A Temporal-Difference Model of Classical Conditioning

Richard S. Sutton
GTE Laboratories Incorporated

Andrew G. Barto
University of Massachusetts

Abstract—Rescorla and Wagner's model of classical conditioning has been one of the most influential and successful theories of this fundamental learning process. The learning rule of their theory was first described as a learning procedure for connectionist networks by Widrow and Hoff. In this paper we propose a similar confluence of psychological and engineering constraints. Sutton has recently argued that adaptive prediction methods called *temporal-difference methods* have advantages over other prediction methods for certain types of problems. Here we argue that temporal-difference methods can provide detailed accounts of aspects of classical conditioning behavior. We present a model of classical conditioning behavior that takes the form of a temporal-difference prediction method. We argue that it is an improvement over the Rescorla-Wagner model in its handling of within-trial temporal effects such as the ISI dependency, primacy effects, and the facilitation of remote associations in serial-compound conditioning. The new model is closely related to the model of classical conditioning that we proposed in 1981, but avoids some of the problems with that model recently identified by Moore et al. We suggest that the theory of adaptive prediction on which our model is based provides insight into the functionality of classical conditioning behavior.

Introduction

The increasing interest in connectionist or parallel distributed processing models of cognitive behavior provides a new rationale for examining animal conditioning behavior. Many of the rules used for adjusting connection weights in connectionist models are the result of postulating that single neuron-like units exhibit simplified analogs of animal behavior in conditioning experiments. Connectionist theories of higher functions therefore provide vehicles for integrating insights from animal learning research into more comprehensive theories of behavior. At the same time, the mathematical theories associated with connectionist learning provide new theoretical perspectives on conditioning behavior.

This research was supported in part by the Air Force Office of Scientific Research through grant AFOSR-87-0030. The authors wish to thank Harry Klopf, Jim Morgan, Jim Kehoe, John Moore, John Desmond, Diane Blazis, and Neil Berthier for sharing their ideas and simulation results with us. We also particularly thank John Moore for reading and providing valuable comments on an earlier draft of this paper.

Viewed at the trial level, classical, or Pavlovian, conditioning is related to supervised associative learning as studied by engineers and computer scientists and embodied in many connectionist learning systems. The system is repeatedly presented with an input pattern, corresponding to a conditioned stimulus (CS), together with a specification of a desired response, which corresponds to the presentation of an unconditioned stimulus (US) and the unconditioned response (UR) that it reflexively elicits. After a number of such CS–US pairings, the CS comes to elicit a conditioned response (CR) that closely resembles the UR or some part of it.* When details occurring within trials are considered, classical conditioning is seen to involve the extraction of predictive relationships among stimuli as if causal rules are being learned.

In a previous paper (Sutton and Barto, 1981), we pointed out that the Rescorla-Wagner model of classical conditioning (Rescorla and Wagner, 1972) is nearly identical to the learning algorithm introduced earlier by engineers Widrow and Hoff (1960), which is used in practical engineering applications (Duda and Hart, 1973; Widrow and Stearns, 1985) as well as in recent connectionist models (e.g., see Rumelhart and McClelland, 1986). That there is this degree of correspondence between psychological models and engineering methods should not be surprising given the similarity of the functional demands made in each case. In this paper, we propose a refinement of this correspondence. We propose a new model of classical conditioning based on a new theory of engineering methods called *temporal-difference methods* (Sutton, 1987). Temporal-difference methods have been shown to be superior in certain respects to the Widrow-Hoff algorithm and to other engineering algorithms for adaptive prediction. Here, we argue that the new model of classical conditioning, which we call the Temporal-Difference, or TD, model, also provides a better account of animal learning data than the Rescorla-Wagner model. In addition, the TD model and the theory of temporal-difference methods provides specific new suggestions about the functional nature of classical conditioning.

The TD model is a minor variant of the Adaptive Heuristic Critic (AHC) algorithm developed by Sutton for temporal credit assignment (Sutton, 1984; Barto, Sutton, and Anderson, 1983) and combined with the error back-propagation method of Rumelhart, Hinton, and Williams (1985) by Anderson (1986). The AHC algorithm itself is closely

---

* For example, a human subject is repeatedly presented with the sound of a bell (CS) followed by a puff of air to his eye (US), which causes him to blink (UR). After several such pairings, the subject blinks immediately (CR) in response to the bell alone.

related to the model of classical conditioning that we proposed in 1981 (Sutton and Barto, 1981; Barto and Sutton, 1982), which we here call the Sutton-Barto, or SB, model, and which was strongly influenced by the work of Klopf (1972, 1982). In this paper, we present the TD model as a substantially modified version of the SB model that solves some of the problems with that model identified by Moore et al. (1986). We show how the TD model performs in simulations of single-CS acquisition and extinction, trace and delay conditioning, blocking, conditioned inhibition, second-order conditioning, and several serial-compound conditioning paradigms. We also discuss what the theoretical basis of the TD model suggests about what animals are doing in classical conditioning. Finally, we briefly mention some of the limitations of the TD model.

Real-Time Models of Classical Conditioning

Whereas many models of classical conditioning (e.g., Rescorla and Wagner, 1972; Mackintosh, 1975; Pearce and Hall, 1980) specify changes in associative strength only as the result of a trial as a whole, the TD and SB models specify changes in associative strengths from moment to moment within trials. We will call models with this property *real-time* models (after Moore and Stickney, 1980; Blazis et al., 1986). Real-time models have also been proposed by, e.g., Gelperin, Hopfield, and Tank (1985), Gluck and Thompson (in press), Hawkins and Kandel (1984), Klopf (1986), Moore et al. (1986) Tesauro (1986), and Wagner (1981).

Real-time models have several kinds of advantages over trial-level models. First, since real-time models distinguish between times within a trial, they can make predictions about the effects of varying the temporal relationships among stimuli within a trial, whereas trial-level models can't. The trial-level Rescorla-Wagner model, for example, does not make predictions about the effect of the inter-stimulus interval between CS and US, even though this is well-known to have a strong effect on conditioning. A second advantage of real-time models is that they are more mechanistic and thus it is easier to see how they might be implemented by physical mechanisms. In particular, they are a step closer to neural models since their behavior can be compared more directly with electrophysiological correlates of learning.

Some real-time models, including the SB model, have been presented in the form of rules for altering the connection weights of a neuron-like adaptive element, and we follow this tradition with our description of the TD model. Although this form of presentation suggests possible relationships to the cellular basis of learning and makes it clear how the model can be used as a learning rule for connectionist networks, it is not essential to the TD model as a model of conditioning behavior. Nor is the realization of the model suggested by this adaptive-element the only way the model could be implemented in a nervous system.

The SB Model

We first describe the SB model and then discuss several of its shortcomings. Following our 1981 paper (Sutton and Barto, 1981) we present it as a set of rules for adjusting the connection weights of a neuron-like element, but we use a slightly different notation. Figure 1 shows a neuron-like adaptive element with $n+1$ input pathways, labeled $x(0),\ldots,x(n)$, and a single output pathway labeled y. For each i, $i = 0,\ldots,n$, $x_t(i)$ denotes the strength of the signal on pathway i at time t; y_t denotes the strength of the output signal at step t. Associated with each input pathway $x(i)$ is a weight $w(i)$ that specifies the efficacy of that pathway; $w_t(i)$ denotes the weight's value at time t. Pathway $x(0)$ is the US pathway and its weight $w(0)$ is positive and constant over time. Patterns of activity over the remaining input pathways represent stimuli that can be associated with the US—the CSs.* Changes in the weights of the CS pathways over time represent changes in the the associative strengths of the CSs with respect to the US. We denote by x_t the input vector at time t consisting of the n components of the CS vector, i.e., $x_t = (x_t(1),\ldots,x_t(n))$. Similarly, w_t denotes the n-component vector of weights of the CS pathways at time t. The element output, y, is assumed to contribute to both the UR and the CR.

