. Each primitive component of a network in our approach is best characterized as a complete, although simple, "reinforcement learning control system" (Mendel & McLaren, 1970) that acquires knowledge about feedback pathways in which it is embedded and uses this knowledge to seek preferred inputs. In providing each component with such capabilities, we have been guided by the proposal of A. H. Klopf (1972, 1979, 1982) that progress in understanding natural intelligence, and progress in artificial intelligence, might be furthered by a study of goal-seeking systems composed of goal-seeking components. Instead of viewing any form of goal-seeking behavior as an emergent property of a system consisting of non-goal-seeking components, Klopf suggests that sophisticated goal-directed behavior arises from interacting components that are self-interested, and exercise strategies for furthering these self-interests. Goal-directed behavior is pushed down the structural hierarchy to basic levels, and higher forms of goal-directed behavior are seen as resulting from the competitive and cooperative interaction of self-interested components. The neural interpretation of this hypothesis is that neurons are similarly sophisticated goal-seeking control systems. In the course of our discussion, we point out similarities between our adaptive elements and goal-seeking strategies known to exist in single-celled organisms such as bacteria. We think that the continued study of the numerous commonalities between bacterial chemotaxis and other simple forms of adaptive behavior in single cells, and the signaling systems of neurons (Koshland, 1979) is a most promising avenue for assessing the hypothesis that neurons are goal-seeking control systems. However, although we present our learning algorithms in terms of neuronlike elements, we are not prepared to argue that all the capabilities of these elements need necessarily reside at the level of single cells.

This article is divided into three major parts. In the first part, we discuss a neuronlike adaptive element that is capable of reproducing some of the details of animal behavior in classical conditioning experiments. We emphasize aspects of classical conditioning that are difficult to achieve by neural models proposed in the past and that seem to have obvious adaptive significance; in particular, we emphasize temporal phenomena involving prediction and expectation. This adaptive element resulted from our attempts to incorporate the sensitivity to temporal succession that seems necessary for goal-seeking control: If actions are to be selected according to their consequences, then temporal factors are important because an action's consequences unfold over time. This adaptive element is not, however, capable of closed-loop control and is not a goal-seeking system in the appropriate sense. In the second part of this article, another type of adaptive element is discussed that is a goal-seeking learning control system closely related to instrumental, rather than to classical, conditioning. We discuss associative networks composed of these elements, how their capabilities differ from associative memories studied in the past, and why these differences are important from a problem-solving perspective. We illustrate the learning capabilities of these networks in several spatial learning tasks. Finally, in the third section, we discuss how the open-loop classical conditioning element and the closed-loop goal-seeking element can interact to provide an approach to a fundamental problem of adaptive system theory known as the "assignment of credit problem": If reward is achieved after a complex series of actions, to which component actions should the credit be assigned (or the blame in the case of penalty)?

ANALOGS OF CLASSICAL CONDITIONING

Many adaptive network theories are based on neuronlike adaptive elements that can behave as single unit analogs of animal classical conditioning (e.g., elements incorporating Hebb's, 1949, postulate). However, there are many features of animal behavior in classical conditioning experiments that are generally not preserved by adaptive element analogs. Although one may validly question the rationale for investigating networks of elements that are *ex-*

act analogs of overt animal associative learning behavior (as surely some properties of this behavior must be due to the effects of higher levels of organization), it seems reasonable to include those characteristics that are most salient in terms of adaptive significance, that are problematic to achieve as emergent properties of organizations of simpler components, and that offer advantages from a problem-solving point of view. Here we describe an adaptive element analog of classical conditioning that preserves features of the anticipatory nature of classical conditioning and is in agreement with data regarding the effects of stimulus context in classical conditioning. We show that these stimulus context effects can be interpreted as the capability to "orthogonalize" input vectors. The element is a temporally refined extension of the Rescorla–Wagner model of classical conditioning (Rescorla & Wagner, 1972) and was presented by Sutton and Barto (1981b) and further discussed by Barto and Sutton (1982).

In a simple classical conditioning experiment, the subject is repeatedly presented with a neutral conditioned stimulus (CS), that is, a stimulus that does not cause responses other than orienting responses, followed after an interval of time (the interstimulus interval, or ISI) by an unconditioned stimulus (UCS), which reflexively causes an unconditioned response (UCR). After a number of such pairings of the CS and the UCS–UCR, the CS comes to elicit a response of its own, the conditioned response (CR), which closely resembles the UCR or some part of it. For example, a dog is repeatedly presented with first the sound of a bell (the CS) and then food (the UCS), which causes the dog to salivate (the UCR). Eventually, just the sound of the bell causes salivation (the CR). This description leaves much unsaid, as we see later, but will suffice as we describe an adaptive element analog.

Fig. 7.1 shows an adaptive element with an input pathway for the UCS and one for each stimulus capable of being associated with the UCS. These latter stimuli are (potential) conditioned stimuli, and we denote them by CS_i, $1 \leq i \leq n$. Let $x_0(t)$ denote the activity of the UCS pathway at time t, and let $x_i(t)$ denote the activity of pathway CS_i, $1 \leq i \leq n$, at time t. The element's output is assumed to contribute to both the UCR and the CR. For our present purposes, we assume that these activity levels at any time are positive real numbers. The associative strength of each CS at time t with respect to the UCS is denoted by $V_{CS_i}(t)$, $1 \leq i \leq n$, and represents the efficacy, or weight, of the corresponding input pathway. The weight of the UCS pathway is fixed at some constant value that we denote by λ. Let $s(t)$ denote the weighted sum of all the inputs at time t, that is,

$$s(t) = \lambda x_0(t) + \sum_{i=1}^{n} V_{CS_i}(t) x_i(t). \tag{1}$$

For our present purposes, it does not matter exactly how the element output is computed, and for simplicity, we assume that at time t it is just $s(t)$.

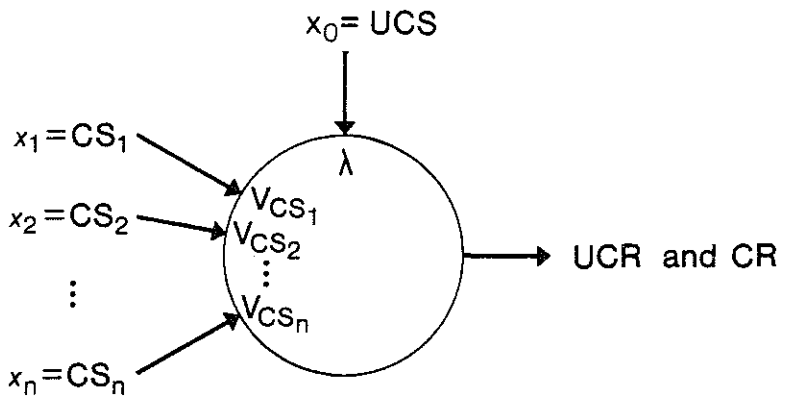

FIG. 7.1 An adaptive element analog of classical conditioning. (Reprinted from Barto & Sutton, 1982).

Several other variables are required in order to define the adaptive element. For each stimulus signal x_i, $1 \leq i \leq n$, we require a separate stimulus trace that we denote by $\bar{x}_i$. By this we mean that activity of variable x_i is reflected in later activity of variable $\bar{x}_i$. This is accomplished by letting $\bar{x}_i(t)$ be a weighted average of the values of x_i for some period of time preceding t. Similarly, we require a trace of the sum s. Let $\bar{s}(t)$ denote a weighted average of the values of s over some interval preceding t. In the computer simulations described as follows, we generated these traces using the first-order linear difference equations

$$\bar{x}_i(t+1) = \alpha \bar{x}_i(t) + (1-\alpha) x_i(t)$$
$$\bar{s}(t+1) = \beta \bar{s}(t) + (1-\beta) s(t)$$

where α and β are constants such that $0 \leq \alpha, \beta < 1$. This process produces exponentially decaying traces with time constants depending on the parameters α and β (Fig. 7.2).

In terms of the two variables s and $\bar{s}$, and the variables x_i, $\bar{x}_i$, and V_{CS_i} for each pathway $1 \leq i \leq n$, the adaptive element successively generates values of the associative strengths, or weights, as follows: for each i, $1 \leq i \leq n$,

$$V_{CS_i}(t+1) = V_{CS_i}(t) + c[s(t) - \bar{s}(t)]\bar{x}_i(t) \tag{2}$$

where c is a positive constant determining the rate of learning.

