

Evaluating the TD model of classical conditioning

Elliot A. Ludvig · Richard S. Sutton · E. James Kehoe

© Psychonomic Society, Inc. 2012

Abstract The temporal-difference (TD) algorithm from reinforcement learning provides a simple method for incrementally learning predictions of upcoming events. Applied to classical conditioning, TD models suppose that animals learn a real-time prediction of the unconditioned stimulus (US) on the basis of all available conditioned stimuli (CSs). In the TD model, similar to other error-correction models, learning is driven by prediction errors—the difference between the change in US prediction and the actual US. With the TD model, however, learning occurs continuously from moment to moment and is not artificially constrained to occur in trials. Accordingly, a key feature of any TD model is the assumption about the representation of a CS on a moment-to-moment basis. Here, we evaluate the performance of the TD model with a heretofore unexplored range of classical conditioning tasks. To do so, we consider three stimulus representations that vary in their degree of temporal generalization and evaluate how the representation influences the performance of the TD model on these conditioning tasks.

Keywords Associative learning · Classical conditioning · Timing · Reinforcement learning

Classical conditioning is the process of learning to predict the future. The temporal-difference (TD) algorithm is an

incremental method for learning predictions about impending outcomes that has been used widely, under the label of *reinforcement learning*, in artificial intelligence and robotics for real-time learning (Sutton & Barto, 1998). In this article, we evaluate a computational model of classical conditioning based on this TD algorithm. As applied to classical conditioning, the TD model supposes that animals use the conditioned stimulus (CS) to predict in real time the upcoming unconditioned stimuli (US) (Sutton & Barto, 1990). The TD model of conditioning has become the leading explanation for conditioning in neuroscience, due to the correspondence between the phasic firing of dopamine neurons and the reward-prediction error that drives learning in the model (Schultz, Dayan, & Montague, 1997; for reviews, see Ludvig, Bellemare, & Pearson, 2011; Maia, 2009; Niv, 2009; Schultz, 2006).

The TD model can be viewed as an extension of the Rescorla–Wagner (RW) learning model, with two additional twists (Rescorla & Wagner, 1972). First, the TD model makes real-time predictions at each moment in a trial, thereby allowing the model to potentially deal with intratrial effects, such as the effects of stimulus timing on learning and the timing of responses within a trial. Second, the TD algorithm uses a slightly different learning rule with important implications. As will be detailed below, at each time step, the TD algorithm compares the current prediction about future US occurrences with the US predictions generated on the last time step. This *temporal difference* in US prediction is compared with any actual US received; if the latter two quantities differ, a *prediction error* is generated. This prediction error is then used to alter the associative strength of recent stimuli, using an error-correction scheme similar to the RW model. This approach to real-time predictions has the advantage of bootstrapping by comparing successive predictions. As a result, the TD model learns whenever there is a change in prediction, and not only when USs are received or omitted. This seemingly subtle difference makes empirical predictions beyond the scope of the RW model. For example, the TD model naturally accounts for second-order conditioning. When an already-established CS occurs, there is an increase in the US prediction and,

E. A. Ludvig (✉)
Princeton Neuroscience Institute and Department of Mechanical
& Aerospace Engineering, Princeton University,
3-N-12 Green Hall,
Princeton, NJ 08542, USA
e-mail: eludvig@princeton.edu

R. S. Sutton
Department of Computing Science, University of Alberta,
Edmonton, AB, Canada
e-mail: rsutton@ualberta.ca

E. J. Kehoe
School of Psychology, University of New South Wales,
Sydney, Australia
e-mail: j.kehoe@unsw.edu.au

thus, a positive prediction error. This prediction error drives learning to the new, preceding CS, producing second-order conditioning (Sutton & Barto, 1990).

In the RW model and many other error-correction models, the associative strength of a single CS trained by itself is a recency-weighted average of the magnitude of all previous US presentations, including trials with no US as the zero point of the continuum (see Kehoe & White, 2002). The relative timing of those previous USs, relative to the CS, does not play a role. So long as the US occurs during the experimenter-designated trial, the US is equivalently included into that running average, which also serves as a prediction of the upcoming US magnitude. In contrast, in TD models, time infuses the prediction process. As was noted above, both the predictions and prediction errors are computed on a moment-by-moment basis. In addition, the predictions themselves can have a longer time horizon, extending beyond the current trial.

In this article, we evaluate the TD model of conditioning on a broader range of behavioral phenomena than have been considered in earlier work on the TD model (e.g., Ludvig, Sutton, & Kehoe, 2008; Ludvig, Sutton, Verbeek, & Kehoe, 2009; Moore & Choi, 1997; Schultz et al., 1997; Sutton & Barto, 1990). In particular, we try to highlight those issues that distinguish the TD model from the RW model (Rescorla & Wagner, 1972). In most situations where the relative timing does not matter, the TD model reduces to the RW model. As outlined in the introduction to this special issue, we focus on the phenomena of timing in conditioning (Group 12) and how stimulus timing can influence fundamental learning phenomena, such as acquisition (Group 1), blocking, and overshadowing (Group 7). To illustrate how the TD model learns in these situations, we present simulations with three stimulus representations, each of which makes different assumptions about the temporal granularity with which animals represent the world.

Model specification

In the TD model, the animal is assumed to combine a representation of the available stimuli with a learned weighting to create an estimate of upcoming USs. These estimated US predictions (V) are generated through a linear combination of a vector (\mathbf{w}) of modifiable weights ($w(i)$) at time step t and a corresponding vector (\mathbf{x}) for the elements of the stimulus representation ($x(i)$):

$$V_t(\mathbf{x}) = \mathbf{w}_t^T \mathbf{x} = \sum_{i=1}^n w_t(i)x(i). \quad (1)$$

This V is an estimate of the *value* in the context of reinforcement learning theory (Sutton & Barto, 1998) and is equivalent to the aggregate *associative strength* central to

many models of conditioning (Pearce & Hall, 1980; Rescorla & Wagner, 1972). We will primarily use the term *US prediction* to refer to this core variable in the model. In the learning algorithm, each element of the stimulus representation (or sensory input) has an associated weight that can be modified on the basis of the accuracy of the US prediction. In the simplest case, every stimulus has a single element that is on (active) when that stimulus is present and off (inactive) otherwise. The modifiable weight would then be directly equivalent to the US prediction supported by that stimulus. Below, we will discuss in detail some more sophisticated stimulus representations.

The US prediction based on available stimuli is then translated into the conditioned response (CR) through a simple response generation mechanism. This explicit rendering of model output into expected behavioral responding allows for more directly testable predictions. There are many possible response rules (e.g., Church & Kirkpatrick, 2001; Frey & Sears, 1978; Moore et al., 1986), but, for our purposes, a simple formalism will suffice. We assume that there is a reflexive mapping from US prediction to CR in the form of a thresholded leaky integrator. The US prediction (V) above the threshold (θ) is integrated in real time with a small decay constant ($0 < \nu < 1$) to generate a response (a):

$$a_t = \nu a_{t-1} + \lfloor V_t(\mathbf{x}_t) \rfloor_{\theta}. \quad (2)$$

The truncated square brackets indicate that only the supra-threshold portion of the US prediction is integrated into the response, which is interpreted as the CR level (see Kehoe, Ludvig, Dudeney, Neufeld, & Sutton, 2008; Ludvig et al., 2009). This response measure can be readily mapped in a monotonic fashion onto either continuous measures (e.g., lick rate, suppression ratios, food cup approach time) or response likelihood measures based on discrete responses. Thus, comparisons can be conducted across experiments on the basis of ordinal relationships, rather than differing preparation-specific levels (cf. Rescorla & Wagner, 1972).

All learning in the model takes place through changes in these modifiable weights. These updates are accomplished through the TD learning algorithm (Sutton, 1988; Sutton & Barto, 1990, 1998). First, the TD or reward-prediction error (δ_t) is calculated on every time step on the basis of the difference between the sum of the US intensity (r_t) and the new US prediction from the current time step ($V_t(\mathbf{x}_t)$), appropriately discounted, and the US prediction from the last time step ($V_t(\mathbf{x}_{t-1})$):

$$\delta_t = r_t + \gamma V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}), \quad (3a)$$

where γ is the discount factor (between 0 and 1). A positive prediction error is generated whenever the world (US plus new predictions) exceeds expectations (the old

US prediction), and a negative prediction error is generated whenever the world falls short of expectations. Alternatively, rearranging terms, the TD error can be expressed as the difference between the US intensity and the change in US prediction:

$$\delta_t = r_t - [V_t(\mathbf{x}_{t-1}) - \gamma V_t(\mathbf{x}_t)]. \quad (3b)$$

This formulation emphasizes the similarity with the RW rule, where a simple difference between the US intensity and a prediction of that intensity drives learning. In this formulation, a positive prediction error occurs when the US intensity exceeds a *temporal difference* in US prediction, and a negative prediction error occurs whenever the US intensity falls below the temporal difference in US prediction. Note that if γ is 0, the prediction error is identical to the RW prediction error, making real-time RW a special case of TD learning.