---

* Tesauro (1986) correctly points out that the original description of the SB model suggests that the model is applicable only when a CS is represented locally by activity on a single input pathway. However, the model obviously also applies to the case of distributed CSs, and we wish to allow that possiblity here. This is also true of the TD model, but in the simulations presented here, locally represented CSs are used for simplicity.

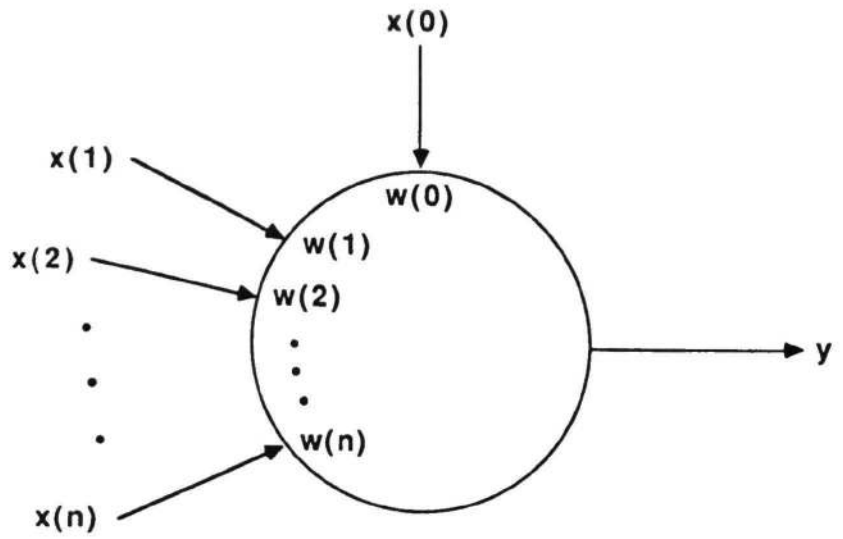

Figure 1. A neuron-like adaptive element used in the SB model. There are n modifiable CS input pathways, $x(1),\ldots,x(n)$, and a pathway $x(0)$ with fixed weight $w(0)$ that corresponds to the US. The element output y corresponds to both the UR and the CR.

The element output at time t is a function of the weighted sum of the inputs at time t:

$$y_t = f\left\{\sum_{i=0}^{n} w_t(i)x_t(i)\right\}, \tag{1}$$

where $f\{\cdot\}$ is some S-shaped function; in our earlier simulations we assumed it was the identity function. We assume that this input/output relationship is instantaneous because the model does not address intrinsic response latencies, which vary across response systems.

The connection weights of the CS pathways are updated at each time step as follows:

$$w_{t+1} = w_t + c(y_t - y_{t-1})\bar{x}_t, \tag{2}$$

where $c > 0$ and $\bar{x}_t$ is the vector of *eligibility traces*, each component of which is a weighted sum of past values of the corresponding input signal.* We compute these traces using the following recursion:

$$\bar{x}_t = \beta\bar{x}_{t-1} + (1-\beta)x_{t-1}, \tag{3}$$

---

* In Sutton and Barto (1981) and Barto and Sutton (1982), the model used an output trace $\bar{y}_t$ in place of y_{t-1} in Equation 2. However, in all the simulations described there we used only $\bar{y}_t = y_{t-1}$, which is the special case of a trace resulting from letting $\beta = 0$ in Equation 3. Because we now believe that this special case is best for reasons made clear in the theory underlying the TD model, we explicitly specify this case in our restatement of the SB model.

where $0 \leq \beta < 1$.[†]

Equations 1, 2 and 3 constitute the SB model. We can describe the learning process as follows: Activity on any input pathway i, $i = 1, \ldots, n$, can immediately influence the element's output, y, if $w(i) \neq 0$, but also causes that pathway to become "tagged" by the stimulus trace $\bar{x}(i)$ as being eligible for modification in the future (for as long as the trace is nonzero). A connection weight changes only if the pathway is eligible and reinforcement occurs, where reinforcement is defined as a deviation of the current output from the immediately preceeding output (for continuous time, reinforcement is the rate-of-change of the output). Figure 2 shows the time courses of the relevant signals for a single trial with an initially neutral CS.

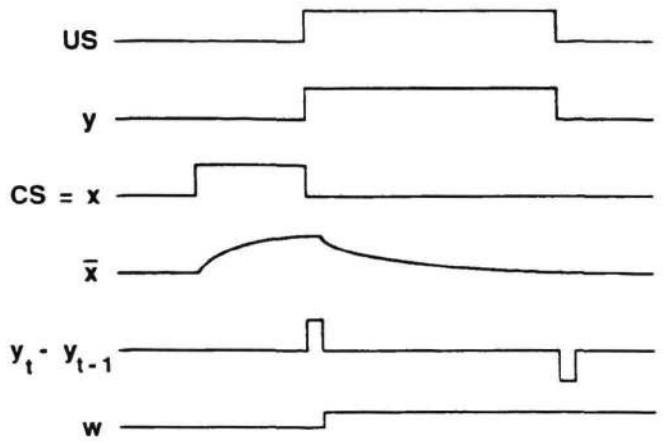

Figure 2. Time courses of element variables in the SB model for a trial in which an initially neutral ($w = 0$) CS is followed by the US.

In Sutton and Barto (1981) and Barto and Sutton (1982) we showed that this model is closely related to the Rescorla-Wagner model and could similarly account for phenomena in classical conditioning such as blocking and conditioned inhibition. Additionally, we showed how it could account for inter-stimulus interval (ISI) effects, anticipatory CRs, and aspects of higher-order and serial-compound conditioning. Recently, a novel prediction of the model concerning blocking and serial-compound conditioning has been tested and

---

[†] In Sutton and Barto (1981), $\bar{x}_t$ was defined as in Equation 3 except that the factor of $(1 - \beta)$ was absent. This factor, which was used in our presentation in Barto and Sutton (1982), simply normalizes the trace in such a way as to ensure that the trace of input that is constant over time will converge to that constant value as $t \rightarrow \infty$.

confirmed by Kehoe, Schreurs, and Graham (in press). That result is discussed further in the section on serial-compound results.

Despite these successes, the SB model suffers from several major problems. In our original presentation of the SB model, we avoided many of the problems by using a US that was very long, which ensured that all CS traces had fallen to zero by the time of US offset. Moore et al. (1986) have since found that if shorter USs are used, the SB model does not generate appropriate conditioning behavior as a function of the CS-US inter-stimulus interval (ISI). For example, if CS onset is simultaneous with or shortly after US onset, then the SB model incorrectly predicts strong inhibitory conditioning to the CS. Even worse, if CS offset is simultaneous with US offset, as in standard delay conditioning, then the unmodified SB model predicts that the CS will fail to acquire a positive association at any ISI.

Moore et al. succeeded in producing a modified version of the SB model, called the Sutton-Barto-Desmond, or SBD, model, that largely solves these and other problems, and also reproduces key features of response topography and CR-related neuronal firing (Moore et al., 1986; Blazis et al., 1986). The primary modifications to the SB model were 1) allowing the effect of the US to vary as a function of current weight values, 2) specifying a particular lagged relationship between CSs and their corresponding signals $x(i)$, and 3) making the trace decay rate β depend on CS duration. Together, these modifications constitute a substantial increase in the complexity of the model. With the TD model, we are attempting to solve the ISI problems of the SB model in a simpler way. The modifications made by Moore et al. to give the SB model a more realistic reponse topography and to relate it to neuronal firings may also be applicable to the TD model, but this has not yet been explored. Space limitations prevent us from making a full comparison of the TD model with the SBD model and with other competing real-time models (e.g., Klopf, 1986, in prep.; Tesauro, 1986; Gluck and Thompson, in press).

