Many theorists emphasize the role of an “internal model of the world” in directing intelligent behavior. Internal models predict the evolution of the environment by imitating its causal flow. They compute prediction signals and use these prediction signals to form novel associative chains. Formation of such predictive chains may contribute to reasoning and planning. Animals seem to learn and use internal models; they learn to anticipate predictable events, and their behavior in latent learning experiments reflects formation of novel associative chains. Despite such behavioral evidence, the neural basis of an internal model is still under debate. In order to investigate possible neural correlates of an internal model, a previous model of animal learning was extended to an internal model approach. As believed to occur in animals, the proposed neural network model computes predictions and uses these predictions to form novel associative chains in a latent learning experiment. Simulated signals resembled anticipatory neural activity which has been reported in cortex, striatum, and midbrain dopamine neurons. Simulated prediction signals were comparable to tonic anticipatory activities in cortex and striatum. Furthermore, simulated reward prediction error signals were comparable to phasic activities of midbrain dopamine neurons. These findings suggest that tonic anticipatory activities can reflect prediction signals and that phasic anticipatory activities can reflect prediction error signals. Furthermore, comparison of the model architecture with biological neural networks suggests that chains of neurons with anticipatory activity underlie formation of novel associative chains. In conclusion, anticipatory activity seems to reflect the processing of an internal model.

Introduction
In a famous experiment by Pavlov (1927), a dog was trained with the ringing of a bell (stimulus) followed by food delivery (reinforcer). In the first trial, the animal salivated when food was presented. After several trials, salivation started when the bell was rung. This finding suggests that the salivation response following the bell ring reflects anticipation of food delivery. A large body of experimental evidence led to the hypothesis that Pavlovian learning is dependent upon the degree of unpredictability of the reinforcer (Rescorla & Wagner 1972; Dickinson 1980). According to this hypothesis, reinforcers become progressively less efficient for behavioral adaptation as their
predictability grows during the course of learning. The difference between the actual occurrence and the prediction of the reinforcer is usually referred to as the “error” in the reinforcer prediction. This concept has been employed in the temporal-difference model (TD model) of Pavlovian learning (Sutton & Barto 1990). The TD model uses reinforcement prediction errors for learning a reinforcer prediction signal. This prediction signal is comparable to anticipatory behavior. As animals seem to optimize the sum of reinforcement over time (Mackintosh 1974; Dickinson 1980), it is the goal of the TD model to compute a desired prediction signal which reflects the sum of future reinforcement. If the reinforcer is “food intake,” this desired prediction signal reflects the sum of available food in the future. After training of the TD model with a stimulus followed by a reinforcer is completed, the prediction error signal increases phasically when the stimulus is presented, and the prediction signal is tonically increased during the intratrial interval. Recent studies relate the TD model to neural information processing, as the reward prediction error of the TD model resembles dopamine neuron activity in situations with unpredicted rewards, fully predicted rewards, reward-predicting stimuli, and unexpectedly omitted rewards. The comparison between basal ganglia anatomy and the architecture of the TD model suggests that cortico-striatonigral pathways are involved in adaptation of dopamine neuron activities (Barto 1995; Houk et al. 1995; Montague et al. 1996; Schultz et al. 1997; Suri and Schultz 1999).

The TD model is limited to associations between stimuli and one reinforcer. However, studies in cognition (Craik 1943; Piaget 1954), brain theory (Arbib 1972) and motor control (Mel 1991; Wolpert et al. 1995) suggest that humans and animals associate events (stimuli, reinforcers or behavioral responses) with other events and use these associations to form novel associative chains. This view is also supported by latent learning experiments consisting of several experimental phases. Such experiments show that animals can profit from unreinforced exploration of a maze in order to find the shortest way to a reward in a later phase of the experiment (Mackintosh 1974; Dickinson 1980). The sensory preconditioning paradigm is a simple experiment which demonstrates latent learning and formation of novel associative chains. This paradigm is composed of three phases: In the first phase, a neutral stimulus A precedes a neutral stimulus B; in the second phase, stimulus B precedes a reinforcer; and in the third phase, stimulus A is presented alone. Animals show an acquired behavioral response to stimulus A in the third phase that resembles the response to the reinforcer (Mackintosh 1974; Dickinson 1980). The similarity between this conditioned response to stimulus A and the unconditioned response to the reinforcer suggests that animals anticipate the occurrence of the reinforcer. This conclusion implies that animals internally form the novel associative chain “stimulus A is followed by stimulus B and stimulus B is followed by the reinforcer.”

Since animals form novel associative chains, it was suggested that animals can learn and use an “internal model” of their environment (Craik 1943; Arbib 1972; Dennett 1978; Sutton & Barto 1981; Sutton & Pinette 1985; Mel 1991; Wolpert et al. 1995). The term “internal model” is popular in engineering sciences and denotes a set of equations that describes the temporal development of a real world process (Garcia et al. 1989). Internal model approaches compute event-specific prediction signals and use these predictions to simulate hypothetical future experience. Therefore, they reproduce the formation of novel associative chains in the sensory preconditioning paradigm (Sutton & Barto 1981).