This TD error is then used to update the modifiable weights for each element of the stimulus representation on the basis of the following update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \delta_t \mathbf{e}_t \quad (4)$$

where α is a learning-rate parameter and \mathbf{e}_t is a vector of eligibility trace levels for each of the stimulus elements. These eligibility traces determine how modifiable each particular weight is at a given moment in time. Weights for recently active stimulus elements will have high corresponding eligibility traces, thereby allowing for larger changes. In the context of classical conditioning, this feature of the model means that faster conditioning will usually occur for elements proximal to the US and slower conditioning for elements remote from it. More generally, the eligibility traces effectively solve the problem of *temporal credit assignment*: how to decide among all antecedent events which was most responsible for the current reward. These eligibility traces accumulate in the presence of the appropriate stimulus element and decay continuously according to $\gamma\lambda$:

$$\mathbf{e}_{t+1} = \gamma\lambda \mathbf{e}_t + \mathbf{x}_t \quad (5)$$

where γ is the discount factor, as above, and λ is a decay parameter (between 0 and 1) that determines the plasticity window. In the reinforcement learning literature, this learning algorithm is known as TD (λ) with linear function approximation (Sutton & Barto, 1998). We now turn to the three stimulus representations with different temporal profiles that provide the features that are used by the TD learning algorithm to generate the US prediction.

Presence representation

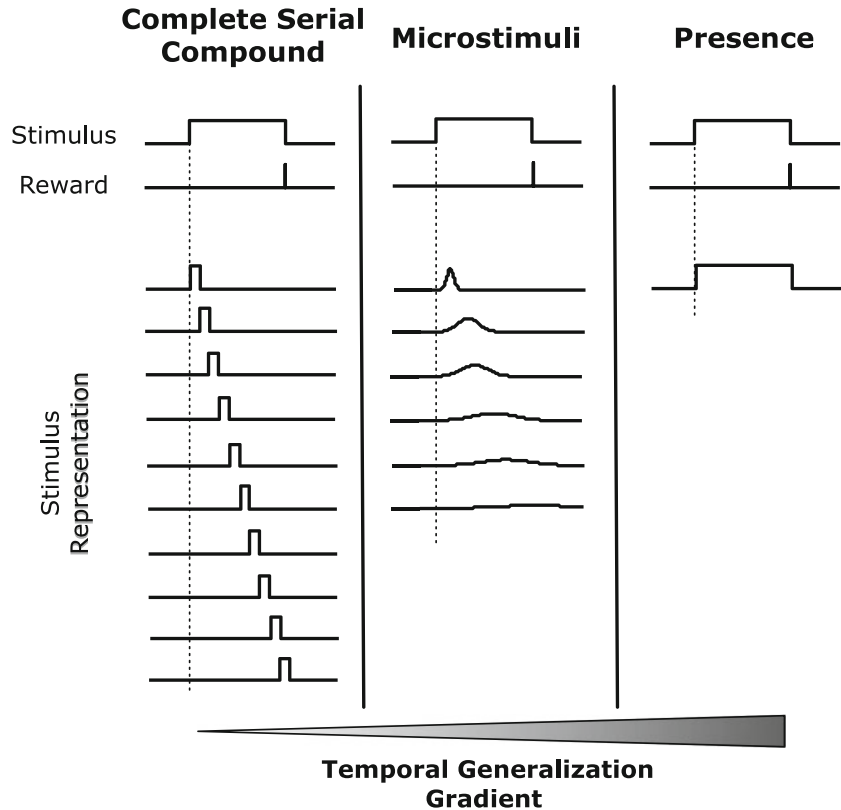
Perhaps the simplest stimulus representation has each stimulus correspond to a single representational element.

Figure 1 depicts a schematic of this representation (right column), along with other more complex representations (see below). This *presence* representation corresponds directly to the stimulus (top row in Fig. 1) and is on when the stimulus is present and off when the stimulus is not present. In Fig. 1, the representations are arranged along a gradient of temporal generalization, and the presence representation rests at one end, with complete temporal generalization between all moments in a stimulus. Although an obvious simplification, this approach, in combination with an appropriate learning rule, can accomplish a surprisingly wide range of real-time learning phenomena (Sutton & Barto, 1981, 1990). Sutton and Barto (1990) demonstrated that the TD learning rule, in conjunction with this stimulus representation, was sufficient to reproduce the effects of interstimulus interval (ISI) on the rate of acquisition, as well as blocking, second-order conditioning, and some temporal primacy effects (e.g., Egger & Miller, 1962; Kehoe, Schreurs, & Graham, 1987). This presence representation suffers from the obvious fault that there is complete generalization (or a lack of temporal differentiation) across all time points in a stimulus. Early parts of a stimulus are no different than the later parts of the stimulus; thus, a graded or timed US prediction across the stimulus is impossible.

Complete serial compound

At the opposite end from a single representational element per stimulus is a separate representational element for every moment in the stimulus (first column in Fig. 1). The motivating idea is that extended stimuli are not coherent wholes but, rather, temporally differentiated into a *serial compound* of temporal elements. In a *complete serial compound* (CSC), every time step in a stimulus is a unique element (separate rows in Fig. 1). This CSC representation is used in most of the TD models of dopamine function (e.g., Montague, Dayan, & Sejnowski, 1996; Schultz, 2006; Schultz et al., 1997) and is often taken as synonymous with the TD model (e.g., Amundson & Miller, 2008; Church & Kirkpatrick, 2001; Desmond & Moore, 1988; Jennings & Kirkpatrick, 2006), although alternate proposals do exist (Daw, Courville, & Touretzky, 2006; Ludvig et al., 2008; Suri & Schultz, 1999). Although clearly biologically unrealistic, in a behavioral model, this CSC representation can serve as a useful fiction that allows examination of different learning rules unfettered by constraints from the stimulus representation (Sutton & Barto, 1990). This stimulus representation occupies the far pole along the temporal generalization gradient from the presence representation (see Fig. 1). There is complete temporal differentiation of every moment and no temporal generalization whatsoever.

Fig. 1 The three stimulus representations (in columns) used with the TD model. Each row represents one element of the stimulus representation. The three representations vary along a temporal generalization gradient, with no generalization between nearby time points in the complete serial compound (left column) and complete generalization between nearby time points in the presence representation (right column). The microstimulus representation occupies a middle ground. The degree of temporal generalization determines the temporal granularity with which US predictions are learned



Microstimulus representation

A third stimulus representation occupies an intermediate zone of limited temporal generalization between complete temporal generalization (presence) and no temporal generalization (CSC). The middle column of Fig. 1 depicts what such a representation looks like: Each successive row presents a *microstimulus* (MS) that is wider and shorter and peaks later.

The MS temporal stimulus representation is determined through two components: an exponentially decaying memory trace and a coarse-coding of that memory trace. The memory trace (y) is initiated to 1 at stimulus onset and decays as a simple exponential:

$$y_{t+1} = dy_t, \quad (6)$$

where d is a decay parameter ($0 < d < 1$). Importantly, there is a memory trace and corresponding set of microstimuli for every stimulus, including the US. This memory trace is coarse coded through a set of basis function across the height of the trace (see Ludvig et al., 2008; Ludvig et al., 2009). For these basis functions, we used nonnormalized Gaussians:

$$f(y, \mu, \sigma) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right), \quad (7)$$

where y is the exponentially decaying memory trace as above, \exp is the exponential function, and μ is the mean

and σ the width of the basis function. These basis functions can be thought of as equally spaced receptive fields that are triggered when the memory trace decays to the appropriate height for that receptive field. The strength x of each MS i at each time point t is determined by the proximity of the current height of the memory trace (y_t) to the center of the corresponding basis function multiplied by the trace height at that time point:

$$x_t(i) = f(y_t, i/m, \sigma)y_t, \quad (8)$$

where m is the total number of microstimuli for each stimulus. Because the basis functions are spaced linearly but the memory trace decays exponentially, the resultant temporal extent (width) of the microstimuli varies, with later microstimuli lasting longer than earlier microstimuli (see the middle column of Fig. 1), even with a constant width of the basis function.

The resulting microstimuli bear a resemblance to the spectral traces of Grossberg and Schmajuk (1989; also Brown, Bullock, & Grossberg, 1999; Buhusi & Schmajuk, 1999), as well as the behavioral states of Machado (1997). We do not claim that the exact quantitative form of these microstimuli is critical to the performance of the TD model below, but rather, we are examining how the general idea of a series of broadening microstimuli with increasing temporal delays interacts with the TD learning rule. Our claim will be that introducing a form of limited temporal generalization

into the stimulus representation and using the TD learning rule creates a model that captures a broader space of conditioning phenomena.

To be clear, here is a fully worked example of what the TD model learns in simple CS–US acquisition. In this example, the CS–US interval is 25 time steps, and we assume a CSC representation and one-step eligibility traces ($\lambda = 0$). First, imagine the time point at which the US occurs (time step 25). On the first trial, the weight for time step 25 is updated toward the US intensity, but nothing else changes. On the second trial, the weight for time step 24 also gets updated, because there is now a temporal difference in the US prediction between time step 24 and time step 25. The weight for time step 24 is moved toward the discounted weight for time step 25 (bootstrapping). The weight for time step 25 gets updated as before after the US is received. On the third trial, the weight for time step 23 would also get updated, because there is now a temporal difference between the US predictions at time steps 23 and 24. . . . Eventually, across trials, the prediction errors (and thus, US predictions) percolate back one step at a time to the earliest time steps in the CS. This process continues toward an asymptote, where the US prediction at time step 25 matches the US intensity and the predictions for earlier time steps match the appropriately discounted US intensity. If we take the US intensity to be 1, because of the discounting, the asymptotic prediction for time step 24 is γ , which is equal to 0 (the US intensity at that time step) + $1 \times \gamma$ (the discounted US intensity from the next step). Following the same logic, the asymptotic predictions for earlier time points in the stimulus are γ^2 , γ^3 , γ^4 , . . . and so on, forming an exponential US prediction curve (see also Eq. 9). Introducing multistep eligibility traces ($\lambda > 0$) maintains these asymptotic predictions but allows for the prediction errors to percolate back through the CS faster than one step per trial.