The process specified by Equations 1 and 2 can be described as follows: Activity on any input pathway i possibly causes an immediate change in the element output s (we have assumed, again for simplicity, that there is no delay through the element) and also causes that pathway to be "tagged" by the

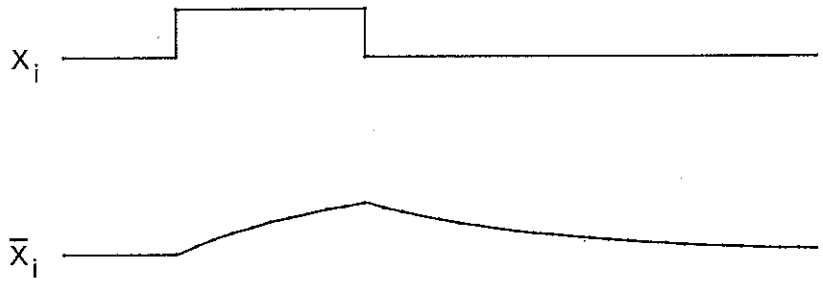

FIG. 7.2 An exponentially decaying stimulus trace. Activity of variable x_i causes a prolonged trace $\bar{x}_i$.

stimulus trace $\bar{x}_i$ as being "eligible" for modification for a certain period of time (the duration of the trace) after the activity on pathway i ceases. A weight is modified only if it is eligible and the current value of s differs from the value of the trace $\bar{s}$ of s. The simplest case, and the one used in our simulations, results from letting $\bar{s}(t) = s(t-1)$ so that $s(t) - \bar{s}(t) = s(t) - s(t-1)$, which is a discrete form of the rate of change of s.

The notion that one set of conditions makes pathway efficacies "eligible" for modification, but that actual modifications occur due to other conditions during periods of eligibility, is a major feature of Klopf's (1972, 1982) theory of neural adaptation. This notion itself is not uncommon among theorists, but the idea of two separate variables, one for signaling the occurrence of events and another for retaining knowledge of these occurrences so that events can be associated with later events, has not been deeply explored. In many neural theories, for example, neural discharges signal the occurrence of stimulus events and also bridge the temporal gap required for conditioning by "reverberating" in some manner. Because it seems advantageous for an organism to be able to perceive events as occurring as closely as possible to their actual times of occurrence, and particularly as early as possible, additional mechanisms must be postulated to distinguish neural activity that is signaling the occurrence of an event from reverberatory neural activity that is storing reflections of past events. In a two-variable system (e.g., x_i and $\bar{x}_i$) these two functions are cleanly separated. Although reverberatory activity is probably important at many levels in the central nervous system, one need not assume that reverberation is the primary mechanism at all levels for spanning the time between the sequential events on which learning depends. We now examine several aspects of our adaptive element's behavior with respect to classical conditioning data and suggest how these aspects of behavior are important from the perspective of problem solving.

Anticipatory Nature of Classical Conditioning

The interval between CS onset and CR onset is called the *CR latency*. For a particular response there is a positive minimum CR latency due to various types of intrinsic delays (e.g., 70–80 msec for rabbit nictitating-membrane response). If the ISI in a conditioning experiment is shorter than the minimum CR latency, then the CR necessarily begins after UCS onset. More usually, however, the ISI is longer than the minimum CR latency, and the CR begins before the UCS onset (although the CR latency tends to change during conditioning procedures [see, for example, Kimmell, 1965]). Being a response to the predictive CS, the CR *anticipates* the UCS and the UCR (Gormezano, 1972; Mackintosh, 1974).

This anticipatory aspect of the CR is a crucial factor in the adaptive significance of the behavior elicited in classical conditioning experiments. Putting on the hat of a designer of an intelligent problem solver, it would seem desirable to have a mechanism that is able to extract predictive regularities in its input so as to make a representation of a predicted event occur at the earliest time at which that event can be predicted with reasonable certainty. A prediction that is available only at the same time as, or later than the event predicted is no more useful in guiding behavior than no prediction at all; and, assuming a competitive environment, the earlier the prediction is available, the better. Moreover, internal predictive representations might act as predictive cues for other internal events, creating the possibility for effectively "compressing" the time scale in a manner similar to what would happen if we were to tape-record something at one tape speed ("real time") and play it back at a higher speed ("faster than real time"). The utility of predictive methods is well established in engineering applications (Box & Jenkins, 1976), and the adaptive advantage to an organism possessing these capabilities is clear.

The fact that anticipatory CRs are possible at all is problematic for many neural theories. For example, many mathematical interpretations of the Hebbian postulate require simultaneous pairing of the UCS and CS signals at the adaptive element, thus implying an optimal ISI of zero. Because the dependency of conditioning on the ISI is generally recognized, delays in the CS pathway are often suggested to bring the behavior closer to animal data (Burke, 1966; Uttley, 1979). Such delays can be used to reproduce the experimental observation that the CS onset must precede the UCS onset in order for conditioning to occur, but they do not by themselves explain the experimental observation that the CR onset generally also occurs before the UCS onset. Such delays necessarily delay CR onset at least until the time of UCS onset, thereby preventing the CR latency from ever being shorter than the ISI required for conditioning. Reverberatory trace mechanisms in the CR

pathway are more satisfactory, but they do not allow for precise temporal localization of the CS.

Let us examine the behavior of the aforementioned adaptive element for a special case of classical conditioning in which the CS and the UCS are rectangular pulses, the CS associative strength is initially zero, and the trace $\bar{s}$ takes the form $\bar{s}(t) = s(t - 1)$. Figure 7.3a shows the adaptive element analog of

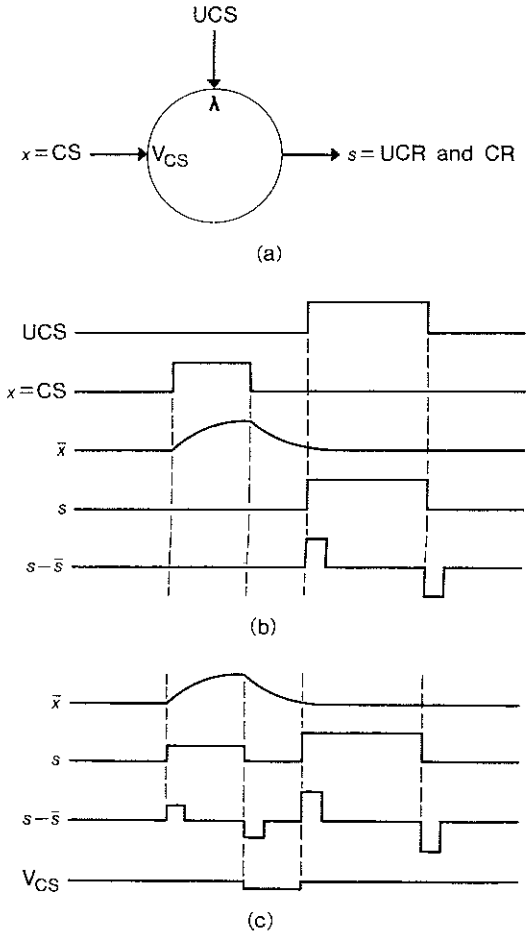

FIG. 7.3 Analog of classical conditioning with a single CS. (a) The adaptive element configuration. (b) Time courses of the model variables for the first trial. (c) Time courses of the model variables after complete conditioning. Note that the element response (s) to the CS anticipates the UCS. (Reprinted from Barto & Sutton, 1982).

this situation, and Figs. 7.3b and 7.3c show the signal time courses we now describe. On the first trial, the occurrence of the CS causes an increase in the eligibility $\bar{x}$ of the CS pathway that persists for some time after CS offset. When the UCS occurs, it causes a positive change in s at its onset and an equal but negative change at its offset. Because eligibility $\bar{x}$ is greater at the time of UCS onset than at the time of UCS offset, V_{CS} is caused to have a net increase: It increases at UCS onset and decreases by a lesser amount at UCS offset (Fig. 7.3b).

On the second trial, V_{CS} is no longer zero so that CS occurrence causes changes in s in addition to those caused by UCS occurrence (Fig. 7.3c). The increase in s at CS onset has no effect on V_{CS} because eligibility is zero (we are assuming that the intertrial interval is long enough to let eligibility decay to zero between trials). The decrease in s at CS offset, however, occurs during high eligibility and therefore causes a decrease in V_{CS}. The UCS causes an increase in V_{CS} as on the first trial, but the net result of both the CS and UCS is less of an increase than on Trial 1. With additional trials, V_{CS} increases until the positive effect of the UCS is counterbalanced by the negative effect of the CS offset. The process therefore stabilizes in the sense that eventually V_{CS} will experience no net change per trial (although it will in general continue to change during these trials). Stability is achieved through negative feedback due to increases in V_{CS} causing increased decreases in V_{CS} at CS offset. Figure 7.4, Trials 0–10, shows a typical acquisition curve plotting the associative strength after each trial[1].