Anticipatory neural activity may follow similar principles as anticipatory behavior. Since dopamine neuron activity reports in many experiments about reward prediction errors (Montague et al. 1996; Schultz et al. 1997; Suri & Schultz 1999; Schultz 1998), dopamine neuron activity may also reflect reward prediction errors in experiments testing formation of novel associative chains. Tonic dopamine concentration in target areas is influenced by phasic responses of dopamine neurons (Schultz 1998). To reproduce experimental findings about effects of dopamine antagonists on associative learning in Schizophrenics, dopamine concentration in the nucleus accumbens (part of striatum) was simulated with an internal model approach (Schmajuk et al. 1996; Gray et al.
This model implies that dopamine concentration is influenced by formation of novel associative chains in a sensory preconditioning paradigm. This model prediction has been confirmed by the finding that striatal dopamine concentration is increased in the third phase of the sensory preconditioning paradigm (Young et al. 1998). Evidence that an internal model approach reproduces dopamine neuron activity would be further strengthened, if neurons in cortex or striatum projecting to dopamine neurons code for reward prediction signals and event-specific prediction signals. Therefore, the current study investigates if anticipatory neural activity in striatum and cortex resembles prediction signals of an internal model.

Anticipatory activity is related to an upcoming event which is prerepresented as a result of a retrieval action of antedating events, in contrast to activity reflecting memorized physical features of a previously experienced event (Wagner 1978). Therefore, anticipatory activity precedes a future event irrespective of the physical features of the antedating events which make this future event predictable (fig. 1A). Phasic activity anticipating rewards was reported in midbrain dopamine neurons (Ljungberg et al. 1992; Schultz et al. 1993a, 1997; Mirenowicz & Schultz 1994). Tonic activity which anticipates stimuli, rewards or the animal’s own actions was termed “anticipatory,” “preparatory,” or “predictive” and has been reported in the striatum (Hikosaka et al. 1989; Alexander & Crutcher 1990ab; Apicella et al. 1992; Schultz & Romo 1992; Kermadi & Joseph 1995; Tremblay et al. 1998; Hollerman et al. 1998), supplementary motor area (Alexander & Crutcher 1990ab; Romo & Schultz 1992), prefrontal cortex (Watanabe 1996), orbitofrontal cortex (Tremblay & Schultz W., 1999), premotor cortex (Mauritz & Wise 1986), and primary motor cortex (Alexander & Crutcher 1990ab).

In order to investigate the relation between anticipatory neural activity and internal models, we propose an adaptive internal model approach that is based on TD models of animal learning (Sutton & Barto 1981, 1990; Sutton & Pinette 1985). We demonstrate that the model forms novel associative chains similar to animals in the sensory preconditioning paradigm. Then we compare simulated predictive signals with anticipatory neural activities.

**Description of Anticipatory Neural Activity**

**Phasic Anticipatory Activity.** Phasic anticipatory activity of about 100 msec duration was reported for midbrain dopamine neurons. These neurons are activated by unpredictable rewards, and they are also activated by a reward following a stimulus for the first time. After repeated presentations of the stimulus followed by the reward, the activation elicited by the reward decreases and entirely disappears after learning is completed. These neurons become activated instead by the conditioned stimulus (Ljungberg et al. 1992; Schultz et al. 1993a, 1997; Mirenowicz & Schultz 1994; Schultz 1998).

**Tonic Anticipatory Activity.** Tonic anticipatory activity was found in subsets of cortical and striatal neurons. Before learning, such neurons often respond to specific events. When this event becomes predictable in the course of learning, these responses become progressively preceded by anticipatory neural activity (Hikosaka et al. 1989; Tremblay et al. 1998). After learning, neural activity is typically increased above base line levels when the predictive event and when the predicted event are presented. During the interstimulus interval, tonic anticipatory activity usually increases gradually until the predicted event occurs or, less frequently, decreases to base line levels (Hikosaka et al. 1989; Alexander & Crutcher 1990ab; Apicella et al. 1992; Kermadi & Joseph 1995; Watanabe 1996). Tonic anticipatory activity usually starts increasing between one and several seconds before the predicted event (Hikosaka et al. 1989; Alexander & Crutcher 1990ab; Kermadi & Joseph 1995).

Tonic anticipatory activity was reported in delayed response tasks. Each correct trial of this task consists of an instruction stimulus, a delay period, a trigger stimulus, a movement, and a reward. After learning, activities of subsets of striatal neurons anticipate the instruction stimulus,
the trigger stimulus, the reward, and the movements of the animal (Apicella et al. 1992; Schultz & Romo 1992; Schultz et al. 1992; Tremblay et al. 1998; Hollerman et al. 1998). Apicella and collaborators (1992) report anticipatory neural activity in the monkey striatum after learning a delayed go-no go task. In this task, the animal had to remember the color of the instruction light to perform correctly when the trigger light was presented. From the 1173 studied striatal neurons, 615 showed some change in activity during task performance. The activity of 193 task-related neurons increased in advance of at least one task component, namely the instruction stimulus (16 neurons), the trigger stimulus (15 neurons), the animal’s movement (56 neurons), or the reward delivery (87 neurons) (fig. 1B). These neurons with anticipatory activity were found in dorsal and anterior parts of caudate and putamen and were slightly more frequent in the proximity of the internal capsule.