For the simulations below, we chose a single set of parameters to illustrate the qualitative properties of the model, rather than attempting to maximize goodness of fit to any single data set or subset of findings. By using a fixed set of parameters, we tested the ability of the model to reproduce the ordinal relationships for a broad range of phenomena, thus ascertaining the scope of the model in a consistent manner (cf. Rescorla & Wagner, 1972, p. 77). The full set of parameters is listed in the [Appendix](#).

Simulation results

Acquisition set

For this first set of results, we simulated the acquisition of a CR with the TD model, explicitly comparing the three stimulus representations described previously (see item

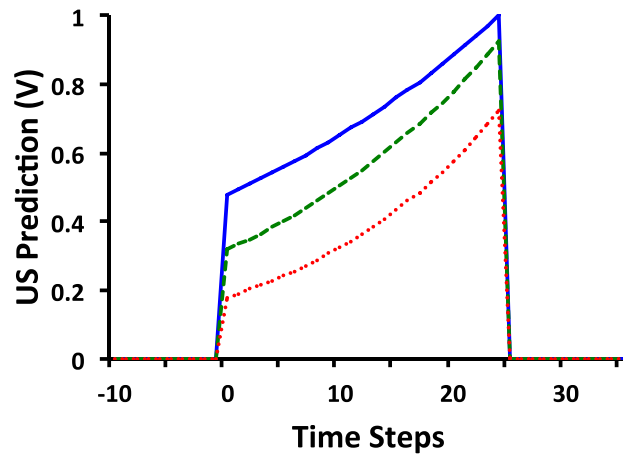
#1.1 listed in the introduction to this issue). We focus here on the timing of the response during acquisition (#12.4) and the effect of ISI on the speed and asymptote of learning (#12.1).

Figure 2 depicts the time course of the US prediction during a single trial at different points during acquisition for the three representations. In these simulations, the US occurred 25 time steps after the CS on all trials. For the CSC representation, the TD model gradually learns a US prediction curve that increases exponentially in strength through the stimulus interval until reaching a maximum of 1 at exactly the time the US occurred (time step 25). This exponential increase is due to the discounting in the TD learning rule (see Eq. 3). That is, at each time point, the learning algorithm updates the previous US prediction toward the US received plus the current discounted US prediction.

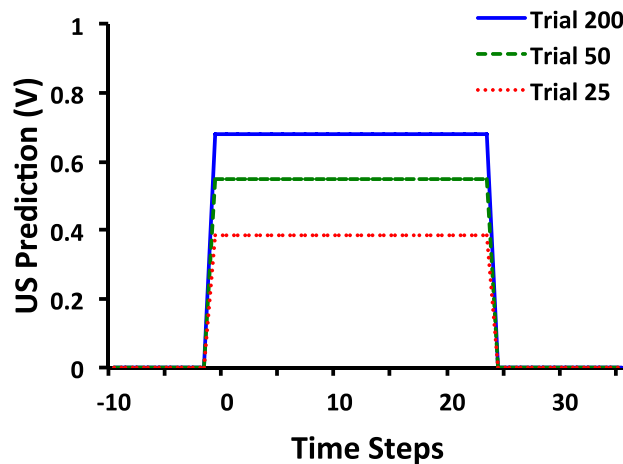
With a CSC representation, each time point has a separate weight, and therefore, the TD model perfectly produces an exponentially increasing US prediction curve at asymptote (Fig. 2a). For the other two representations, the representations at different time steps are not independent, so the updates for one time step directly alter the US prediction at other time steps through generalization. Given this interdependence, the TD learning mechanism produces an asymptotic US prediction curve that best approximates the exponentially increasing curve observed with the CSC, using the available representational elements. With the presence representation (Fig. 2b), there is only a single weight per stimulus that can be learned. As a result, the algorithm gradually converges on a weight that is slightly below the halfway point between the US intensity (1) and the discounted US prediction at the onset of the stimulus ($\gamma^t = .97^{25} \approx .47$, where t is the number of time steps in the stimulus). The US prediction is nearly constant throughout the stimulus, modulo a barely visible decrease due to nonreinforcement during the stimulus interval. This asymptotic weight is a product of the small negative prediction errors at each time step when the US does not arrive and the large positive prediction error at US delivery. With a near-constant US prediction, the TD model with the presence representation cannot recreate many features of response timing (see also Fig. 4b).

Finally, as is shown in Fig. 2c, the TD algorithm with the MS representation also converges at asymptote to a weighted sum of the different MSs that approximates the exponentially increasing US prediction curve. Even with only six MSs per stimulus, a reasonable approximation is found after 200 trials. An important feature of the MS representation is brought to the fore here. The US also acts as a stimulus with its own attendant MSs. These MSs gain strong negative weights (because the US is not followed by another US) and, thus, counteract any residual prediction about the US that would be produced by the CS MSs during the post-US period.

A. Complete Serial Compound



B. Presence



C. Microstimulus

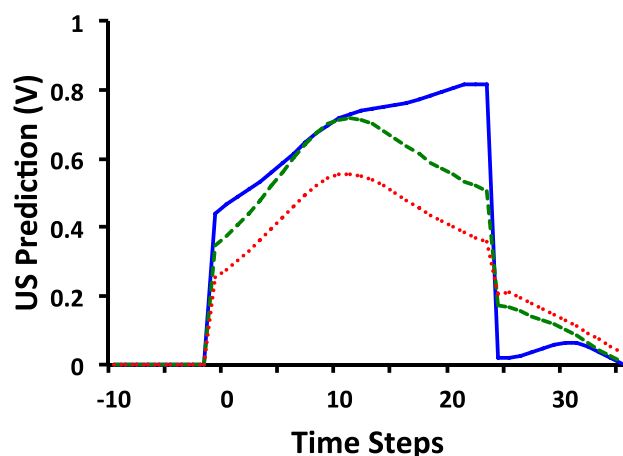


Figure 3 shows how changing the ISI influences the acquisition of responding in the TD model with the three stimulus representations (#12.1). Empirically, the shortest

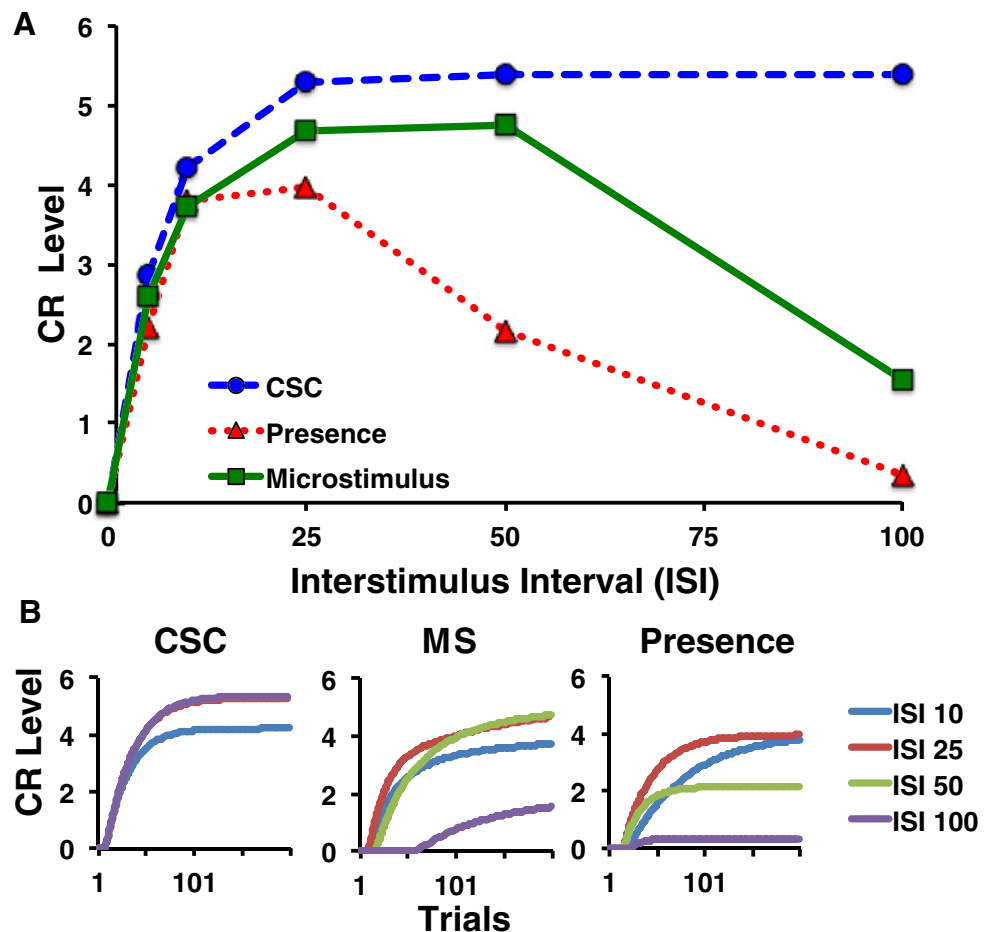
Fig. 2 Time course of US prediction over the course of acquisition for the TD model with three different stimulus representations. **a** With the complete serial compound (CSC), the US prediction increases exponentially through the interval, peaking at the time of the US. At asymptote (trial 200), the US prediction peaks at the US intensity (1 in these simulations). **b** With the presence representation, the US prediction converges to an almost constant level. This constant level is determined by the US intensity and the length of the CS–US interval. **c** With the microstimulus representation, at asymptote, the TD model approximates the exponential decaying time course depicted with the CSC through a linear combination of the different microstimuli