The Temporal-Difference (TD) Model

A key desirable feature of the SB model and some other models (Gelperin, Hopfield, Tank, 1985; Hawkins and Kandel, 1984; Klopf, 1986, in prep.; Moore et al., 1986, Tesauro,

1986) is that reinforcement is caused by the onsets and offsets of previously conditioned CSs. Since the US is treated exactly like a previously conditioned CS in the SB model, the US's reinforcing effects also occur at its onset and offset. Experimentally, however, it seems as if simply the presence of the US is reinforcing rather than changes in its presence. This is the basic difference between the SB model and the TD model—in the TD model, US presence itself is directly reinforcing, not its initiation and termination.

We define the TD model by referring to the adaptive element shown in Figure 1. The element's output at time t is

$$y_t = r_t + P(w_t, x_t),$$

where r_t denotes the value at time t of a signal indicating the presence and strength of the US (i.e., r_t is the same as $w_t(0)x_t(0)$ of the SB model) and $P(w_t, x_t)$ is defined by

$$P(w,x) = \begin{cases} \sum_{i=1}^{n} w(i)x(i), & \text{if } \sum_{i=1}^{n} w(i)x(i) > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The weights are updated according to the rule

$$w_{t+1} = w_t + c\Big(r_t + \gamma P(w_t, x_t) - P(w_t, x_{t-1})\Big)\bar{x}_t, \tag{5}$$

where $c > 0$, $0 < \gamma < 1$, and $\bar{x}_t$ is as defined by Equation 3.

This model is similar to the SB model and basically works in the same manner, but it differs from that model in several crucial ways. First, note that the sum P plays a role in the weight update equation similar to the role the output y plays in the SB model (Equation 2): Changes in P over time are critical determinants of weight changes. But here the sum P does not include a contribution from the US as the sum y does in the SB model (Equation 1). The US directly contributes to weight changes through the term r_t in Equation 5. Consequently, in the TD model, the presence of the US (signaled by a nonzero value of r_t), rather than its onset and offset, acts as reinforcement. This is accomplished while retaining the feature of the SB model whereby a CS with an existing association generates reinforcement at its onset and offset (through the CS's contribution to P).

A second major feature distinguishing the TD model from the SB model concerns the parameter γ. The theoretical interpretation of this parameter is discussed in a later section. Here it suffices to point out that this parameter causes a CS with an existing

associative strength to generate reinforcement throughout its presence and not just at its onset and offset. In Equation 5, if P is constant over time, then to the extent that γ is less than 1, reinforcement is still generated. The strength of this reinforcement is proportional to the strength of the CS's existing association, but of opposite sign. The choice of γ determines the relative importance of reinforcement generated by CSs with existing associations due to their constant presence, and due to their onsets and offsets. γ is usually chosen to be near 1 (e.g., $\gamma = .95$ in all simulations described here), so that the presence of a CS generates much less reinforcement than does its onset or offset.

Basic Results

In this section we present simulation results showing the behavior of the TD model in a range of basic conditioning paradigms—single-CS acquisition and extinction, ISI curves for trace and delay conditioning, blocking, conditioned inhibition, and the lack of extinction of conditioned inhibitors. We regard such results as basic because they do not involve complicated temporal relationships between CSs and because previous models have demonstrated each of these abilities. Nevertheless, to our knowledge only the SBD model (Moore et al., 1986; Blazis, 1986) has previously demonstrated all of these abilities.

The parameter values used in all simulations were $c = .01$, $\beta = .8$, and $\gamma = .95$. These values were chosen so as to approximately match ISI data for the rabbit nictitating membrane response (as discussed below), under the interpretation that each time step corresponds to .05 seconds. When a stimulus was present, the corresponding input signal (x or r) was set to 1, and when the stimulus was absent, the signal was set to 0. The time interval between trials was long enough for all traces to fall to zero. Since no stimuli were presented during the inter-trial interval, it is clear from Equation 5 that no weight changes will occur during the bulk of this time. Thus, most of the inter-trial interval was simulated simply by setting the traces to zero.

Figure 3 shows the behavior of the TD model in a single-CS acquisition and extinction paradigm. The temporal relationships among stimuli during the acquisition phase of the experiment are shown in Figure 3A. During extinction, only the CS was presented. Over acquisition trials, the CS gains associative strength in a negatively accelerated way, asymptotically approaching a fixed value. During extinction, associative strength is lost

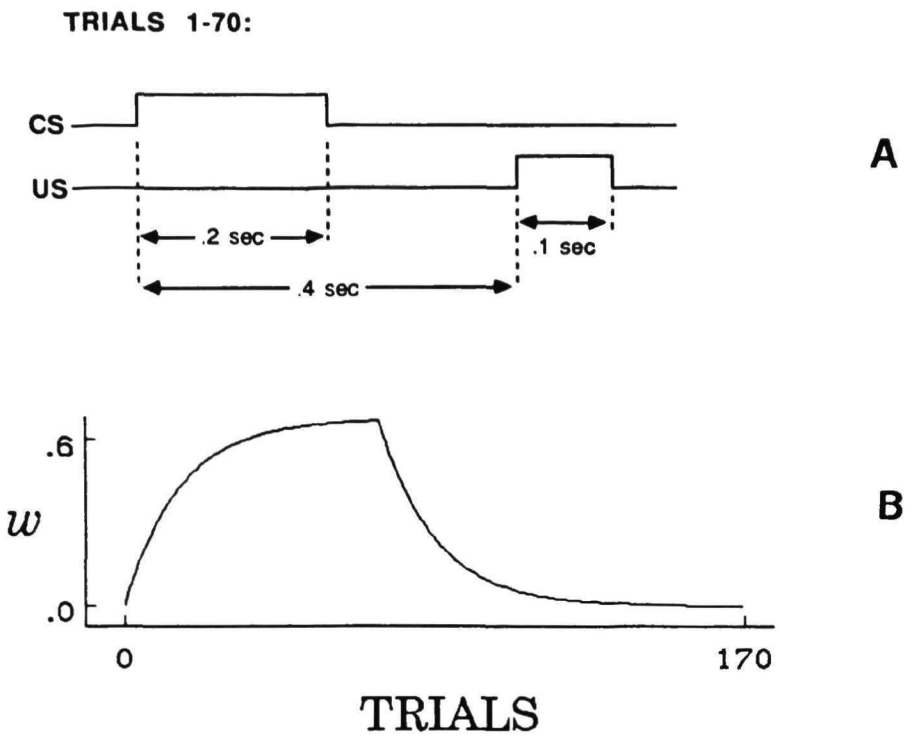

Figure 3. Simulation of Single-CS Acquisition and Extinction in the TD Model. A) Timing relationships between stimuli during acquisition. **B)** The behavior of the weight corresponding to the CS during acquisition (trials 1-70) and extinction (trials 71-170). During extinction, the CS is presented not followed by a US. The time intervals are given in seconds under the interpretation that each time step corresponds to .05 seconds.

in a similar manner.

Figure 4 shows the ISI curves produced by the TD model in trace and delay conditioning experiments. These curves show the final associative strength generated by the TD model after 80 CS-US pairings as a function of the inter-stimulus interval between CS and US. The general shape of these curves is independent of parameter settings, but not important details such as how rapidly associative strength declines as the ISI increases. Roughly speaking, β determines the rate of decline in trace conditioning, and, for fixed β, γ determines the rate of decline in delay conditioning. The parameter values given above were selected to approximate the ISI data for rabbit NMR conditioning shown in Figure 5.