Fig. 7.3c shows that the value of s shows a response to the CS and the UCS. The later response is assumed to contribute to the UCR whereas the earlier one is assumed to contribute to the CR. Thus, the CR component anticipates the UCS onset and the UCR onset. Here, the CR latency is zero because we have assumed that there is no delay in the input/output response of the element, but the ISI must be positive for conditioning to occur. The basis for this anticipatory behavior is clearly the prolonged eligibility trace. If an event regularly precedes another event by an amount of time spannable by this trace's duration, then the association between these events can be "recorded," in a sense, by the adaptive element and "played back" at a much faster time scale.

The adaptive element is also capable of doing something more subtle than this. Because activity on any input pathway with nonzero weight causes changes in s, this activity can cause changes in the weights of other pathways. Thus, a previously conditioned CS can act as a UCS for a second CS. This also can occur in a Hebbian element analog of classical conditioning, but

[1]This acquisition curve is strictly negatively accelerated whereas experimental acquisition curves generally have an initial period of positive acceleration. Extensions of models similar to our adaptive element have been proposed to remedy this (Frey & Sears, 1978).

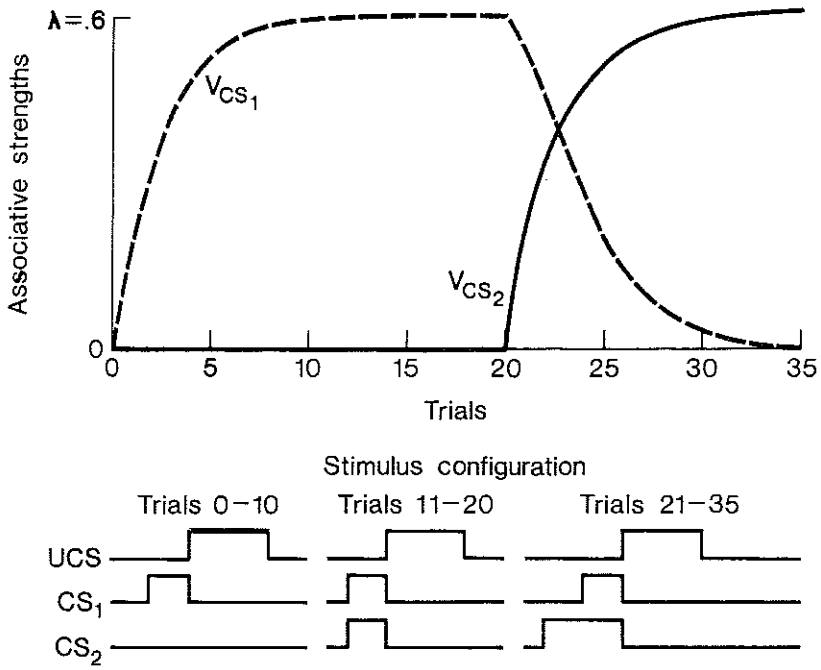

FIG. 7.4 Analog of a blocking experiment. Trials 0–10: Acquisition curve for a single CS in Part I of the blocking experiment. Trials 11–20: Part II of the blocking experiment. CS_1 and CS_2 are identically paired; the formation of an association from CS_2 is blocked due to prior conditioning to CS_1. Trials 21–35: Interaction of stimulus context and anticipatory effects. Blocking is reversed because CS_2 is an earlier predictor of UCS occurrence. (Reprinted from Barto & Sutton, 1982).

when coupled to the anticipatory capabilities of our element, some novel consequences appear. Figure 7.5 shows a simulated experimental arrangement in which each trial consists of a temporal sequence of four CSs (i.e., a serial compound CS) followed by a UCS. Only the CS that occurs immediately before the UCS (i.e., CS_1) initiates an eligibility trace that reaches far enough into the future to permit conditioning to occur. At first, then, only the associative strength of CS_1 increases. As an association from CS_1 is forming, however, CS_1 occurrence causes changes in s and thereby acts as a UCS for the preceding CS, that is, for CS_2. In turn, CS_2 acts as a UCS for CS_3, etc. Figure 7.5 shows the acquisition curves of this higher-order conditioning process. During this process, the CR onset moves back in time from the time of CS_1 onset to the earlier time of CS_4 onset. Kehoe, Gibbs, Garcia, and Gormezano (1979) observed a strong effect of this nature for rabbit nictating-membrane response (see also Gormezano & Kehoe, in press). Chaining of associations in

this manner (by a single element) permits conditioning to occur for ISIs much longer than those that can be spanned by a single eligibility trace, provided there are regularly occurring intervening events. Under such conditions, the anticipatory CR will tend to begin at the earliest time at which the UCS can be predicted with reasonable certainty irrespective of the eligibility trace duration. We discuss the significance of this capability from a problem-solving perspective in more detail in a later section ("Assignment of Credit").

Stimulus Context Effects and Orthogonalization

In classical conditioning experiments the associative strength of the stimuli that act as context for a CS on a trial can nullify or even reverse the effect of the occurrence of the UCS on that trial. In this section, we discuss two examples of stimulus context effects, known as *blocking* and *conditioned inhibition*, and show how the adaptive element already described is able to produce these effects. We then explain this by relating our element to the Rescorla-Wagner model of classical conditioning and discuss the significance of this behavior from a problem-solving point of view. In particular, we observe that the stimulus context effects that animals exhibit can be inter-

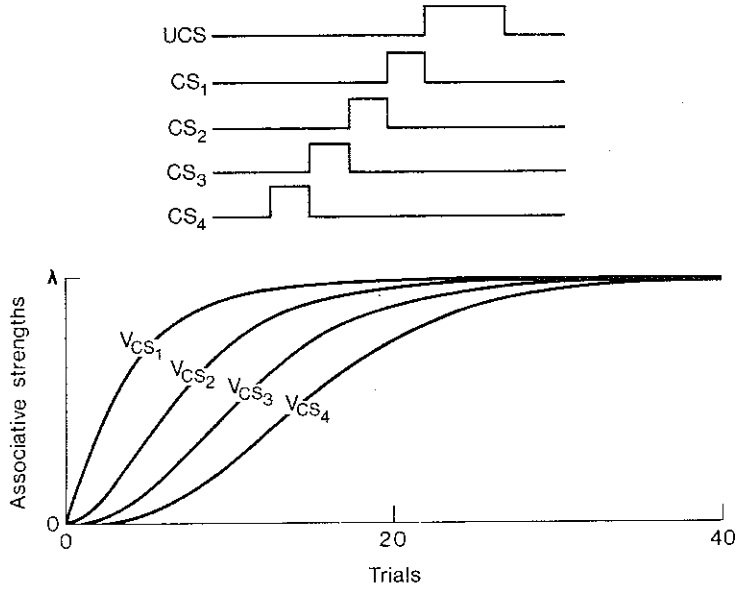

FIG. 7.5 Serially compound CSs. As their associative strengths increase, later CSs serve as UCSs for earlier CSs. (Reprinted from Barto & Sutton, 1982).

preted as the result of a process that "orthogonalizes" stimulus vectors, a process of considerable practical importance.

A typical blocking experiment consists of two parts. In Part I, one stimulus, CS_1, is paired with the UCS at an appropriate ISI until the associative strength between CS_1 and the UCS reaches its asymptotic value. In Part II, CS_1 continues to be paired with the UCS, but another stimulus, CS_2, co-occurs with CS_1. Although CS_2 is appropriately paired with the UCS in Part II, it conditions very poorly, if at all, compared to a control group lacking prior Part I conditioning to CS_1 (see, for example, Hilgard & Bower, 1975). The results of a simulation of blocking using our adaptive element are illustrated in Trials 0–20 of Fig. 7.4. For the first 10 trials, CS_1 was presented alone and followed by the UCS, and for Trials 11–20, CS_2 was presented identically paired with CS_1, and both were followed by the UCS. During trials 11–20, changes in V_{CS_2} were blocked because s did not change while the CS_2 pathway was eligible.

Conditioned inhibition is another stimulus context effect involving at least two CSs, denoted CS+ and CS−. Suppose the occurrence of CS+ alone is always followed by the UCS, but the co-occurrence of CS+ and CS− is never followed by the UCS. For this paradigm, the associative strength V_{CS+} increases so that CS+ produces a CR, but V_{CS-} becomes negative so that a CR does not follow the co-occurrence of CS+ and CS−; CS− becomes a conditioned inhibitor of the CR. Figure 7.6 shows the results of a simulation of this procedure using our adaptive element.