and longest ISIs are often learned more slowly and to a lower asymptote (e.g., Smith, Coleman, & Gormezano, 1969). In this simulation, there were six ISIs (0, 5, 10, 25, 50, and 100 time steps). The top panel (Fig. 3a) depicts the full learning curves for each representation and the four longer ISIs, and the bottom panel displays the maximum level of responding in the TD model on the final trial of acquisition for each ISI (trial 200). With the CSC representation, the ISI has limited effect. Once the ISI gets long enough, the CSC version of the TD model always converges to the same point. That similarity is because the US prediction curve is learned at the same speed and with the same shape independently of ISI (see also Fig. 4).

With the presence representation, there is a substantial decrease in asymptotic CR level with longer ISIs and a small decrease at short ISIs—similar to what has been shown previously (e.g., Fig. 18 in Sutton & Barto, 1990). The decrease with longer ISIs occurs because there is only a single weight with the presence representation. That weight is adjusted on every trial by the many negative prediction errors early in the CS and the single positive prediction error at US onset. With longer ISIs, there are more time points (early in the stimulus) with small negative prediction errors. With only a single weight to be learned and no temporal differentiation, the presence representation thereby results in lower US prediction (and less responding) with longer ISIs. With very short ISIs, in contrast, the high learned US prediction does not quite have enough time to accumulate through the response generation mechanism (Eq. 2), producing a small dip with short ISIs (which is even more pronounced for the MS representation). There is an interesting interaction between learning rate and asymptotic response levels with the presence representation. Although the longer ISIs produce lower rates of asymptotic conditioning, they are actually learned about more quickly because the eligibility trace, which helps determine the learning rate, accumulates to a higher level with longer stimuli (see right panel of Fig. 3b). In sum, with the presence representation, longer stimuli are learned about more quickly, but to a lower asymptote.

With the MS representation, a similar inverted-U pattern emerges. The longest ISI is now both learned about most slowly and to a lower asymptote. The slower learning

Fig. 3 **a** Conditioned response (CR) level after 200 trials as a function of interstimulus interval (ISI) for the three different representations. The complete serial compound (CSC) representation produces a higher asymptote with longer ISIs, whereas the other two representations produce more of an inverted-U-shaped curve, in better correspondence with the empirical data. **b** Learning curves as a function of ISI for the different representations. The learning curves as a whole show a similar pattern to the asymptotic levels, with the key exception that the presence representation produces an interaction between learning rate and asymptotic levels. MS = microstimulus representation



emerges because the late MSs are shorter than the early MSs (see Fig. 1) and, therefore, produce lower eligibility traces. The lower asymptote emerges because the MSs are inexact and also extend beyond the point of the US. The degree of inexactness grows with ISI. As a result, the maximal US prediction does not get as close to 1 with longer ISIs, and a lower asymptotic response is generated. In this simulation with 200 trials, it is primarily the low learning rate due to low eligibility that limits the CR level with long ISIs, but a similar qualitative effect emerges even with more trials (or a larger learning rate; not shown). Finally, as with the presence representation, the shortest ISI produces less responding because of the accumulation component of the response generation mechanism.

Timing set

For our second set of simulations, we consider in greater detail the issue of response timing during conditioning (cf. Items #12.4, #12.5, #12.6, and #12.9 as listed in the Introduction to this special issue). In these simulations, there were four different ISIs: 10, 25, 50, and 100 time steps. Simulations were run for 500 trials, and every 5th trial was a probe trial. On those probe trials, the US was not presented,

and the CS remained on for twice the duration of the ISI to evaluate response timing unfettered by the termination of the CS (analogous to the peak procedure from operant conditioning; Roberts, 1981; for examples of a similar procedure with eyeblink conditioning, see Kehoe et al., 2008; Kehoe, Olsen, Ludvig, & Sutton, 2009).

Figure 4 illustrates the CR time course for the different stimulus representations and ISIs. As can be seen in the top row, with a CSC representation (left column), the TD model displays a CR time course that is sharply peaked at the exact time of US presentation, even in the absence of the US. In line with those results, the bottom row (Fig. 4b) shows how the peak time is perfectly aligned with the US presentation from the very first CR that is emitted by the model. This precision is due to the perfect timing inherent in the stimulus representation and is incongruent with empirical data, which generally show a less precise timing curve (e.g., Kehoe, Olsen, et al., 2009; Smith, 1968). In addition, the time courses are exactly the same for the different ISIs, only translated along the time axis. There is no change in the width or spread of the response curve with longer ISIs, again in contrast to the empirical data (Kehoe, Olsen, et al., 2009; Smith, 1968). Note again how the maximum response levels are the same for all the ISIs (cf. Fig. 3).

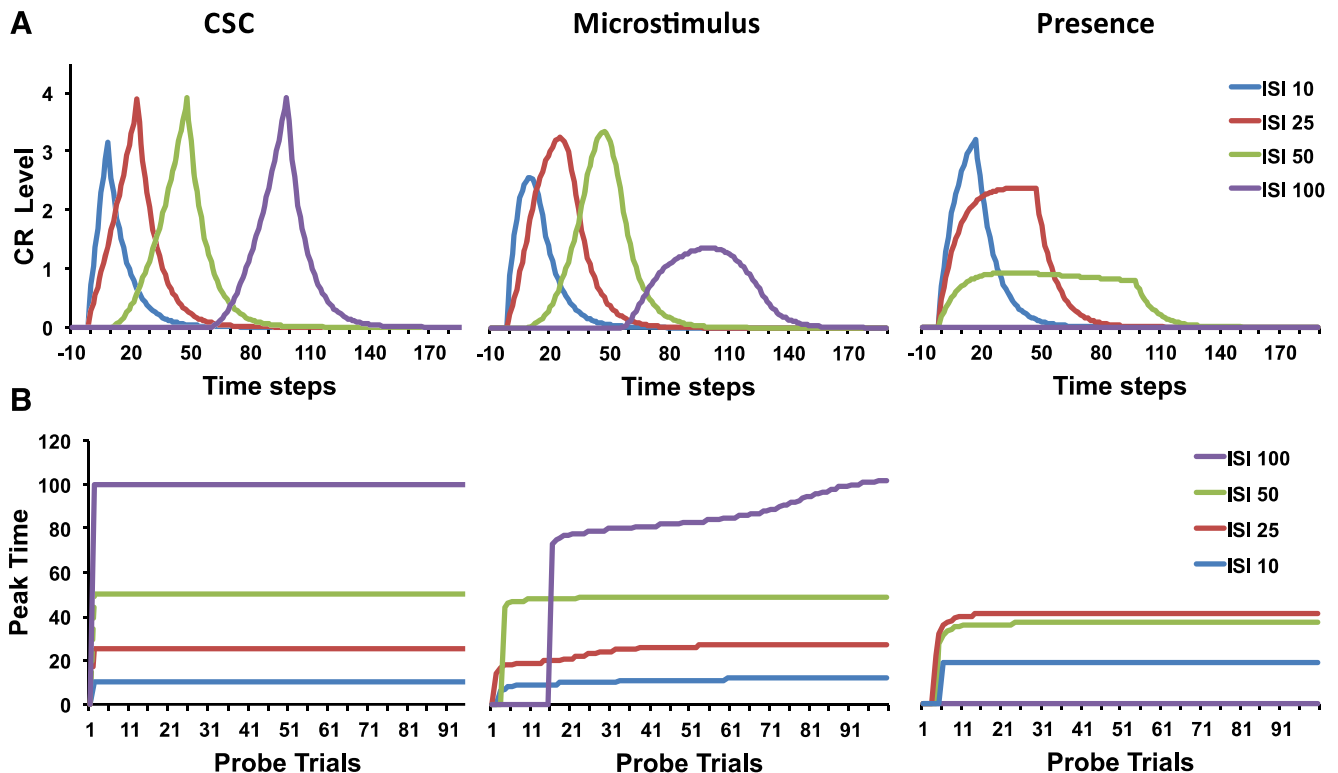


Fig. 4 Timing of the conditioned response (CR) on probe trials with different interstimulus intervals (ISIs). **a** Time course of responding on a single probe trial. For all three stimulus representations, the response peaks near the time of usual US presentation. The key difference is the sharpness in these response peaks; there is too much temporal specificity for the complete serial compound

(CSC) representation and too little for the presence representation. **b** Peak times over the course of learning. Over time, the peak response time changes very little for the CSC and presence representations. For the microstimulus representation, the peak times tend to initially occur a little too early but gradually shift later as learning progresses

With the MS representation (middle column in Fig. 4), model performance is greatly improved. The response curve gradually increases to a maximum around the usual time of US presentation and gradually decreases afterward for all four ISIs. The response curves are wider for the longer ISIs (although not quite proportionally), approximating one aspect of scalar timing, despite a deterministic representation. The times of the simulated CR peaks (bottom row) are also reasonably well aligned with actual CRs from their first appearance (see Drew, Zupan, Cooke, Couvillon, & Balsam, 2005; Kehoe et al., 2008). In contrast to the CSC representation, the simulated peaks occur early during the initial portion of acquisition but drift later as learning progresses. This effect is most pronounced for the longest ISI—in line with the empirical data for eyeblink conditioning (Kehoe et al., 2008; Vogel, Brandon, & Wagner, 2003). At asymptote, the CR peaks occur slightly (1 or 2 time steps) later than the usual time of US presentation, due to a combination of the estimation error inherent in the coarse stimulus representation and the slight lag induced by the leaky integration of the real-time US prediction (Eq. 2).