The TD model exhibits complete blocking if first-stage training is conducted until asymptotic associative strength is achieved and if the CS added in the second stage has

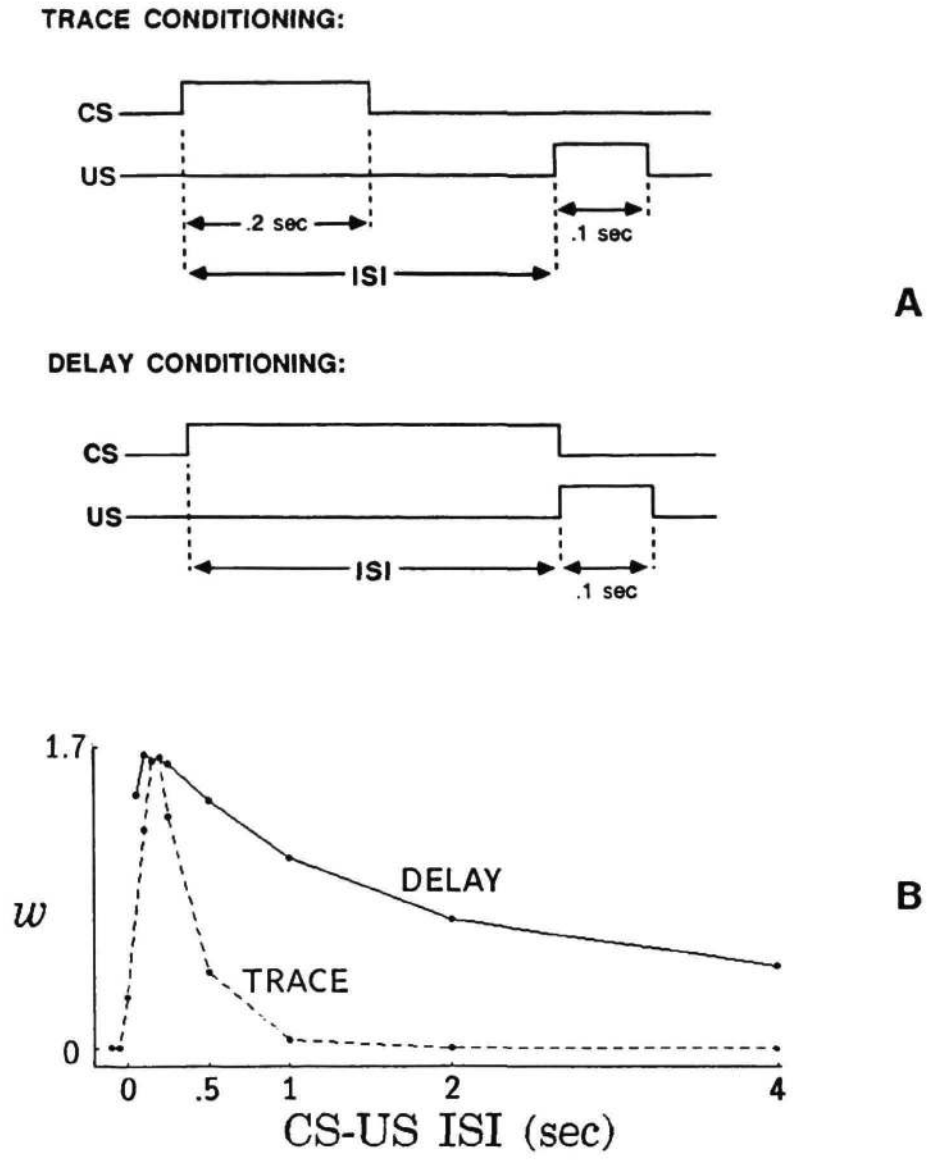

Figure 4. Effect of the CS-US Inter-Stimulus Interval in Trace and Delay Conditioning of the TD Model. **A**) Timing relationships between stimuli in trace and delay acquisition trials. **B**) Resultant CS weight after 80 acquisition trials as a function of ISI.

exactly the same time course as the first CS. This is apparent from inspection of Equation 5—the weights for the two CSs experience exactly the same increments during a second-stage trial; if the weight of the first CS no longer experiences any net change, then neither will the weight of the added CS.

Figure 6 shows the behavior of the TD model in a conditioned inhibition (CI) training regime. In CI, reinforced and unreinforced trials of the two types shown in Figure 6A are

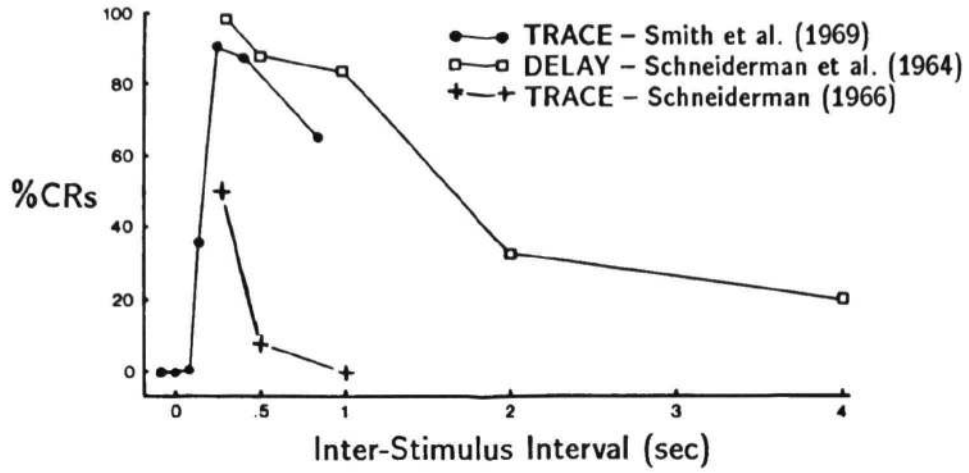

Figure 5. Effect of the CS-US ISI in Trace and Delay Conditioning of the Rabbit Nictitating Membrane Response (NMR). The time course of the ISI dependency varies widely between species and response systems. The parameter values used here in the TD model were chosen so that the model's ISI dependency, shown in Figure 4, approximately matches this rabbit NMR data.

intermixed. CS^+ is followed by the US except in the presence of CS^-. CS^+ is found experimentally to become positively conditioned whereas CS^- becomes a conditioned inhibitor, that is, it tends to inhibit CRs. This result is also found in the simulation. In the extinction phase of the CI experiment shown in Figure 6, both stimuli were presented individually without the US. The result shown is also the same as that found experimentally: The association to the excitor extinguishes, but the association to the inhibitor does not (Zimmer-Hart and Rescorla, 1974). Moore et al. (1986) showed that the SB model will reproduce the desired behavior if the output y is prevented from becoming negative (this corresponds to a particular choice for f in Equation 1), and this is essentially what we have done in the TD model by using a threshold operation in Equation 4.

Serial-Compound Results

Real-time conditioning models are interesting primarily because they make predictions for a wide range of situations that cannot be represented by trial-level models. These situations involve conditionable stimuli that occur together but not strictly simultaneously.