Perhaps the best way to explain how our adaptive element produces these effects is to relate it to the Rescorla–Wagner model, which was devised to describe these effects in animal behavior (Rescorla & Wagner, 1972). The Rescorla–Wagner model is based on the view that learning occurs only when expectations are violated. According to this view, for example, blocking occurs because Part I training creates an expectation of the UCS that is not disrupted in Part II. When the activity trace $\bar{s}$ in Equation 2 is interpreted as providing the expected value of the actual activity s, then Equation 2 resembles the Rescorla–Wagner model because it implies that eligible pathways are modified whenever the actual value of s differs from the expected value $\bar{s}$. The term $s - \bar{s}$ is a measure of how strongly the current activity confirms or contradicts the previously formed expectation. Sutton and Barto (1981b) discuss the Rescorla–Wagner model and these correspondences in detail. Anticipatory aspects of classical conditioning and ISI dependency are not addressed by the Rescorla–Wagner model because, unlike our element, it is a trial-level model that does not distinguish between different times within each trial.

It is a striking fact that the Rescorla–Wagner model, which was formulated to describe compactly a wide variety of effects observed in animal learning experiments, is identical to an algorithm for iteratively computing the inverse

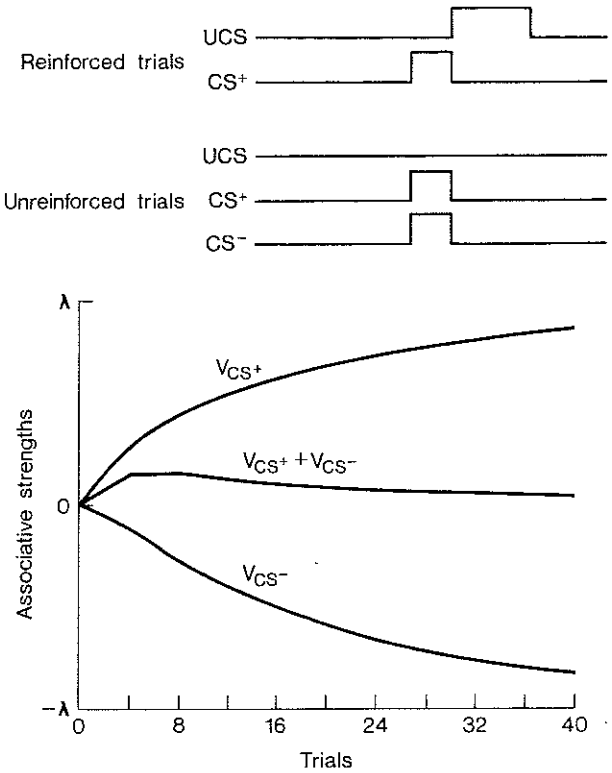

FIG. 7.6 Analog of a conditioned inhibition experiment. CS+ is always followed by the UCS, but the co-occurrence of CS+ and CS− is never followed by the UCS. CS− becomes a conditioned inhibitor of the CR. (Reprinted from Barto & Sutton, 1982).

of a linear transformation, a process having many practical problem-solving applications. This algorithm has a long history in mathematics and appeared in the form of an adaptive element developed by Widrow and Hoff (1960), which they called an "adaline" (for *ada*ptive *line*ar). Closely related adaptive elements are those used in Rosenblatt's "perceptron" (1962) and Uttley's "informon" (1979). Consider a set $X = \{X^\alpha, 1 \leq \alpha \leq k\}$ of stimulus patterns $X^\alpha = (x_1^\alpha, \ldots x_n^\alpha)$ and an associated set of real numbers $Z = \{z^\alpha, 1 \leq \alpha \leq k\}$ where each z^α is the adaline response desired for stimulus pattern X^α. The weights of an adaline change as follows: for $1 \leq i \leq n$,

$$w_i(t+1) = w_i(t) + c[z(t) - \sum_{j=1}^{n} w_j(t)x_j(t)]x_i(t)$$

where $z(t) \in Z$ is the reference or "teacher" signal that provides the desired response to input pattern $X(t) = (x_1(t), \ldots, x_n(t)) \in X$, and c is a positive con-

stant. If the set X of input vectors is linearly independent and an adaline is trained by presenting the adaline with sufficient repetitions of the pairs (X^α, z^α), $1 \leq \alpha \leq k$, it will eventually respond with z^α when presented with X^α alone, $1 \leq \alpha \leq k$. In other words, it will form a weight vector $W^* = (w_1^*, \ldots, w_n^*)$ such that

$$[w_1^*, \ldots, w_n^*] \begin{bmatrix} x_1^\alpha \\ \vdots \\ x_n^\alpha \end{bmatrix} = z^\alpha$$

for $1 \leq \alpha \leq k$.

Widrow and Hoff (1960) proposed associative memory networks similar to those discussed by Anderson and Kohonen in this volume but consisting of adalines (although Widrow's work considerably predated this use of the term *associative memory*). Amari (1977a, 1977b) and Kohonen and Oja (1976) discuss similar networks. An associative memory network consisting of adalines does not require orthogonal input, or "key," vectors in order to obtain perfect recall peformance. Amari (1977a, 1977b) calls this "orthogonal learning" because nonorthogonal patterns are "orthogonalized" by the network. Moreover, if the set X is not even linearly independent, the system will form weights so as to minimize the mean square error. The process is, in fact, an algorithm for computing a linear regression or, more technically, for finding the Moore–Penrose pseudoinverse of a linear transformation. Duda and Hart (1973) provide a good overview of this general theory in the context of pattern classification.

Both the stimulus context effects of blocking and conditioned inhibition can be seen as instances of "orthogonalization." For a form of blocking, one has the stimulus vector $X^1 = (1, 0)$ representing the occurrence of CS_1 alone and the vector $X^2 = (1, 1)$ representing the co-occurrence of CS_1 and CS_2. These are clearly linearly independent but not orthogonal. The responses desired are $z^1 = z^2 = \lambda$ (as the UCS, and hence the UCR, occurs on both CS_1 alone and $CS_1 + CS_2$ trials). An adaline will form the weight vector $W^* = (\lambda, 0)$ giving

$$[\lambda, 0] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = [\lambda, 0] \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \lambda$$

Equivalently, the process solves the matrix equation

$$[w_1, w_2] \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = [\lambda, \lambda]$$

for w_1 and w_2 by effectively finding the inverse of the 2×2 matrix whose columns are the stimulus patterns X^1 and X^2. Blocking appears because w_2 turns out to be zero.

For conditioned inhibition, one has the vectors $X^1 = (1, 0)$ for CS+ occurrence and $X^2 = (1, 1)$ for the co-occurrence of CS+ and CS−. These are the same linearly independent but nonorthogonal vectors that represent the blocking experiment. The desired responses are $z^1 = \lambda$ and $z^2 = 0$ because the UCS is absent for CS−. An adaline will produce the weight vector $W^* = (\lambda, -\lambda)$, showing that CS+ eventually excites the element and CS− eventually inhibits it. Again, the process solves a matrix equation.

These stimulus context effects, and others that we have not discussed, provide evidence that animals "orthogonalize" their stimulus patterns during classical conditioning experiments. We think the independent discovery of this orthogonalization algorithm, in one case to describe animal behavior and in the other case to provide solutions to practical problems, is a remarkable instance of how purely theoretical problem-solving considerations can illuminate the adaptive significance of animal behavior. The adaptive element defined by Equations 1 and 2 orthogonalizes input patterns by virtue of its similarity to an adaline (and hence to the Rescorla–Wagner model) while also preserving some of the anticipatory aspects of classical conditioning. We have not yet thoroughly explored how these two aspects of our element's behavior interact, but an example of this interaction is provided by the results shown in Fig. 7.4, Trials 21–35. Here blocking is reversed because CS_2 begins earlier than a previously conditioned CS_1, suggesting that stimulus context effects occur insofar as they are consistent with the tendency to extract the earliest predictors of the UCS. We know of no attempts to perform this experiment on an animal preparation.

We have also not yet thoroughly explored the possibilities suggested by the use of our classical conditioning element in the associative memory paradigm discussed by Anderson and Kohonen in this volume. In one study, however, we used these elements to form a predictive associative memory that served as an internal model to evaluate proposed, but not overtly executed actions (Sutton & Barto, 1981a; see Fig. 7.7). We illustrated how this configuration was able to account for some of the difficult features of an experiment demonstrating "latent learning" in animals. We now focus on another type of adaptive element that was used in the "action selector" component shown in Fig. 7.7.