Finally, the TD model with the presence representation does very poorly at response timing, as might be expected

given that no temporal differentiation in the US prediction curve is possible. The simulated CRs peak too late for the short ISIs (10 and 25 time steps), too early for the medium ISI (50 time steps), and there is no response curve at all for the longest ISI (100 time steps). The late peaks for the short ISI reflect the continued accumulation of the US prediction by the response generation mechanism (Eq. 2) well past the usual time of the US as the CS continues to be present. The disappearance of a response for the longest ISI is due to the addition of probe trials in this simulation (cf. Fig. 3). With only a single representational element, the probe trials are particularly detrimental to learning with the presence representation. Not only is no US present on the probe trials, but also the CS (and thus, the presence element) is extended. This protracted period of unreinforced CS presentation operates as a doubly long extinction trial, driving down the weight for the lone representational element and reducing the overall US prediction below the threshold for the response generation mechanism. Indeed, making the probe trials longer would drive down the US prediction even further, potentially eliminating responding for the shorter ISIs as well. The other representations do not suffer from this shortcoming, because the decline in the weights of elements during the extended

portion of the CS generalizes only weakly, if at all, to earlier portions of the CS.

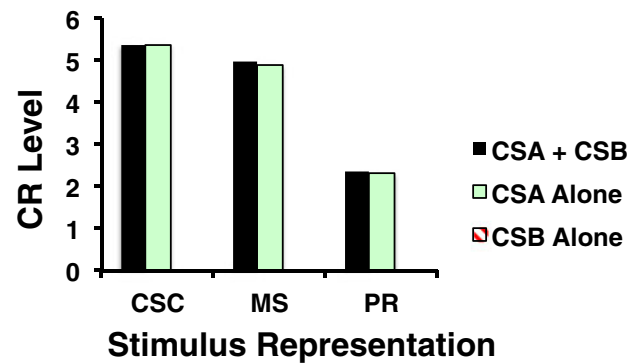
Cue competition set

For this third set of simulations, we examined how the TD model deals with a pair of basic cue competition effects. As listed in the introduction to this special issue, these effects include blocking (#7.2) and overshadowing (#7.5), with a focus on how stimulus timing plays a role in modulating these effects (#12.7; #7.11). Figure 5 depicts simulated responding in the model following three variations of a blocking experiment in which one stimulus (CSA) was given initial reinforced training at a fixed ISI and then a second stimulus (CSB) was added, using the same ISI, a shorter ISI, or a longer ISI. Thus, in the resulting compound, the onset of the added, blocked stimulus (CSB) occurred at the same time, later, or earlier than the blocking stimulus (CSA) (cf. Jennings & Kirkpatrick, 2006; Kehoe, Schreurs, & Amodei, 1981; Kehoe et al., 1987). In the simulations, CSA was first trained individually for 200 trials, and then CSA and CSB were trained in compound for a further 200 trials. Following this training, probe trials were run with CSA alone, CSB alone, or the compound of the two stimuli.

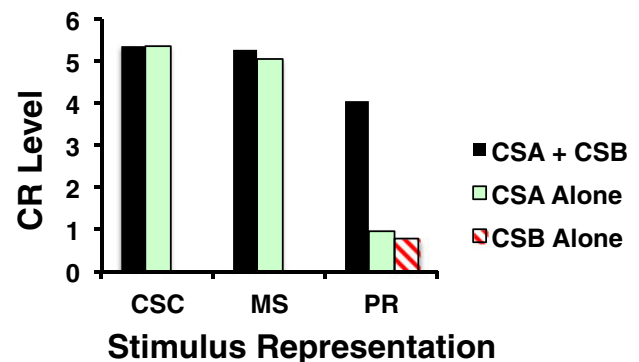
The top panel of Fig. 5 (Fig. 5a) shows responding on the probe trials when both CSs had identical ISIs (50 time steps). With all three representations, and in accord with the empirical data (e.g., Cole & McNally, 2007; Jennings & Kirkpatrick, 2006; Kehoe et al., 1981), there is complete blocking of responding to CSB but high levels of responding to both CSA and the CSA + CSB compound stimulus. With the CSC representation, the blocking occurs because the US is perfectly predicted by the pretrained CSA after the first phase of training; thus, there is no prediction error and no learning to the added CSB. With the presence and MS representations, the US is not perfectly predicted by CSA after the first phase (cf. Fig. 2). As a result, when the US occurs, there is still a positive prediction error, which causes an increase in the weights of eligible CS elements. The increment in the weights induced by this prediction error, however, is perfectly cancelled out by the ongoing, small negative prediction errors during the compound stimulus on the next trial. No net increase in weights occurs to either stimulus, resulting in blocking to the newly introduced CSB (see the section on acquisition above).

If, instead, the added CSB starts later than the pretrained CSA (see Fig. 5b), the simulated results change somewhat, but only for the TD model with the presence representation. In these simulations, during the second phase, CSB was trained with an ISI of 25 time steps, and CSA was still trained with an ISI of 50 time steps. As a result, the CSB started 25 time steps after CSA and lasted half as long. In this case, there is still full blocking of CSB with the CSC

A Identical Timing



B Blocked CSB Later



C Blocked CSB Earlier

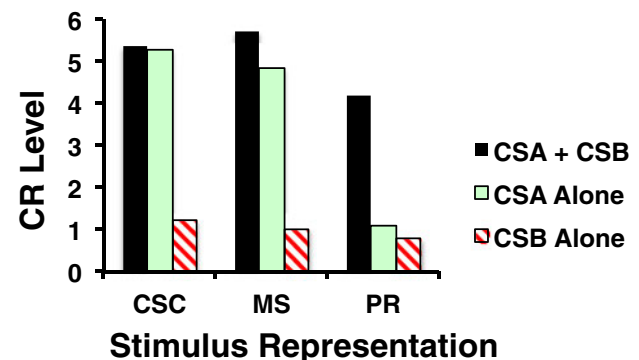


Fig. 5 Blocking in the TD model with different stimulus representations. CSA is the pretrained “blocking” stimulus, and CSB is the “blocked” stimulus introduced in the later phase. **a** When the timing of the two stimuli is identical in both phases, blocking is perfect with all three stimulus representations, and there is no conditioned response (CR) to CSB alone. **b** When the blocked CSB starts later, there is still full blocking with the CSC and MS representations. For the presence representation, the later, shorter stimulus can serve as a better predictor of the US and, thus, steals some of the associative strength from the earlier stimulus. **c** When the blocked CSB starts earlier, all three representations show an attenuation of blocking of the CSB, but there is an additional decrease in response to the CSA for the PR and MS representations. CSC = complete serial compound; PR = presence; MS = microstimulus

and MS representations, but not with the presence representation. With the latter representation, there is some responding to the blocked CSB and a sharp decrease in responding to CSA, as compared with the condition with matched ISIs (Fig. 5a). This responding to CSB occurs because CSB is a shorter stimulus that is more proximal to the US. As a result, there is the same positive prediction error on US receipt but fewer negative prediction errors during the course of the compound stimulus. Over time, CSB gradually “steals” the associative strength of the earlier CSA.

Empirically, when the onset of the added CSB occurs later than the onset of the pretrained CSA (as in Fig. 5b), acquisition of responding to the added CSB is largely blocked in eyeblink, fear, and appetitive conditioning (Amundson & Miller, 2008, Experiment 1; Jennings & Kirkpatrick, 2006, Experiment 2), as predicted by the TD model with the CSC or MS representations. When measured, responding to the pretrained CSA has shown mixed results. In rabbit eyeblink conditioning, responding to CSA has remained high (Kehoe et al., 1981), consistent with the representations that presume limited temporal generalization (MS and CSC). In appetitive conditioning in rats, however, CSA has suffered some loss of responding (Jennings & Kirkpatrick, 2006, Experiment 2), more consistent with a greater degree of temporal generalization (presence).

Finally, Fig. 5c depicts the results of simulations when the onset of the added CSB precedes the pretrained CSA during compound training. In this simulation, the CSA was always trained with an ISI of 25 time steps, and the CSB was trained with an ISI of 50 time steps. In this case, blocking was attenuated in the model with all three stimulus representations. For the TD model, this attenuation derives from second-order conditioning (#11.2): CSB, with its earlier onset, comes to predict the onset of the blocking stimulus CSA. In TD learning, the change in the US prediction at the onset of CSA (see Eq. 3) produces a prediction error that changes the weights for the preceding CSB. For the presence and MS representations, this second-order conditioning has an additional consequence. Because the elements of the stimulus representation for the added CSB overlap with those of CSA, the response to CSA diminishes (significantly more so for the presence representation).