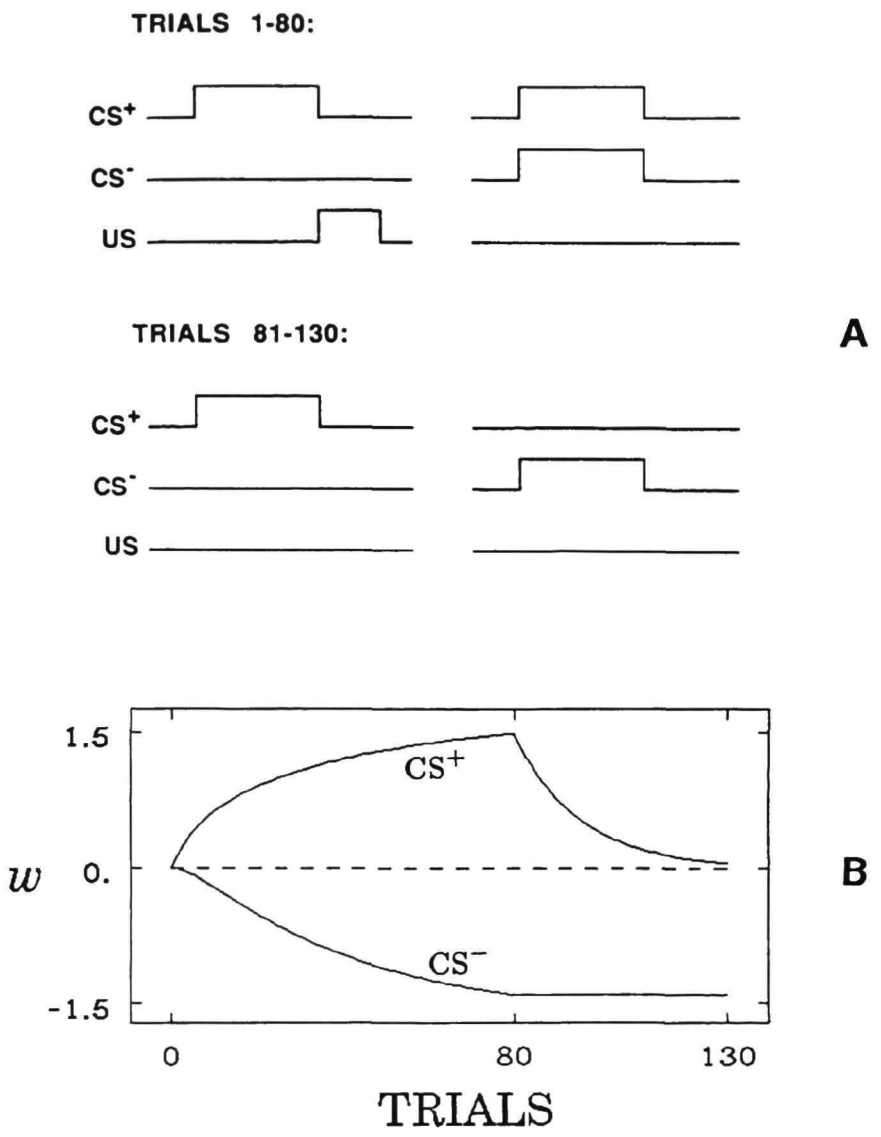

Figure 6. Conditioned Inhibition and its Extinction in the TD Model. A) Time traces showing the two kinds of trials presented alternately in a conditioned inhibition experiment (trials 1-80) and in a subsequent attempt to extinguish the resultant associations (trials 81-130). **B)** Behavior over trials of the weights associated with CS^+ and CS^-. During acquisition, the weight for CS^+ becomes positive, while the weight for CS^- becomes negative. The association to CS^+, but not to CS^-, is extinguished by nonreinforcement. Both CSs are .2 seconds in duration and the US is .1 second in duration.

Any such compound stimulus whose components do not both begin and end at the same time is called a serial-compound stimulus. It should be recognized that almost all learning involves serial-compound stimuli, either because the animal distinguishes earlier and later

portions of a stimulus that may be viewed as a single stimulus by the experimenter, or because the animal's behavior gives rise to a predictable sequence of situations leading to reinforcement, as in maze running. Kehoe (1982) surveys the theoretical issues and empirical results relevant to serial-compound conditioning.

One of the theoretical issues arising in serial-compound conditioning concerns the facilitation of remote associations. It has been found that if an empty trace interval between the CS and the US is filled with a second CS to form a serial compound stimulus, then conditioning to the first CS is facilitated. Figure 7B shows the behavior of the TD model in a simulation of such an experiment, the timing details of which are shown in Figure 7A. Consistent with the experimental results, the model shows facilitation of both the rate of conditioning and the asymptotic level of conditioning of the first CS due of the presence of the second CS.

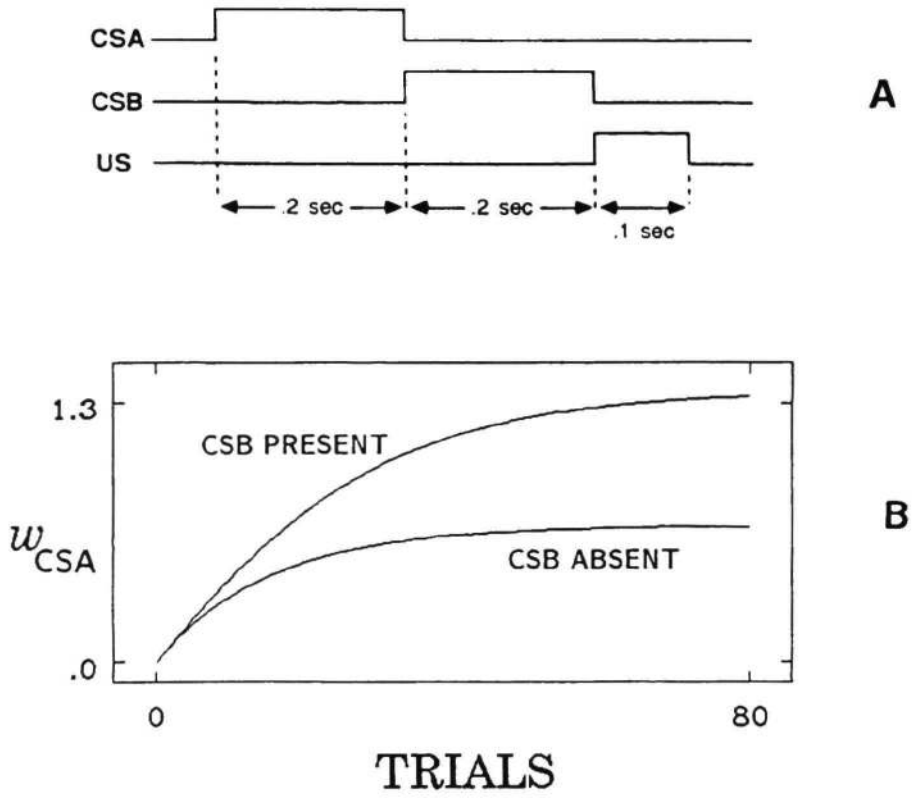

Figure 7. Facilitation of a Remote Association by an Intervening Stimulus in the TD Model. A) Temporal relationships among stimuli within a trial. **B)** The behavior over trials of CSA's weight when CSA is presented in a serial compound, as in **A**, and when presented in an identical temporal relationship to the US, only without the presence of CSB.

The stimulus context effects such as blocking and conditioned inhibition that the Rescorla-Wagner model is so successful at reproducing involve effects on the conditioning of one CS due to the presence of others. However, since it is a trial-level model, the Rescorla-Wagner model does not take into account the temporal relationships between the CSs, which are known to be capable of producing dramatic behavioral consequences. One of the best-known early demonstrations of this is the Egger-Miller (1962) experiment that involved two overlapping CSs in a delay configuration as shown in Figure 8A. Although CSB is in a better temporal relationship with the US, the presence of CSA reduces conditioning to CSB substantially as compared to controls in which CSA is absent. Figure 8B shows the same result being generated by the TD model in a simulation of this experiment.

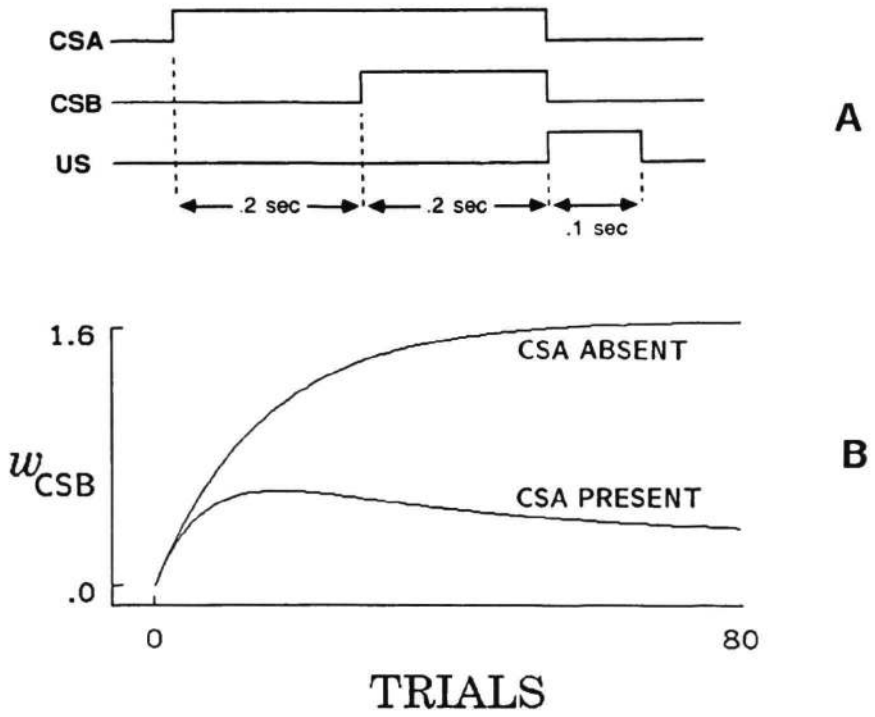

Figure 8. The Egger-Miller or Primacy Effect in the TD Model. A) Temporal relationships among stimuli within a trial. **B)** The behavior over trials of CSB's weight when CSB is presented with and without CSA.