GOAL-SEEKING ADAPTIVE ELEMENTS

The adaptive element described in the preceding section operates in a completely open-loop mode: Its operation does not depend in any way on its being able to influence its input signals, as is appropriate because the classical conditioning paradigm was designed to prevent response contingencies (although in practice it may be impossible to remove all such contingencies). Instrumental (cued operant) conditioning, on the other hand, is learning that

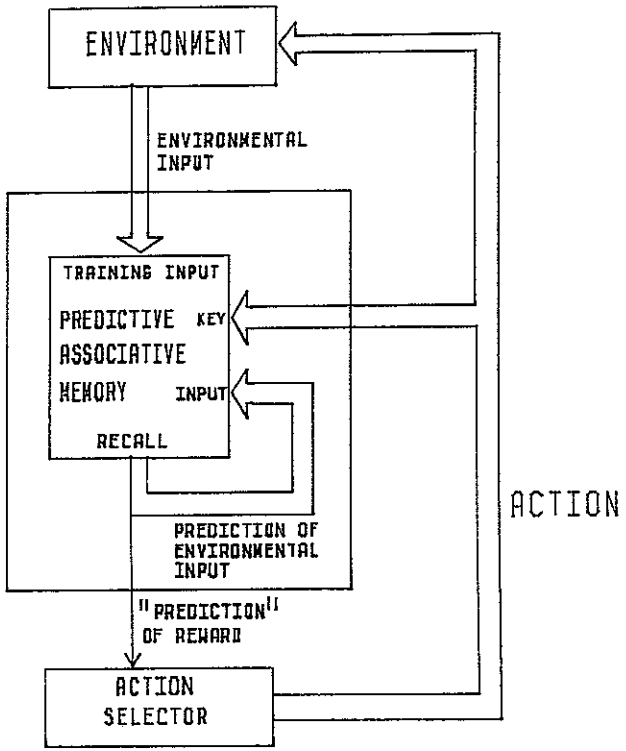

FIG. 7.7 The use of a predictive associative memory as an internal model enabling proposed actions to be evaluated before they are executed. (Reprinted from Sutton & Barto, 1981a).

occurs in experimental paradigms that do involve response contingencies. Reinforcement may be given or withheld depending on the animal's response. If a system can exert such control over its input, it is possible to speak of goal-seeking behavior in which, for example, the system acts so as to obtain appetitive stimuli and avoid aversive stimuli. Despite common belief to the contrary, nontrivial forms of response-contingent learning have received very little attention from adaptive network theorists[2]. Recognizing this, Klopf (1972, 1982) proposed that neurons may operate as analogs of instrumental

[2]This may seem a surprising comment, and an adequate defense of it is beyond the scope of the present chapter. Although the "error-correction" methods employed by the adaline or perception, for example, are often considered to be analogous to "trial-and-error" learning, they are not. These methods search in the space of weight vectors but not in the space of possible actions. In Barto and Sutton (1981b) we discuss this in more detail.

conditioning rather than classical conditioning and suggested how this may be accomplished. What follows is a discussion of some of our studies of networks of such goal-seeking components.

The psychological literature on instrumental conditioning and on the relationship between instrumental and classical conditioning is extremely complex. Rather than attempting to carefully integrate our studies of closed-loop learning rules with this literature, as we attempted to do for the open-loop case of classical conditioning, we have instead concentrated on the problem-solving potential of such rules. Here we describe some simulation experiments intended to illustrate these capabilities in a vivid and intuitively satisfying manner. This adaptive element was presented by Barto, Sutton, and Brouwer (1981) and the experiments described here were presented by Barto and Sutton (1981b).

This adaptive element has n input pathways x_i, $1 \leq i \leq n$, a specialized "payoff" pathway z, and an output pathway y. We let $x_i(t)$, $1 \leq i \leq n$, $z(t)$, and $y(t)$ respectively denote the activity on these pathways at time t. As usual, a variable weight with value $w_i(t)$ at time t is associated with each pathway x_i, $1 \leq i \leq n$. Let

$$s(t) = \sum_{i=1}^{n} w_i(t)x_i(t).$$

The output of the element at time t is

$$y(t) = \begin{cases} 1 \text{ if } s(t) + \text{NOISE}(t) > 0 \\ 0 \text{ else} \end{cases} \tag{3}$$

where NOISE(t) is a normally distributed random variable with mean zero. The weights change according to the following equation:

$$w_i(t) = w_i(t-1) + c[z(t) - z(t-1)]y(t-1)x_i(t-1) \tag{4}$$

for $1 \leq i \leq n$, where c is a positive constant determining the rate of learning.

This adaptive element *searches* for the action that will lead to the largest payoff obtainable in the situations signaled by its stimulus patterns. Suppose the payoff provided to the element at time t is a function of the element's action at time $t-1$ and the stimulus pattern $X(t-1) = (x_1(t-1), \ldots, x_n(t-1))$ present at time $t-1$; that is $z(t) = f[y(t-1), X(t-1)]$. The element is to learn to perform the action $y(t-1)$ in response to the pattern $X(t-1)$ that maximizes $z(t)$. The element searches for this action by trying its various responses to each pattern and settling on the one that turns out to be best. The element need never be directly instructed as to which response is best for each pattern. If the consequences of an action are not returned to the element in one time step as we have assumed here, it is appropriate to replace the terms $z(t-1)$, $y(t-1)$, and $x_i(t-1)$ in Equation 4 with prolonged

traces of these signals such as those used in the classical conditioning element described in the previous section.

The random component in the element's response (Equation 3) is essential to this process. Responses are made randomly but are biased in one direction or the other by the sum s. Because s depends on the input patterns through the weights, the weights determine how this probabilistic bias conditionally depends on each input pattern. According to Equation 4, if the element "fired" at $t - 1$ (i.e., $y(t - 1) = 1$) in the presence of nonzero input activity on pathway i (i.e., $x_i(t - 1) > 0$), perhaps due to an excitatory effect of signal x_i or perhaps by chance, *and* this was followed by an increase in payoff (i.e., $z(t) - z(t - 1) > 0$), then firing in the presence of signal x_i is made more likely by incrementing weight w_i. Similarly, the firing probability is decreased if the payoff decreases. The noise in the response, then, is essential to the learning process because it generates trials in the absence of any preestablished influence from sensory input and continues to generate trials as this influence is established. Conducting a search in this probabilistic manner also permits the element to improve its performance (in terms of the amount of payoff received) even if the environment provides payoff in a nondeterministic manner, a property whose importance will become more clear when we consider a network of these elements.

Hill Climbing and Chemotaxis

The adaptive element just described implements an elaboration of a goal-seeking strategy that occurs in certain simple organisms. Fraenkel and Gunn (1961) discuss a number of methods used by animals for finding and remaining near light or dark areas, warm or cool areas, or, in general, for approaching attractants and avoiding repellents. One of the most primitive mechanisms is a strategy that they called klinokinesis, the most intensely studied example of which occurs in the behavior of various types of bacteria such as *Escherichia coli, Salmonella typhimurium,* or *Bacillus subtilis.* This manifestation of klinokinesis, known as bacterial chemotaxis, was discovered in the 1880s and was recently reviewed by Koshland (1979). These bacteria propel themselves along relatively straight paths by rotating (!) flagella. With what at first appears to be a random frequency, they reverse flagellar rotation, which causes a momentary disorganization of the flagellar filaments. This causes the bacterium to stop and tumble in place. As the disorganized flagellum continues to rotate in the new direction, its filaments twist together again, causing the bacterium to move off in some random new direction. If the attractant is getting stronger, the probability of reversing flagellar rotation decreases, thereby increasing the probability that the bacterium will continue to move in the same direction; whereas if the attractant level drops, the probability increases that the bacterium's flagellum will reverse and cause

the bacterium to swim off in a randomly chosen new direction. Runs in directions leading up the attractant gradient therefore tend to be longer than runs in directions leading down the gradient. As a result of this strategy, bacteria are able to find and remain in the vicinity of the peaks of attractant distributions. Selfridge (1978) points out the general utility of this basic mechanism, which he calls "run-and-twiddle"—if things are getting better, keep doing whatever you are doing; if things are getting worse, do something (anything!) else. It is a very effective strategy, particularly when gradient information is very noisy.

To see how the adaptive element defined by Equations 3 and 4 implements an analog of this procedure, consider an element that receives only a single input signal, say x_0, in addition to the payoff, and assume that x_0 has a constant value of 1, that is, $x_0(t) = 1$ for all t. The weight w_0 associated with this signal changes according to Equation 4, where the term $x_i(t - 1) = 1$ for all t. The payoff level $z(t)$ represents the level of attractant sensed by the element at time t. Thus, if "firing" is followed by an increase in attractant level, then firing is made more likely. Note that we can consider the single constant input and its weight as a convenient means of specifying a variable threshold (so that the constant input need not really be supplied from the element's environment). If upper and lower bounds were imposed on the value of w_0 and if the learning constant c were large enough, then a single "move" up or down the attractant gradient would respectively cause the element to "continue doing what it was doing" or to "do something else" (with a high probability).