In the empirical data, when the onset of the added CSB occurred earlier than the pretrained CSA, responding to the added CSB showed little evidence of blocking (Amundson & Miller, 2008, Experiment 2; Cole & McNally, 2007; Jennings & Kirkpatrick, 2006, Experiment 2; Kehoe et al., 1987). Independently of stimulus representation, the TD model correctly predicts attenuated blocking of responding to CSB in this situation (Fig. 5c), but not an outright absence of blocking. Responding to the pretrained CSA, in contrast, showed progressive declines after CSB was added in eyeblink and fear conditioning (Cole & McNally, 2007; Kehoe et

al., 1987), consistent with the presence representation, but not in appetitive conditioning (Jennings & Kirkpatrick, 2006), more consistent with the CSC and MS representations.

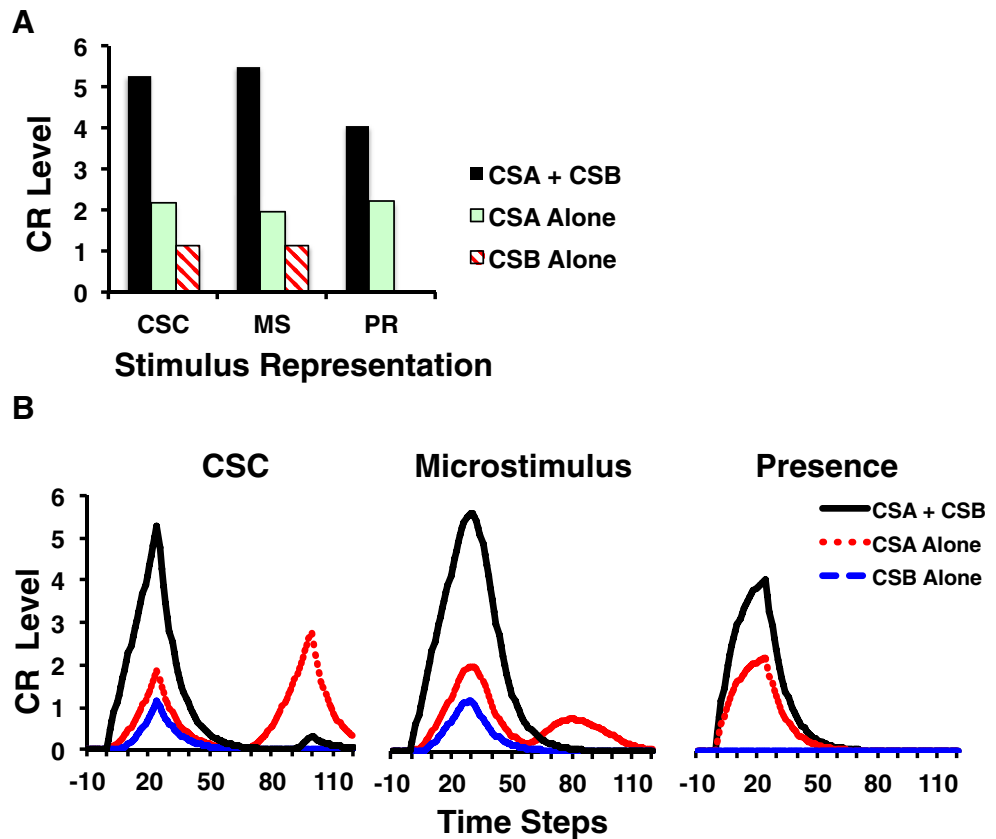
In these simulations of blocking, the presence representation is the most easily distinguishable from the other two representations, which presuppose less-than-complete temporal generalization. Most notably, with a presence representation, the TD model predicts that responding will strongly diminish to the pretrained stimulus (CSA) when the onset of the added stimulus (CSB) occurs either later or earlier than CSA. In contrast, the CSC and MS predict that responding to the pretrained CSA will remain at a high level.

We further consider one additional variation on blocking, where the ISI for the blocking stimulus CSA changes between the elemental and compound conditioning phases of the blocking experiment (e.g., Exp. 4 in Amundson & Miller, 2008; Schreurs & Westbrook, 1982). Empirically, in these situations, blocking is attenuated with the change in ISI. Once again, in these simulations, the blocking CSA was first paired with the US for 200 trials, but with an ISI of 100 time steps. Compound training also proceeded for 200 trials, and both stimuli had an ISI of 25 time steps during this phase.

Figure 6 shows the responding of the TD model to the different CSs on unreinforced probe trials presented at the end of training in this modified blocking procedure. With both the CSC and MS representations, but not with the presence representation, blocking is attenuated, and the blocked CSB elicits responding when presented alone, as in the empirical data. For these two representations, during compound conditioning, the US occurs earlier than expected, producing a large positive prediction error and driving learning to the blocked CSB. A surprising result emerges when the time course of responding is examined on these different probe trials (Fig. 6b). For the CSB alone and the compound stimulus (CSA + CSB), responding peaks around the time when the US would have occurred with the short ISI from the second phase. For the CSA alone, however, there is a secondary peak that corresponds to when the US would have occurred with the long ISI from the first phase. This secondary peak is restricted to the CSA-alone trials because the later temporal elements from CSB pick up negative weights to counteract the positive US prediction from CSA, effectively acting as a conditioned inhibitor in the latter portion of the compound trial. To our knowledge, this model prediction has not yet been tested empirically.

A second cue competition scenario that we simulated is overshadowing (#7.5)—often observed when two stimuli are conditioned in compound. In the overshadowing simulations below, the overshadowed CSB always had an ISI of 25 time steps. We included four overshadowing conditions in these simulations, where the overshadowing CSA had

Fig. 6 Blocking with a change in interstimulus interval (ISI). CSA is the pretrained “blocking” stimulus, and CSB is the “blocked” stimulus introduced in the later phase. **a** Performance on probe trials at the end of the blocking phase. Blocking was attenuated with the change in ISI for the CSC and MS representations, as indicated by the conditioned response (CR) level to CSB alone. **b** The time course of responding to CSB and the combined stimulus (CSA + CSB) shows a single peak at the time the US would ordinarily have been presented in the second phase (25 time steps). The time course for CSA alone shows a secondary peak later in the trial for the two representations that allow for temporal differentiation (CSC and MS). CSC = complete serial compound; PR = presence; MS = microstimulus



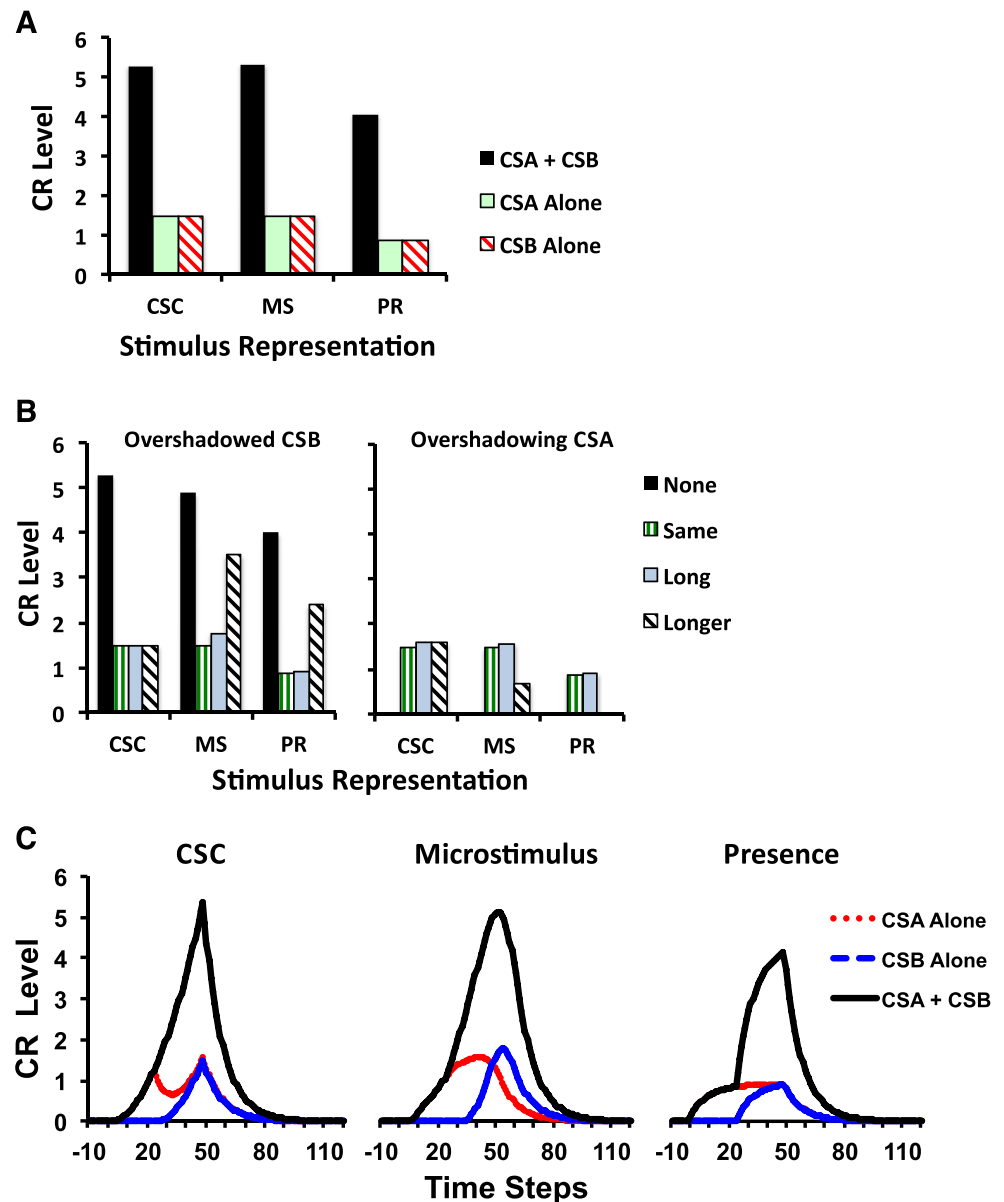
ISIs of 25 time steps (same), 50 time steps (long), or 100 time steps (longer) or was omitted altogether (none). Training proceeded for 200 trials, and 3 probe trials were included at the end of training: CSA alone, CSB alone, and a compound stimulus (CSA + CSB together).