In Sutton and Barto (1981), we presented simulation results with the SB model for an experiment similar to the Egger-Miller experiment discussed above. The experiment we simulated differed from the Egger-Miller experiment in that CSB was given prior training until it was fully associated with the US. When CSA was subsequently introduced, the pre-established association to CSB decreased to zero as training continued. Although

we did not realize it at the time, this is a novel and surprising prediction of the SB model. Why should a well-trained CS that continues to be paired with the US in a good temporal relationship lose associative strength just because a new CS is introduced with no initial association and in a poor temporal relationship? This is a situation in which one might expect the original CS to block and limit association to the new CS. However, the SB model predicts a decrement in the other direction. Recently, Kehoe, Schreurs, and Graham (in press) have tested and confirmed the prediction that CSB will lose associative strength under these conditions. They also note that alternative theories do not make this prediction and have considerable difficulty in explaining this result. The behavior of the TD model under these conditions is shown in Figure 9. This behavior is in slightly better accord with the data than is the SB model's behavior, in that the association to CSB is reduced after the introduction of CSA, but not completely eliminated.

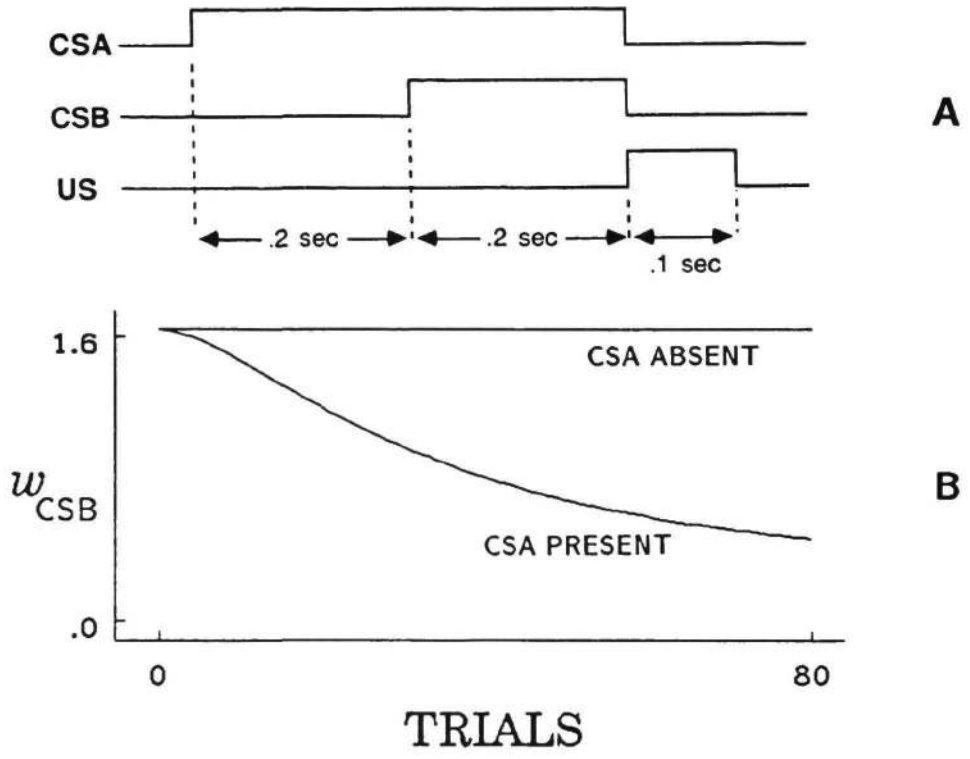

Figure 9. Temporal Primacy Overriding Blocking in the TD Model. A) Temporal relationships between stimuli. **B)** The behavior over trials of CSB's weight when CSB is presented with and without CSA. The only difference between this simulation and that shown in Figure 8 was that here CSB started out fully conditioned—CSB's weight was initially set to 1.653, the final level reached when CSB was presented alone for 80 trials, as in the "CSA-absent" case in Figure 8.

Figure 10 shows the behavior of the TD model in a second-order conditioning experiment. In the first phase (not shown in the figure), CSB is pretrained with the US. In the second phase, CSA is paired with CSB in the sequential arangement shown in Figure 10A, in the absense of the US. Experimentally, CSA is found to acquire associative strength even though it is never paired with the US. In the TD model, CSA first acquires a substantial association and then this association and the original one to CSB are extinguished. This is the same pattern seen experimentally.

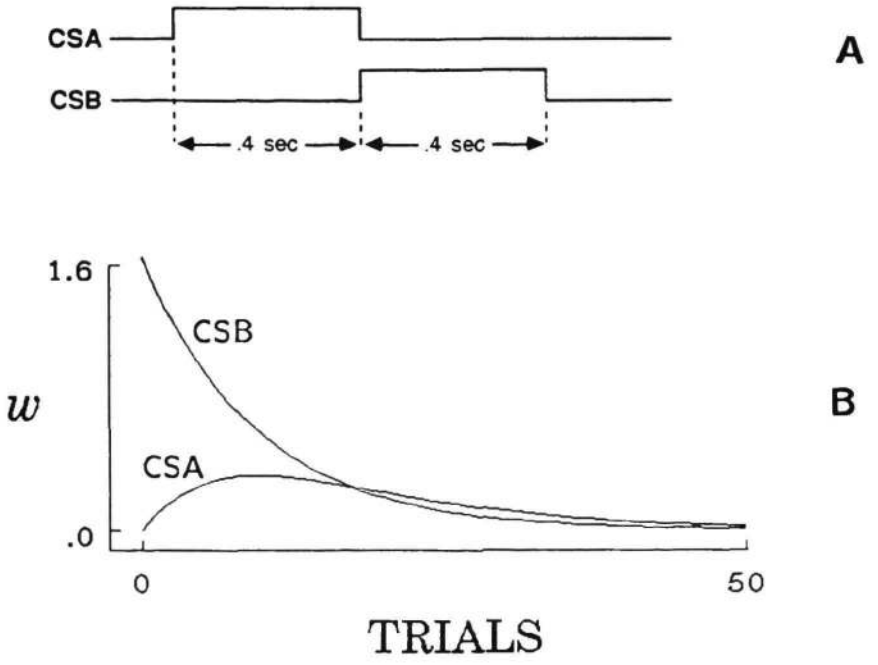

Figure 10. Second-Order Conditioning of the TD Model. **A**) Temporal relationships between stimuli. **B**) The behavior of the weights associated with CSA and CSB over trials. The second stimulus, CSB, has an initial weight of 1.653 at the beginning of the simulation.