Rather than directly simulating a spatial version of "running" and "twiddling" using a single element, we simulated an "organism" whose locomotion is controlled by four adaptive elements, each controlling movement in one of the four cardinal directions; it moves north if Element 1 fires, south if Element 2 fires, etc. In case two elements fire simultaneously, then an appropriate compound move is made, for example, northwest. A sort of "reciprocal inhibition" is used to reduce the probability that the north and south or the east and west elements fire together. We assume that each move is a fixed distance and is completed in a single time step. Clearly, we were not attempting to model in any detailed manner the motor control system of an actual organism, and we have not optimized this hill-climbing strategy. Figure 7.8 shows the simulated organism's trail in an environment containing a "tree" as the center of an attractant distribution that decreases linearly with distance from the tree. The organism (shown as an asterisk) approaches the tree and remains in its vicinity.

Associative Search

The simulated organism climbing the attractant distribution in Fig. 7.8 is not forming long-term memory traces. If we were to move it back to its starting

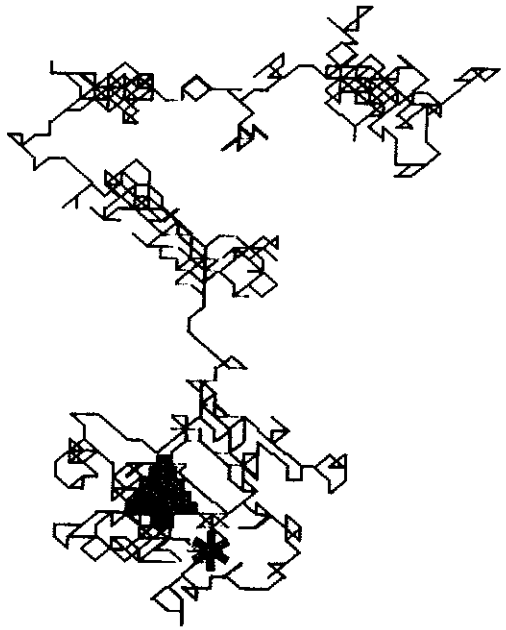

FIG. 7.8 Chemotacticlike behavior of a network of goal-seeking adaptive elements. The "organism," shown as an asterisk, started in the upper right and generated the trail shown as it climbed an attractant distribution whose peak is marked by the location of the "tree." (Reprinted from Barto & Sutton, 1981b).

position, it would take just as long (on the average) to move toward the tree; nothing was learned during the first excursion. This suggests that the other input pathways to the elements controlling locomotion might provide information that can be used to guide the hill-climbing procedure and that their weights might provide useful long-term memory traces. The following simulation experiments were designed to explore the coupling of associative learning capabilities with chemotacticlike behavior. To the spatial environment already described, we added four "landmarks," each of which emits a distinctive "odor" that decays with distance from the landmark (Fig. 7.9a). These odors are neutral in the sense that they are not attractants or repellents but can serve as cues as to location in space.

Figure 7.9b shows the network of four adaptive elements that controls movement in the manner described. These input pathways are labeled vertically on the left according to the landmarks to which they respond. The location of the organism, then, determines the input pattern it receives. The shaded input pathway N in Fig. 7.9b indicates that the organism is near the north neutral landmark. Given the presence of these other signals, there is no

longer a need for the constant input x_0 (although the system still works if it is present). The arrangement of input and output pathways used in Fig. 7.9b permits us to show the connection weights as circles centered on the intersections of input pathways and the vertical output element "dendrites." We show positive weights as hollow circles and negative weights as solid circles. The sizes of the circles indicate the relative magnitudes of the corresponding weights. The uppermost "tree" input is the payoff pathway z, which has no associated weights. This network is an example of what we have called an "associative search network" (Barto et al., 1981). The matrix of weights forms an associative memory, but unlike those discussed by others, it need not be directly told what associations to store. Instead, it stores the successful results of the chemotacticlike search. With sufficient experience, the system can learn to respond to the configuration of signals at each place with the action that is optimal for that place.

Figure 7.10 illustrates the performance of this system. In this case, noise has been added to the attractant level in order to make the hill-climbing task more difficult. Figure 7.10a shows the trail of an inexperienced organism that starts near the nothern neutral landmark. It eventually remains in the vicinity of the tree. Figure 7.10b shows the trail produced by replacing the organism at its original starting point after it has undergone the experience shown in Fig. 7.10a. It now proceeds directly to the tree, clearly benefiting from its earlier experiences. Figure 7.11a shows the network after learning. Nonzero weights have appeared so that, for example, proximity to the northern landmark causes a high probability of movement south because the "odor" of the northern landmark excites the element that causes movement south and inhibits the one that causes movement north. Figure 7.11b shows the results of learning as a vector field in which each vector shows the average direction that the organism will take on its *first* step from any place. The vector field is the organism's map of its environment (it is never literally present in the environment). Moreover, it should be clear that the organism would follow this map even if the tree and its attractant distribution were to be removed (so long as the neutral landmarks remained). Although the problem is simple enough for this network to solve by forming a linear associative mapping, it illustrates how adaptively significant behavior can be achieved naturally by combining associative learning with chemotacticlike strategies. Further discussion of this example is provided in Barto and Sutton (1981b).

Although some accounts of learning in the cybernetic literature essentially equate learning and hill climbing, here we see an example of a hill climber that learns. This is very important from a problem-solving perspective. Search is an essential element of almost any problem-solving task (see, for example, Minsky, 1963), but it is often essential to minimize explicit search in order to gain efficiency. The landmark-guided hill-climbing example illustrates how the results of explicit searches can be transferred to an associative

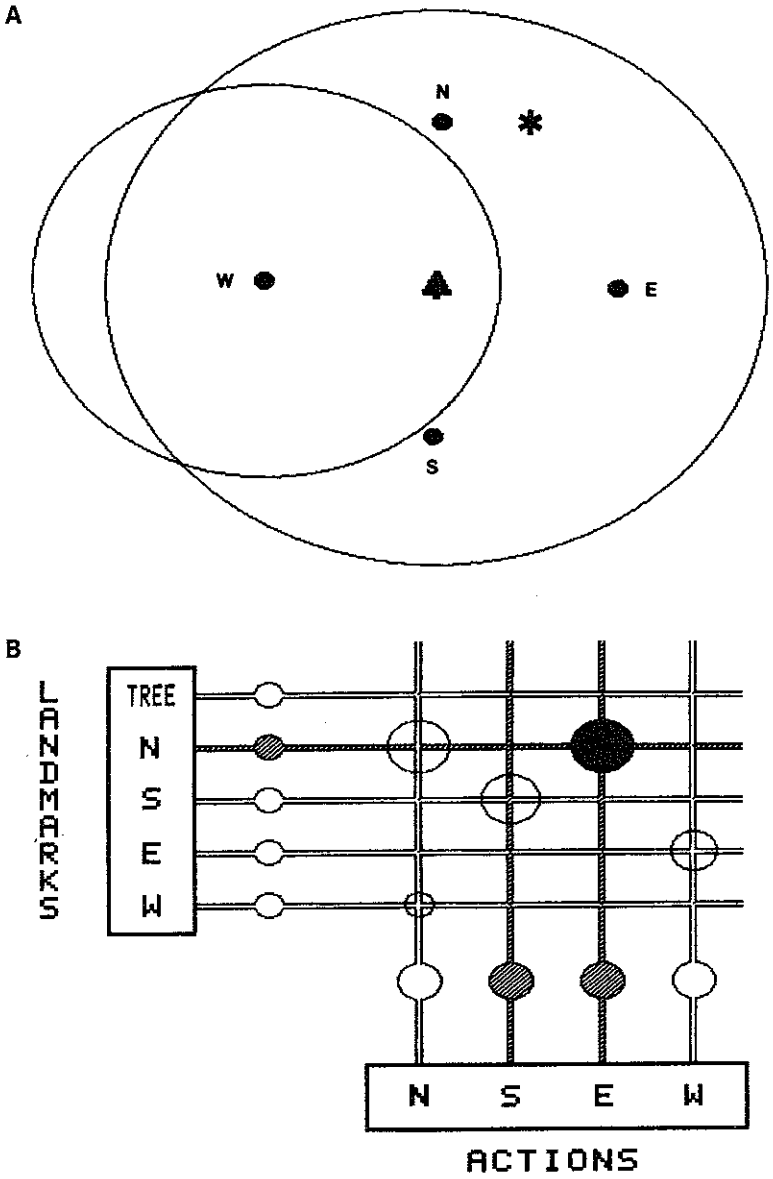

FIG. 7.9

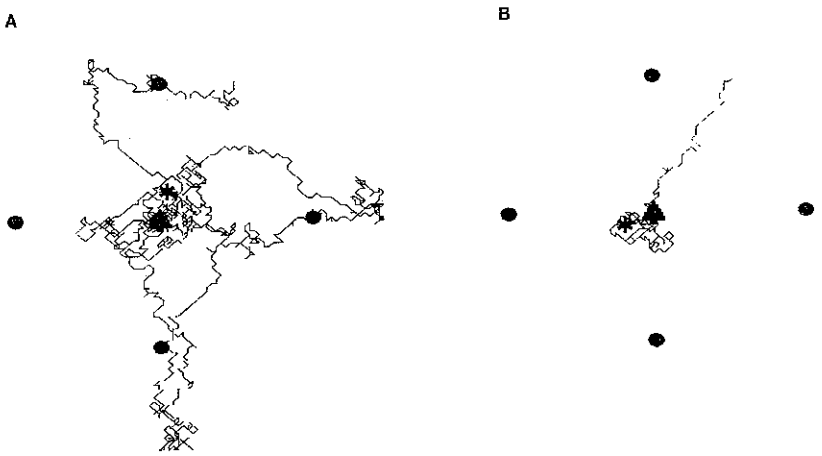

FIG. 7.10 Chemotacticlike behavior combined with associative learning. (a) The trail of an inexperienced organism that starts near the northern neutral landmark. Hill climbing is difficult because noise has been added to the attractant level, but the organism eventually remains in the vicinity of the attractant peak. (b) The trail produced by an experienced organism. After the experience shown in (a), the organism is placed in its original starting position. It now proceeds directly to the tree, clearly benefiting from its previous experience.