Figure 7 plots responding in the TD model with different representations on these overshadowing simulations. When the timing of the two CSs are equated, there is a significant decrement in the level of maximal responding to each individual CS, as compared with the compound CS (Fig. 7a) or as compared with an individual CS trained alone (the none condition in Fig. 7b). When the timing of the overshadowing CSA is varied, so that the CSA now starts 25 time steps before the onset of CSB but still coterminates with CSB at US onset, there is a near-equivalent amount of overshadowing of CSB, independent of the stimulus representation in the model (left panel in Fig. 7b; see Jennings, Bonardi, & Kirkpatrick, 2007). If the overshadowing CSA is made even longer, we see a divergence in predicted degree of responding to the overshadowed CSB. With the CSC representation, the overshadowing is exactly equivalent no matter the length of the CSA. This equivalence arises because there is always the same number of representational elements from CSA that overlap and, thus, compete with the representational elements from CSB. With the presence and MS representations, a longer CSA produces less overshadowing. In these cases, the representational elements from CSA that

overlap with CSB are so broad that they support a lower level of conditioning by themselves and are thus less able to compete with CSB. The empirical data from appetitive conditioning show little change in overshadowing due to the duration of the CSA, but only a limited range of relative durations (2:1 and 3:1; Jennings et al., 2007). Thus, it remains somewhat of an open empirical question as to whether very long CSA durations would lead to reduced overshadowing, as predicted by both the MS and presence representations with the TD model.

The different representations also produce different predicted time courses of responding to the overshadowing and overshadowed stimuli. Figure 7c shows the CR time course during the probe trials after overshadowing with asynchronous stimuli (the long condition in Fig. 7b). For the CSB alone and the compound stimulus, the response curves are quite similar with the different representations (modulo the quirks in the shape of the timing function highlighted in Fig. 4), with a gradually increasing response that peaks around the time of the US. With the presence and MS representations, there is a small kink upward in responding when the second stimulus (CSB) turns on during the compound trials (compare Fig. 2 in Jennings et al., 2007), because the CSB provides a stronger prediction about the upcoming US. For the CSA-alone trials, however, the CR time courses are different for each of the representations. The CSC representation predicts a two-peaked response, the MS

Fig. 7 Overshadowing with the TD model. **a** Regular overshadowing. When both CSs start and end at the same time, there is a reduced conditioned response (CR) level to both individual stimuli with all three stimulus representations. **b** Overshadowing with asynchronous stimuli. When the CSA is twice the duration of the CSB (long), there is comparable overshadowing to the synchronous (same) condition. When the CSA is four times the duration of the CSB (longer), overshadowing is sharply reduced for the presence and MS representations, but not for the CSC. **c** Time course of responding during asynchronous overshadowing. Both the MS and CSC representations predict a leftward shift in the time course of responding to the overshadowing CSA, as opposed to CSB, and the time of US presentation. CSC = complete serial compound; PR = presence; MS = microstimulus



representation predicts a slight leftward shift in the time of maximal responding relative to the US time, and the presence representation predicts flat growth until CS termination. The empirical data seem to rule out the time course predicted by a CSC representation but are not clear in distinguishing the latter two possibilities (Jennings et al., 2007).

This overshadowing simulation also captures part of the information effect in compound conditioning (#7.11; Egger & Miller, 1962); with a fully informative and earlier CSA present, responding to CSB is sharply reduced (Fig. 6b). In addition, making CSA less informative by inserting CSA-alone trials during training reduces the degree to which responding to CSB is reduced (simulation not shown, but see Sutton & Barto, 1990).

Discussion

In this article, we have evaluated the TD model of classical conditioning on a range of conditioning tasks. We have examined how different temporal stimulus representations modulate the TD model's predictions, with a particular focus on those findings where stimulus timing or response timing are important. Across most tasks, the microstimulus representation provided a better correspondence with the empirical data than did the other two representations considered. With only a fixed set of parameters, this version of the TD model successfully simulated the following phenomena from the focus list for this special issue: 1.1, 7.2, 7.5, 7.11, 12.1, 12.4, 12.5, 12.6, 12.7, and 12.9.

A valuable feature of the TD model is that the learning algorithm has a normative grounding. There is a well-defined function that characterizes what the TD learning algorithm converges toward or what can be thought of as the goal of the computation. Equation 9 expresses more precisely how the TD model aims to generate predictions that are based on the *return*, which is the summed expectations of impending USs over time, and not just the US at the end of a trial (Sutton & Barto, 1998). The impending USs are discounted by their relative imminence; predicted USs are weighted so that imminent USs contribute more strongly to the prediction than do temporally distant USs in an exponentially discounted fashion (see Sutton & Barto, 1990). More formally, this target prediction of the future USs is the return (R_t):

$$R_t = \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}, \quad (9)$$

where r_t is the US intensity at time step t and γ is the discount factor (between 0 and 1), as in Eq. 3. The return from time step t (R_t) is thus the target for the US prediction using the features available at that time ($V_t(\mathbf{x}_t)$). In the TD model, this target prediction is *what* the animal is trying to learn about the world, and the TD learning algorithm is the proposed mechanism for *how* the animal does so. The animal's goal is thereby construed as making real-time US predictions that are as close as possible to the time course of the target prediction above.

Although not a causal explanation (this lies in the mechanism described in Eqs. 1–8), such a teleological interpretation can be very helpful in understanding the functioning of the proximal learning mechanism. For example, let us return to the question of why the US prediction takes a given time course with the MS representation (Fig. 2c). In this case, a teleological interpretation is that the TD learning algorithm is trying to best approximate the target US prediction, which is an exponentially weighted average of future USs. This target US prediction is exactly recreated by the US prediction curve for the complete serial compound (Fig. 2a), which can approach the target curve without constraints. In contrast, the TD model with MSs will find the best linear weighting of the MSs that approximates this target curve.

A notable feature of the MS TD model is that good timing results emerge with only a small handful of deterministic representational elements per stimulus (e.g., Fig. 4). The MS TD model exhibits proportional timing (peaks at the right time), a graded response curve, an increase in the width of the response curve with ISI (although subproportional), well-aligned peak times from early in conditioning, and inhibition of delay. Because the TD learning algorithm finds the best linear combination of representation elements to approximate the target US prediction curve, there is no need to fully cover the space with basis functions that each

have unique maximal response times. This sparse coverage approach differs from most other timing models in the spectral timing family (e.g., Grossberg & Schmajuk, 1989; Machado, 1997; but see Buhusi & Schmajuk, 1999), improves upon earlier versions of the MS TD model that used many more MSs per stimulus (Ludvig et al., 2008; Ludvig et al., 2009), and stills one of the main criticisms of this class of models, that they suffer from the “infinite of the possible” (Gallistel & King, 2009). In addition, as further support for such an approach to learning and timing, a spectrum of MS-like traces have recently been found during temporally structured tasks in the basal ganglia (Jin, Fujii, & Graybiel, 2009) and hippocampus (MacDonald, Lepage, Eden, & Eichenbaum, 2011).

The presence representation, however, did produce a better correspondence with some aspects of the data from asynchronous blocking (Fig. 5). For example, the decrease in responding to the blocking CSA when preceded by the blocked CSB is predicted only by the presence representation (Fig. 5c). In previous work, we have examined a hybrid representation that uses both an MS spectrum and a presence bit to model some differences between trace and delay conditioning (Kehoe, Olsen, Ludvig, & Sutton, 2009b; Ludvig et al., 2008). Such a hybrid representation can have the advantages of both constituent representations, but the interaction between the representations quickly gets complicated even in simple situations, perhaps limiting its explanatory value (see Ludvig et al., 2009).