Figure 11 shows the ISI curve for the TD model in second-order conditioning. It plots the associative strength after 100 trials as a function of the CSA–CSB ISI. This ISI curve differs significantly from the CS–US ISI curve shown in Figure 4 primarily in that here simultaneous presentation results in the formation of a large negative association instead of a small positive one. Recall that the TD model treats the reinforcement due to USs and previously conditioned CSs differently: US signals directly cause reinforcement whereas *changes* in the signals of previously conditioned CSs cause reinforcement. Thus, in simultaneous presentation, a US's reinforcement is delivered thoughout the presentation,

whereas a previously conditioned CS delivers reinforcement only at its onset, and negative reinforcement at its offset, so that a simultaneously paired CS will be much more affected by the negative reinforcement than by the positive reinforcement.

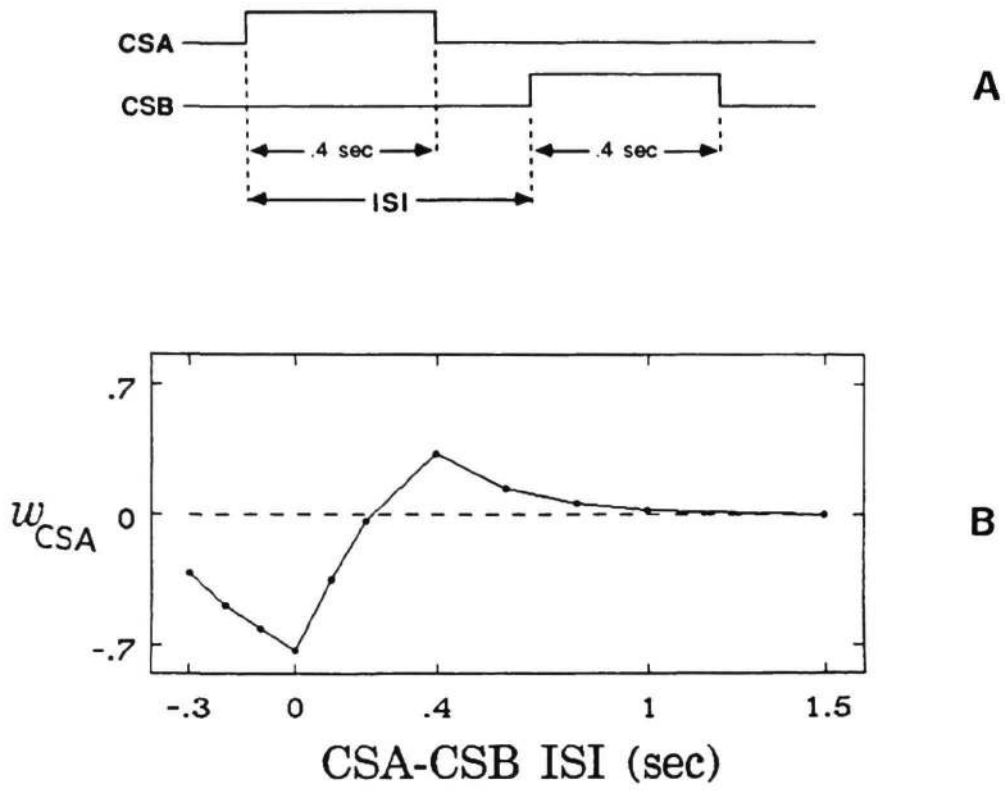

Figure 11. Effect of the CSA-CSB ISI on Second-Order Conditioning of TD Model. A) Temporal relationships between stimuli. **B)** Resultant value of CSA's weight after 10 trials as a function of CSA-CSB ISI.

Experimentally, second-order conditioning is observed to occur with both simultaneous and sequential CSA–CSB pairings. To explain this observation in terms of the TD model we must appeal to indirect associations, which are outside the scope of the model per se. That is, the model clearly predicts that no direct CSA → US association will develop, but does not preclude the development of both CSA → CSB and CSB → US associations, which together could have the same effect. This explanation of second-order conditioning is in fact partially confirmed experimentally. One observed difference between simultaneous and sequential second-order conditioning is that the association to CSA is eliminated by extinguishing CSB in simultaneous second-order conditioning, but not in sequential second-order conditioning (Rescorla, 1980). This suggests that simultaneous second-order conditioning in fact does not result in a direct CSA → US association.

Theoretical Basis of the TD Model

In addition to providing an account of the range of classical conditioning phenomena described above, the TD model has a theoretical basis that suggests an account of the functionality of these phenomena. Sutton (1987) has developed a class of methods for adaptive prediction called temporal-difference (TD) methods and has shown that they have certain advantages over other prediction methods for problems having a certain structure. The advantages of TD methods include reductions in memory requirements, a more even distribution of computation over time, and better generalization from past experience to new situations. If classical conditioning involves prediction, as many believe it does, then TD methods are likely candidates for the underlying learning procedure. Here we provide a brief introduction to the theory as it relates to the TD model.

At each time step t, the subject receives a pattern of CSs represented by the stimulus vector x_t, from which it forms a prediction $P(w, x_t)$, using its current weight vector w. But what does $P(w, x_t)$ predict? Clearly, $P(w, x_t)$ should tell the subject something about the values of the US signal r in the near future. For example, $P(w, x_t)$ might predict something like

$$E\left\{\sum_{k=1}^{N} r_{t+k}\right\},$$

where N is the number of steps remaining in the current trial. The sum is a natural way to have the ideal prediction vary with the intensity, duration, and number of USs occurring on the trial, and the expected value provides a principled way to deal with statistical variation from trial to trial.

However, the particular sum given above, in which all the r_{t+k} values in the rest of the trial are given equal weight, is problematic for two reasons. First, trials and trial boundaries are generally in the mind of the experimenter and unknown to the subject. Second, experimentally the association formed to a CS depends strongly on the time elapsing between it and the US—the more closely the US follows the CS, the stronger the association it will support. This last observation suggests that subjects are predicting a sum in which greater weight is given to r_{t+k} values for smaller values of k. Although there are many ways of varying the weighting with time, the TD model is based on an exponential weighting in which the weight of each r_{t+k}, $k \geq 1$, is γ^{k-1}, for $0 < \gamma < 1$. That is, the TD model is based on the hypothesis that the subject attempts to adjust w

so that, at each time t:

$$P(w, x_t) \approx E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\right\}. \tag{6}$$

The parameter γ is called the *discount rate* because is determines the rate at which later values of r are discounted.

Although the theorems so far obtained for TD methods (Sutton, 1987) do not apply to predicting the quantity given by Equation 6, TD theory nevertheless provides a methodology for constructing a TD learning method specialized for predicting this quantity. The distinguishing feature of TD methods is that the error term they use is the difference between temporally successive predictions. $P(w, x_{t-1})$ and $P(w, x_t)$ are temporally successive predictions, but it is not appropriate to use their difference directly as an error because they are predictions of two different quantities, $P(w, x_{t-1})$ of $E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right\}$, and $P(w, x_t)$ of $E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\right\}$. However, these two predictions are closely related as follows:

$$\begin{aligned} P(w, x_{t-1}) &\approx E\left\{\sum_{k=0}^{\infty} \gamma^k rt + k\right\} \\ &= E\left\{r_t + \sum_{k=1}^{\infty} \gamma^k r_{t+k}\right\} \\ &\approx r_t + \gamma E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\right\} \\ &\approx r_t + \gamma P(w, x_t). \end{aligned}$$

Thus, $r_t + \gamma P(w, x_t)$ is a prediction of the same quantity predicted by $P(w, x_{t-1})$, but it is available one time step later and is based on slightly better information—on the newly-available actual value of r_t and on the new stimulus vector x_t. It is thus the difference between these two predictions, that is, $(r_t + \gamma P(w, x_t)) - P(w, x_{t-1})$, that is used as a reinforcement or error in the TD model's update rule (Equation 5).

The TD model proposed here is not the first model of classical conditioning to be based on changes or temporal differences in net associative strength. This mechanism is a key part of the SB model, and also of the models proposed by Hawkins and Kandel (1984), Gelperin, Hopfield and Tank (1985), Klopf (1986, in prep.), Moore et al. (1986), and Tesauro (1986). What is different about the TD model is that the precise way temporal differences are used is based on a formal, engineering theory of prediction, coupled with a specific proposal for the quantity being predicted.