long-term store so that in future encounters with similar (but not necessarily identical) situations the system need only access the store in order to find out what to do. The associative search network shows how all of this can be accomplished without centralized control. It is thus an improvement over a nonlearning search method while also offering an improvement over the usual storage methods for associative memories because the optimal responses need not be known a priori by the environment, the system, or the

FIG. 7.9 *(Opposite page)* (a) A spatial environment in which the attracting "tree" is surrounded by four other landmarks. The landmarks each possess a distinctive "odor" that can be sensed at a distance but that is not an attractant. Odor distributions decrease linearly from their associated landmarks and become undetectable at a certain distance (indicated for landmark '*W*' by the surrounding circle). (b) A network of goal-seeking adaptive elements. The five input pathways are labeled vertically on the left according to the landmarks to which they respond. The shaded input pathway *N* indicates that the organism is near the north neutral landmark. The four output pathways controlling actions are labeled horizontally at the bottom according to the direction of movement they cause. The shaded output elements indicate that a southeast movement is being made. The associative matrix weights are displayed as circles centered on the intersections of the horizontal input pathways and vertical output pathways. Positive weights are shown as hollow circles, and negative weights are shown as solid circles. (Reprinted from Barto & Sutton, 1981b).

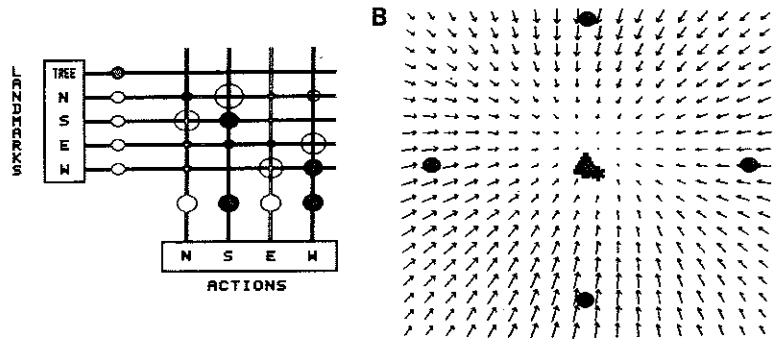

FIG. 7.11 Associative memory contents after learning. (a) The network showing the weights. Nonzero weights have appeared so that, for example, proximity to the northern landmark causes a high probability of moving south because the "odor" of the northern landmark excites action *S* and inhibits action *N*. (b) A vector field representation of the associative memory's contents. Each vector shows the most likely direction that the organism will move on its *first* step from any place. Note the generalization to places it has never visited. (Reprinted from Barto & Sutton, 1981b).

system's designer. Future research will focus on networks that combine associative learning with search strategies that are more sophisticated than simple *hill climbing*.

Neural Signaling and Bacterial Chemotaxis

Koshland (1979) suggests that study of the numerous commonalities between bacterial chemotaxis and other forms of adaptive behavior in single-celled organisms, and the signaling systems of neurons may provide insight into neural mechanisms. Like bacteria, neurons possess receptors that detect chemical signals from their environments. A bacterium's sensory-processing system produces signals that control its motor response by altering the probability of flagellar reversal. Neurons similarly respond to chemically mediated afferent signals and produce action potentials as "motor" responses. Koshland (1979) hypothesizes that many features of bacterial chemotaxis can be accounted for by a model in which random variations in the concentration of a hypothetical tumble regulator substance X are modulated by changes in attractant concentrations. Flagellar reversal occurs whenever the concentration of X exceeds a threshold. Suppose X is formed at rate V_f and decomposed at rate V_d. If an increase in the level of attractant sensed causes a fast increase in V_f and a slower increase in V_d, then the intracellular concentration of X will show a transient increase to any sustained increase in attractant level, and a transient decrease to any sustained decrease (Fig. 7.12), thus

causing the appropriate hill-climbing behavior. This is the same sort of "differentiation" accomplished by the term $s(t) - \bar{s}(t)$ of the classical conditioning element (Equation 2) and the term $z(t) - z(t - 1)$ of the goal-seeking element (Equation 4). More specifically, the value of the term $s(t)$ + NOISE(t) in Equation 3 functionally corresponds to the concentration of the hypothetical substance X in Koshland's model. Mechanisms similar to those suggested by Koshland for bacterial chemotaxis could provide a basis for neurons to exhibit related behavior.

It is an intriguing hypothesis that neurons implement goal-seeking strategies related to those of single-celled organisms. Perhaps it will prove useful to view neurons as swimming (in a metaphorical sense, of course) in an environment of contingencies determined by the nervous system of which they are a part and the organism and its environment to whose survival they contribute. Important aspects of a neuron's behavior may involve its ability to influence its own input when operating in its usual environment. This influence may extend through the environment external to the entire organism, as well as through local internal feedback loops. In order to experimentally investigate this hypothesis, single neurons would need to be studied in closed-loop control situations in which their efferent activity could influence, perhaps after considerable delay, their afferent activity according to experimentally known and controllable transformations.

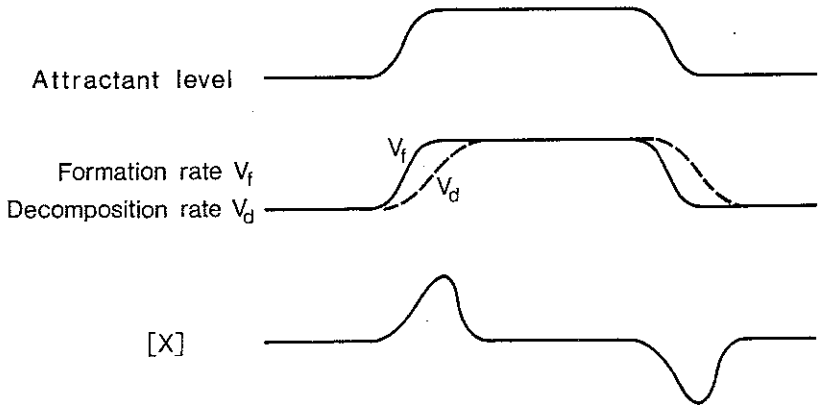

FIG. 7.12 Hypothetical mechanism for detecting attractant gradient in bacterial chemotaxis (from Koshland, 1979). The formation and decomposition rates V_f and V_d of a hypothetical substance X are influenced by the attractant level. Both V_f and V_d follow the attractant level sensed, but V_d changes more slowly than V_f. This results in the concentration of X responding to changes in attractant levels. (After Barto & Sutton, 1981c).

ASSIGNMENT OF CREDIT

We have described two types of adaptive elements that share many basic features but whose behaviors have a different character, one closely related to classical conditioning, and the other related to instrumental conditioning and bacterial chemotaxis. We have argued that both types of behavior would confer adaptive advantages to any organism possessing them, but we have not suggested how these forms of behavior might be related. Here, we propose that this relationship can be understood in terms of what has been called the "assignment of credit problem." Suppose success is achieved by a complex mechanism after operating over a considerable period of time (for example, a chess-playing program wins a game). To what particular decisions made by what particular components should the successs be attributed? And, if failure results, what decisions deserve blame? The magnitude of this problem is most forcefully appreciated by those actually attempting to construct systems capable of learning to improve performance in complex tasks. This is closely related to the problem known as the "mesa" or "plateau" problem (Minsky, 1963; Minsky & Selfridge, 1960). The performance evaluation function available to a learning system may consist of large level regions in which hill-climbing degenerates to exhaustive search. Only a few of the situations obtainable by the learning system and its environment are known to be desirable, and these situations may occur rarely.