In these simulations, we have necessarily focused on those findings that both have not been shown before for the TD model and particularly distinguish the TD model from the RW model. Many other results, however, have been previously demonstrated for the TD model or follow trivially from the RW model. For example, Sutton and Barto (1990) demonstrated second-order conditioning (#11.2; see their Fig. 23), extinction (#2.1 and #9.1), and conditioned inhibition (#5.1) in the TD model. In addition, by constraining the US prediction to be nonnegative, they also simulated the failure of conditioned inhibition to extinguish (#5.3; see their Fig. 19; see also Ludvig et al., 2008; Ludvig et al., 2009). Their simulations used a presence representation, but those results hold equally well for the other two representations considered here. Other phenomena follow straightforwardly from the similarity of the TD learning rule to the RW learning rule when questions of timing are removed (see Ludvig et al., 2011, for some discussion of this point): Overexpectation (#7.8), unblocking by increasing the US (#7.3), and superconditioning (#7.9) are all predicted by any TD model.

Other extensions to the TD model have been proposed that expand the reach of the model to other conditioning phenomena that we have not considered here. For example, Ludvig and Koop (2008) proposed a scheme for learning a predictive representation with a TD model that allows for

generalization (#3.1) between situations on the basis of their anticipated future outcomes. With this representation, they showed how a TD model could exhibit sensory preconditioning (#11.1 and #11.4), mediated conditioning (#11.5), and acquired equivalence (e.g., Honey & Hall, 1989). Pan, Schmidt, Wickens, and Hyland (2008) proposed a different extension to the TD model, which supposed separate excitatory and inhibitory weights that decayed at different rates. They showed that this formulation produced both spontaneous recovery (#10.6) and rapid reacquisition (#9.2).

No model is perfect, and the TD model is no exception. Several of the major classes of phenomena under consideration in this special issue lie beyond the explanatory power of current TD models, including most of the phenomena grouped under discrimination (Group 4), preexposure (Group 8), and recovery (Group 10). Future research will hopefully provide new angles for integrating these results with the TD model. These extensions will require new formalisms that may attach additional components to the TD model, such as memory- or model-based learning (Daw, Niv, & Dayan, 2005; Ludvig, Miran, Sutton, & Kehoe, 2012; Sutton, 1990), step-size adaptation (Pearce & Hall, 1980; Sutton, 1992), or additional configural representational elements (Pearce, 1987, 1994). These new developments will likely feature prominently in the next generation of computational models of animal learning.

Appendix

Simulation details

The following parameters were used in all simulations:

Learning Rule:

Discount factor (γ) = .97

Eligibility trace decay rate (λ) = .95

Step size (α) = .05

Response Model:

Response threshold (θ) = .25

Response decay (ν) = .9

Stimulus Representations:

Memory decay constant (d) = .985

Number of microstimuli (m) = 6

Width of microstimuli (σ) = .08

Salience of presence element (x) = .2

Other:

US was always magnitude 1 and lasted a single time step.

Trial duration was always 300 time steps.

Author's Note Preparation of this manuscript was supported by Alberta Innovates–Technology Futures and the National Science and Engineering Research Council of Canada.

References

- Amundson, J. C., & Miller, R. R. (2008). CS–US temporal relations in blocking. *Learning & Behavior*, 36, 92–103.
- Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, 19, 10502–10511.
- Buhusi, C. V., & Schmajuk, N. A. (1999). Timing in simple conditioning and occasion setting: A neural network approach. *Behavioural Processes*, 45, 33–57.
- Church, R. M., & Kirkpatrick, K. (2001). Theories of conditioning and timing. In R. R. Mowrer & S. B. Klein (Eds.), *Contemporary learning: Theory and applications* (pp. 211–253). Hillsdale, NJ: Erlbaum.
- Cole, S., & McNally, G. P. (2007). Temporal-difference prediction errors and Pavlovian fear conditioning: Role of NMDA and opioid receptors. *Behavioral Neuroscience*, 121, 1043–1052.
- Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, 18, 1637–1677.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8, 1704–1711.
- Desmond, J. E., & Moore, J. W. (1988). Adaptive timing in neural networks: The conditioned response. *Biological Cybernetics*, 58, 405–415.
- Drew, M. R., Zupan, B., Cooke, A., Couvillon, P. A., & Balsam, P. D. (2005). Temporal control of conditioned responding in goldfish. *Journal of Experimental Psychology Animal Behavior Processes*, 31, 31–39.
- Egger, M. D., & 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.
- Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain*. Medford, MA: Wiley-Blackwell.
- Grossberg, S., & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, 2, 79–102.
- Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, 15, 338–346.
- Jennings, D. J., Bonardi, C., & Kirkpatrick, K. (2007). Overshadowing and stimulus duration. *Journal of Experimental Psychology: Animal Behavior Processes*, 33, 464–475.
- Jennings, D. J., & Kirkpatrick, K. (2006). Interval duration effects on blocking in appetitive conditioning. *Behavioural Processes*, 71, 318–329.
- Jin, D. Z., Fujii, N., & Graybiel, A. M. (2009). Neural representation of time in cortico-basal ganglia circuits. *Proceedings of the National Academy of Sciences*, 106, 19156–19161.
- Kehoe, E. J., Ludvig, E. A., Dudeney, J. E., Neufeld, J., & Sutton, R. S. (2008). Magnitude and timing of nictitating membrane movements during classical conditioning of the rabbit (*Oryctolagus cuniculus*). *Behavioral Neuroscience*, 122, 471–476.

- Kehoe, E. J., Ludvig, E. A., & Sutton, R. S. (2009a). Magnitude and timing of CRs in delay and trace classical conditioning of the nictitating membrane response of the rabbit (*Oryctolagus cuniculus*). *Behavioral Neuroscience*, 123, 1095–1101.
- Kehoe, E. J., Olsen, K. N., Ludvig, E. A., & Sutton, R. S. (2009b). Scalar timing varies with response magnitude in classical conditioning of the nictitating membrane response of the rabbit (*Oryctolagus cuniculus*). *Behavioral Neuroscience*, 123, 212–217.
- Kehoe, E. J., Schreurs, B. G., & Amodei, N. (1981). Blocking acquisition of the rabbit's nictitating membrane response to serial conditioned stimuli. *Learning and Motivation*, 12, 92–108.
- Kehoe, E. J., Schreurs, B. G., & Graham, P. (1987). Temporal primacy overrides prior training in serial compound conditioning of the rabbit's nictitating membrane response. *Animal Learning & Behavior*, 15, 455–464.
- Kehoe, E. J., & White, N. E. (2002). Extinction revisited: Similarities between extinction and reductions in US intensity in classical conditioning of the rabbit's nictitating membrane response. *Animal Learning & Behavior*, 30, 96–111.
- Ludvig, E. A., Bellemare, M. G., & Pearson, K. G. (2011). A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives. In E. Alonso & E. Mondragon (Eds.), *Computational neuroscience for advancing artificial intelligence: Models, methods and applications* (pp. 111–144). Hershey, PA: IGI Global.
- Ludvig, E. A., & Koop, A. (2008). Learning to generalize through predictive representations: A computational model of mediated conditioning. In *From Animals to Animats 10: Proceedings of Simulation of Adaptive Behavior (SAB-08)*, 342–351.
- Ludvig, E. A., Mirian, M. S., Sutton, R. S., & Kehoe, E. J. (2012). *Associative learning from replayed experience*. Manuscript submitted for publication.
- Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, 20, 3034–3054.
- Ludvig, E. A., Sutton, R. S., Verbeek, E. L., & Kehoe, E. J. (2009). A computational model of hippocampal function in trace conditioning. *Advances in Neural Information Processing Systems (NIPS-08)*, 21, 993–1000.
- MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontinuous events. *Neuron*, 71, 737–749.
- Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological Review*, 104, 241–265.
- Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, 9, 343–364.
- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16, 1936–1947.
- Moore, J. W., & Choi, J. S. (1997). The TD model of classical conditioning: Response topography and brain implementation. In J. W. Donahoe & V. P. Dorsel (Eds.), *Neural-network models of cognition, biobehavioral foundations (Advances in Psychology* (pp. Vol. 121, pp. 387–405). Amsterdam: North-Holland/Elsevier.
- 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 inter-stimulus intervals. *Behavioral Brain Research*, 21, 143–154.
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53, 139–154.
- Pan, W. X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2008). Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *Journal of Neuroscience*, 28, 9619–9631.
- Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review*, 94, 61–73.
- Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, 101, 587–607.
- Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87, 532–552.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.
- Roberts, S. (1981). Isolation of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, 7, 242–268.
- Schreurs, B. G., & Westbrook, R. F. (1982). The effects of changes in the CS-US interval during compound conditioning upon an otherwise blocked element. *Quarterly Journal of Experimental Psychology*, 34B, 19–30.
- Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, 57, 87–115.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275, 1593–1599.
- Smith, M. C. (1968). CS-US interval and US intensity in classical conditioning of the rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology*, 66, 679–687.
- 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.
- Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91, 871–890.
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3, 9–44.
- Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *International Conference on Machine Learning (ICML)*, 7, 216–224.
- Sutton, R. S. (1992). Adapting bias by gradient descent: An incremental version of delta-bar-delta. *National Conference on Artificial Intelligence*, 10, 171–176.
- 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. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. W. Moore (Eds.), *Learning and computational neuroscience* (pp. 497–537). Cambridge, MA: MIT Press.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Vogel, E. H., Brandon, S. E., & Wagner, A. R. (2003). Stimulus representation in SOP: II. An application to inhibition of delay. *Behavioural Processes*, 62, 27–48.