Limitations and Conclusion

Neither the SB model nor the TD model are complete models of classical conditioning. Among the major classes of phenomena that are beyond the scope of these models and which have been treated by other models are configuration and patterning phenomena (e.g., Kehoe, 1986, and Granger and Schlimmer, 1986), attentional and stimulus selection effects, learning to learn, and learned salience/associability changes (e.g., Moore and Stickney, 1980; Schmajuk and Moore, 1986; Kehoe, 1986), sensory preconditioning and other effects of indirect associations (e.g., Schmajuk and Moore, 1986), CR topography (e.g., Moore et al., 1986; Frey and Sears, 1978), and stimulus preprocessing issues (e.g., Gelperin, Hopfield, and Tank, 1985). Some of these phenomena may be addressable with connectionist mechanisms such as backpropagation (Rumelhart, Hinton, and Williams, 1985) learning-rate adjustment rules (e.g., Frey and Sears, 1978; Sutton, 1986; Barto and Sutton, 1981, Appendix C), and recurrent networks (e.g., Sutton and Barto, 1981a; Sutton and Pinette, 1985).

Although animal learning is complex and subtle, with different processes operating at different levels and time scales, its regularities are far more striking than its variations. Although one theory that explains all animal learning remains a goal, most progress in this area has been made by focussing on identifiable component processes of animal learning. Against this background, the TD model actually represents a substantial integration, since its behavior subsumes nearly all the behavior of the trial-level Rescorla-Wagner model but additionally generates predictions and explanations for within-trial phenomena. The simulations of the TD model described in this paper, together with the theoretical basis of the TD model, suggest that these phenomena might be regarded as consequences of an adaptive process for predicting a discounted sum of future values of the US signal.

References

Anderson, C.W. 1986. Learning and problem solving with multilayer connectionist systems. Ph.D. dissertation, Dept. of Computer and Information Science, University of Massachusetts.

Barto, A.G., Sutton, R.S. 1981. Goal seeking components for adaptive intelligence: An

initial assessment. *Air Force Wright Aeronautical Laboratories/Avionics Laboratory Technical Report AFWAL-TR-81-1070*, Wright-Patterson AFB, Ohio.

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

Barto, A.G., Sutton R.S., Anderson, C.W. 1983. Neuronlike elements that can solve difficult learning control problems. *IEEE Trans. on Systems, Man, and Cybernetics, SMC-13*, No. 5, 834–846.

Blazis, D.E.J., Desmond, J.E., Moore, J.W., Berthier, N.E. 1986. Simulation of the classically conditioned nictitating response by a neuron-like adaptive element: A real-time variant of the Sutton-Barto model. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 176-186.

Duda, R.O., Hart, P.E. 1973. *Pattern Classification and Scene Analysis.* New York: Wiley.

Egger, D.M., Miller, N.E. 1962. Secondary reinforcement in rats as a function of information value and reliability of the stimulus. *Journal of Experimental Psychology 64*: 97–104.

Frey, P.W., Sears, R.J. 1978. Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychological Review 85*: 321–348.

Gelperin, A., Hopfield, J.J., Tank, D.W. 1985. The logic of *Limax* learning. In: *Model Neural Networks and Behavior*, A. Selverston, Ed. New York: Plenum Press.

Gluck, M.A., Thompson, R.F. In press. Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review.*

Hawkins R.D., Kandel, E.R. 1984. Is there a cell-biological alphabet for simple forms of learning? *Psychological Review 91*: 375–391.

Kehoe, E.J. 1982. Conditioning with serial compound stimuli: Theoretical and empirical issues. *Experimental Animal Behavior 1*: 30–65.

Kehoe, E.J. 1986. A layered network model for learning-to-learn and configuration in classical conditioning. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 154–175.

Kehoe, E.J., Schreurs, B.G., Graham, P. In press. Temporal primacy overrides prior training in serial compound conditioning of the rabbit's nictitating membrane response.

Klopf, A.H. 1972. Brain function and adaptive systems—A heterostatic theory. Air Force Cambridge Research Laboratories Special Report No. 133 (AFCRL-72-0164). Also DTIC

Report AD 742259 available from the Defense Technical Information Center, Cameron Station, Alexandria, VA 22304.

Klopf, A.H. 1982. *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence.* New York: Harper & Row / Hemisphere.

Klopf, A.H. 1986. A drive reinforcement model of single neuron function: An alternative to the Hebbian neural model. In J.S. Denker (Ed.) *Neural Networks for Computing*, AIP Conference Proceedings *151*, New York: American Institute of Physics, 265–270.

Klopf, A.H. In preparation. A neuronal model of classical conditioning.

Mackintosh, N.J. 1975. A theory of attention: Variation in the associability of stimuli with reinforcement. *Psychological Review 82*: 276–298.

Moore, J.W., Desmond, J.E., Berthier, N.E., Blazis, D.E.J., Sutton, R.S., Barto, A.G. 1986. Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing and interstimulus intervals. *Behavioral Brain Research 21*: 143–154.

Moore, J.W., Stickney, K.J. 1980. Formation of attentional-associative networks in real time: Role of the hippocampus and implications for conditioning. *Physiological Psychology 8*: 207–217.

Pearce, J.M., Hall, G. 1980. A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review 87*: 532–552.

Rescorla, R.A. 1980. Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology: Animal Behavioral Processes 6*: 339–351.

Rescorla, R.A., Wagner, A.R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical Conditioning II*, A.H. Black and W.F Prokasy, Eds., 64–99. New York: Appleton-Century-Crofts.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. 1985. Learning internal representations by error propagation. Institute for Cognitive Science Technical Report 8506, UCSD, La Jolla, CA 92093. Also in Rumelhart and McClelland (1986), 318–362.

Rumelhart, D.E., McClelland, J.L. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* Cambridge, MA: MIT Press.

Schlimmer, J.C., Granger, R.H. 1986. Simultaneous configural classical conditioning. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 141-153.

Schmajuk, N.A., Moore, J.W. 1986. A real-time attentional-associative network for clas-

sical conditioning of the rabbit's NMR. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 794–807.

Schneiderman, N. 1966. Interstimulus interval function of the nictitating membrane response of the rabbit under delay and trace conditioning. *Journal of Comparative and Physiological Psychology 62*: 397–402.

Schneiderman, N., Gormezano, I. 1964. Conditioning of the nictitating membrane of the rabbit as a function of the CS-US interval. *Journal of Comparative and Physiological Psychology 57*: 188–195.

Smith, M.C., Coleman, S.R., Gormezano, I. 1969. Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous and forward CS-US intervals. *Journal of Comparative and Physiological Psychology 69*: 226–231.

Sutton, R.S. 1984. Temporal credit assignment in reinforcement learning. Ph.D. dissertation, Dept. of Computer and Information Science, University of Massachusetts. Available from the author or as Technical Report #84-2.

Sutton, R.S. 1987. Learning to predict by the methods of temporal differences. Technical Report TR87-509.1, GTE Labs, Waltham, MA.

Sutton, R.S., Barto, A.G. 1981. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review 88*: 135–171.

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

Sutton, R.S., Pinette, B. 1985. The learning of world models by connectionist networks. *Proceedings of the Seventh Annual Conf. of the Cognitive Science Society*, 54–64.

Tesauro, G. 1986. Simple neural models of classical conditioning. *Biological Cybernetics 55*: 187–200.

Wagner, A.R. 1981. SOP: A model of automatic memory processing in animal behavior. In: *Information Processing in Animals: Memory Mechanisms*, N.E. Spear and R.R. Miller, Eds., 5–48. Hillsdale, NJ: Erlbaum.

Widrow B., Hoff, M.E. 1960. Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, 96–104.

Widrow, B., Stearns, S.D. 1985. *Adaptive Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall.

Zimmer-Hart, C.L., Rescorla, R.A. 1974. Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology 86*: 837–845.