An approach to one aspect of this problem is illustrated by the network of goal-seeking components described previously. At each time step, each element produces a component of a total output pattern. If a pattern produces an increase in the performance evaluation (i.e., if the organism moves up the attractant gradient), to what element or elements should success be attributed? The network solves this problem by assigning credit to *any* element that happened to fire, whether or not its firing was actually causal in producing success. The probabilistic nature of the search procedure, however, allows any misleading consequences of this strategy to be averaged out with repeated trials (and we are reminded of the philosophical problem of truly distinguishing causality from correlation). More technically, part of each element's operation implements what is known as a stochastic learning automaton optimization method (see, for example, Narendra & Thathachar, 1974) and is capable of improving its performance under the uncertainty produced by the unknown and random influences of the other elements on its own payoff. Of course, the larger the network, the more trials will be required in general for credit to be apportioned correctly. Thus, this method alone will not suffice for large networks. Another part of the solution may be to permit interconnections to form between elements and to effectively assign credit to linked assemblies of elements rather than to individual elements.

Our experiments with layered networks of goal-seeking elements suggest that this approach indeed works, but a complete discussion is beyond the scope of the present chapter.

Another aspect of the assignment of credit problem concerns temporal factors. The utility of making a certain action may depend on the sequence of actions of which it is a part, and an indication of improved performance may not occur until the entire sequence has been completed. The landmark learning task presented here does not illustrate this problem because we assumed that an action was always evaluated in a single time step. An approach to this problem has been discussed by Minsky (1963) and has been used successfully in Samuel's (1959) famous learning checkers-playing program. The idea is to interpret predictions of future reward as rewarding events themselves. In other words, neutral stimulus events can themselves become reinforcing if they regularly occur before events that are intrinsically reinforcing. This phenomenon is observed in animal learning experiments in which neutral stimuli can become "secondary reinforcers" if they predict "primary reinforcement." This has two consequences. First, a prediction of eventual reward can reinforce the actions that precede that prediction, thereby eliminating the delay in obtaining useful evaluative feedback. Second, a prediction of reward can provide reinforcement to the learning process by which the predictions themselves are formed, permitting the formation, via associative transfer, of predictions of predictions, etc. This is, in fact, the mechanism employed by the classical conditioning element described in the second section of this chapter. Its anticipatory behavior, coupled with its ability to produce higher-order conditioning, is ideally suited for providing evaluative information to a goal-seeking system that is more useful than information directly available from its environment. This view parallels the CR-mediational theories of instrumental conditioning proposed by animal learning theorists (Gormezano & Kehoe, in press). Moreover, the classical conditioning element turns out to implement an algorithm remarkably similar to a part of the actual algorithm used by Samuel in his checkers-playing program. We are currently investigating systems that combine both types of adaptive elements and that face control tasks characterized by variably delayed reinforcement, and it may be possible to devise a single relatively simple element that combines both types of behavior.

CONCLUSION

In this article, we have described some of the results of a research program intended to reexamine the potential for networks of neuronlike adaptive elements to provide a computational substrate for solving nontrivial problems.

We have highlighted examples of how adaptively significant features of animal behavior and pure problem-solving considerations converge: the anticipatory nature of classical conditioning and the necessity to construct internal evaluation criteria to solve problems involving variably delayed reinforcement; stimulus context effects of classical conditioning and the utility of orthogonalizing stimulus patterns for associative storage; bacterial chemotaxis and the necessity of search in problem solving. We have described an adaptive element that preserves some of these features of classical conditioning and an element that combines the goal-seeking nature of chemotaxis with associative learning. Networks of the latter type of element conduct searches, store the results of these searches, and access these results to aid future searches. They also eliminate the necessity for the learning system's environment to know the optimal associations. Further, this is accomplished without centralized control. Our present research is directed toward extending these capabilities in order to produce networks that are able to solve problems that have proved resistant to standard problem-solving methods.

ACKNOWLEDGMENTS

This research was supported by the Air Force Office of Scientific Research and the Avionics Laboratory (Air Force Wright Aeronautical Laboratories) through contracts F33615-77-C-1191 and F33615-80-C-1088. The authors wish to thank M. A. Arbib, W. L. Kilmer, D. N. Spinelli, A. H. Klopf, J. W. Moore, and O. G. Selfridge for their many valuable criticisms and contributions.

REFERENCES

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

Amari, S. Neural theory of association and concept-formation. *Biol. Cybernetics,* 1977, *26,* 175–185. (b)

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

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

Barto, A. G., & Sutton, R. S. Landmark learning: An illustration of associative search. *Biol. Cybernetics,* 1981, *42,* 1–8. (b)

Barto, A. G., & Sutton, R. S. Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research,* 1982, *4,* 221–235. (c)

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

Box, G., & Jenkins, G. *Time series analysis: Forecasting and control.* San Francisco: Holden Day, 1976.

Burke, W. Neuronal models for conditioned reflexes. *Nature,* 1966, *210,* 269–271.

Duda, R. O., & Hart, P. E. *Pattern classification and scene analysis.* New York: Wiley, 1973.

Fraenkel, G. S., & Gunn, D. L. *The orientation of animals: Kineses, taxes and compass reactions.* New York: Dover, 1961.

Frey, P. W., & Sears, R. J. Model of conditioning incorporating the Rescorla–Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychol. Rev.,* 1978, *85,* 321–340.

Gormezano, I. Investigations of defense and reward conditioning in the rabbit. In A. H. Black & W. F. Prokasy (Eds.), *Current research and theory.* New York: Appleton–Century–Crofts, 1972.

Gormezano, I., & Kehoe, E. J. Associative transfer in classical conditioning to serial compounds. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analysis of behavior. Vol. 3: Acquisition.* Cambridge: Ballinger, in press.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Hilgard, E. R., & Bower, G. H. *Theories of learning* (4th ed.). Englewood Cliffs, N.J.: Prentice–Hall, 1975.

Kehoe, J. E., Gibbs, C. M., Garcia, E., & Gormezano, I. Associative transfer and stimulus selection in classical conditioning of the rabbit's nictitating membrane response to serial compound CSs. *Journal of Experimental Psychology: Animal Behavior Processes,* 1979, *5,* 1–18.

Kimmell, H. D. Instrumental inhibitory factors in classical conditioning: In W. F. Prokasy (Ed.), *Classical conditioning.* New York: Appleton–Century–Crofts, 1965.

Klopf, A. H. *Brain function and adaptive systems—A heterostatic theory.* Air Force Cambridge Research Laboratories Research Report AFCRL-72-0164, Bedford, Mass., 1972. (A summary appears in Proc. Int. Conf. Syst., Man., Cybern., IEEE Syst., Man, Cybern., Soc., Dallas, Texas, 1974)

Klopf, A. H. Goal-seeking systems from goal-seeking components: Implications for AI. The *Cognition & Brain Theory Newsletter,* 1979, *3,* 2.

Klopf, A. H. *The hedonistic neuron: A theory of memory, learning and intelligence.* Washington, D.C.: Hemisphere Publishing Corp., 1982.

Kohonen, T., & Oja, E. Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biol. Cybernetics,* 1979, *21,* 85–95.

Koshland, D. E. Jr. A model regulatory system: Bacterial chemotaxis. *Physiol. Rev.,* 1979, *59,* 811–862.

Mackintosh, N. J. *The psychology of animal learning.* New York: Academic Press, 1974.

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

Minsky, M. L. Steps toward artificial intelligence. In E. A. Feigenbaum, & J. Feldman (Eds.), *Computers and thought.* New York: McGraw–Hill, 1963.

Minsky, M. L., & Selfridge, O. G. Learning in random nets. In C. Cherry (Ed.), *Information theory: Fourth London symposium.* London: Butterworths, 1960.

Narendra, K. S., & Thathachar, M. A. L. Learning automata—a survey. *IEEE Trans. Syst., Man, Cybern. SMC-4,* 1974, *4,* 323–334.

Rescorla, R. A., & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory.* New York: Appleton–Century–Crofts, 1972.

Rosenblatt, F. *Principles of neurodynamics.* New York: Spartan Books, 1962.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM J. Res. and Dev.,* 1959, *3,* 210–229.

Selfridge, O. G. *Tracking and trailing: Adaptation in movement strategies.* Unpublished draft, August 1, 1978.

Sutton, R. S., & Barto, A. G. An adaptive network that constructs and uses an internal model of its world. *Cognition & Brain Theory,* 1981, *4,* 217–246. (a)

Sutton, R. S., & Barto, A. G. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol. Rev.,* 1981, *88,* 135–170. (b)

Uttley, A. M. *Information transmission in the nervous system.* London: Academic Press, 1979.

Widrow, G., & Hoff, M. E. Adaptive switching circuits. In *1960 IRE WESCON Convention Record,* 1960, Part 4, 96–104.