Tremblay and Schultz (Tremblay & Schultz 1999; Tremblay & Schultz, in preparation) trained monkeys in delayed response tasks in which each instruction stimulus indicated presentation of a specific reward (two liquids with different taste). Instruction stimulus A was followed by reward X, instruction stimulus B was followed by the same reward X, and instruction stimulus C was followed by reward Y. Neural activity was recorded in six-layered parts of orbitofrontal areas 11 and 14 and rostral area 13. These neurons showed three principal types of activation, namely responses to instructions (15% of 1095 tested neurons), responses following reward (8%), and sustained activations preceding reward (9%). The pre-reward activations began several seconds before the reward and subsided <1 sec after reward delivery, apparently reflecting the upcoming reward rather than preceding events. All three principal types of orbitofrontal activation discriminated between different liquid rewards.

Tonic anticipatory activities can be specific for anticipated stimuli (Hikosaka et al. 1989; Alexander & Crutcher 1990b; Apicella et al. 1992; Kermadi & Joseph 1995; Tremblay et al. 1998; Hollerman et al. 1998). Such stimulus-specific anticipatory neural activity was reported by Kermadi & Joseph (1995) who analyzed neural activity of 2100 neurons in the caudate nucleus. During the instruction phase of the task, a sequence of three visual targets was presented and the monkey was required to fixate on a central fixation point. In the behavioral phase, the monkey had to press three levers in the order indicated by the instruction phase. Six different sequences of the three targets were presented in the instruction phase. As these six sequences were always presented in the same order, the three target stimuli in each trial were completely predictable. From 125 instruction-related neurons, activity of 81 neurons preceded presented stimuli. Activity of 46 neurons anticipated the offset of the central fixation point, activity of 7 neurons anticipated the illumination of any target, 17 neurons anticipated the illumination of the first target, and 11 neurons anticipated the onset of specific targets. In a majority (35 neurons), the responses to specific targets were modulated by the rank of the target in the sequence or by complex relationships with other targets. Anticipatory activity started increasing about 1 second before stimulus onset and progressively increased until it reached the peak at the onset of the anticipated stimulus. A neuron with activity anticipating the specific target L in the specific sequence ULR is shown in fig. 1C.

Other Anticipatory Activity. In some studies anticipatory activity was reported which does not clearly fit in our distinction between phasic (about 100 msec duration) and tonic anticipatory activity (seconds). Activity of hippocampal place cells anticipates the future position of the rat by about 100 msec (Muller & Kubie 1989; Mehta et al. 1997). Each of the head-direction cells in the anterior thalamus has a characteristic anticipation time between 0-50 msec by which it anticipates the future head direction (Blair et al. 1997). Event-specific anticipatory activity can also depend on the future behavior of the animal. Such activity that anticipates the retinal consequences of intended eye movements by about 100 msec was reported in frontal eye fields (Goldberg & Bruce 1990; Umeno & Goldberg 1997), superior colliculus (Walker et al. 1995), and parietal cortex (Duhamel et al. 1992).
Fig. 1 (A) Illustration defining the expression “activity anticipating event X.” The three well-trained trials types “event A followed by event X”, “event B followed by event X”, and “event C followed by event Y” are assumed to be separated by sufficiently long intertrial intervals and presented randomly intermixed. Event A (left) and event B (middle), which both precede event X, increase the activity of the neuron. Event C, which precedes event Y, elicits no activity (right side). This last control trial shows that the anticipatory activity is specific for event X. Furthermore, this control trial indicates that the activity is not related to a common physical feature of the events A, B, and C, and therefore does not reflect memorization of these preceding events. The responses following events X and event Y are not shown as they are not relevant for the definition. Events A and B were termed “predictive” events and event X the “predicted” or “anticipated” event. (B) Population activity of expectation- and preparation-related striatal neurons (figure from Apicella et al. 1992). Top: activation of 16 neurons preceding instruction onset. The intertrial interval was 4-7 seconds. Middle: activation of 44 neurons preceding the trigger stimulus in go trials. Most, but not all, preinstruction activity depended on the go or no-go condition. The histogram is split, because the intervals between instruction and trigger varied from 2.5 to 3.5 sec. Neurons responding to the trigger stimulus, activated during movement, or activated before instruction or reward are excluded. Bottom: activation of 68 neurons preceding reward in no-go trials. The activation began to a modest extend already before trigger onset, gained increasingly in amplitude after trigger onset and reached its peak when the reward was delivered. In each display, histograms for each neuron normalized for trial number are added and the resulting sum is divided by the number of neurons. (C) Activity of this neuron in caudate anticipated and responded to stimulus L only if stimulus L was illuminated in the sequence ULR (figure from Kermadi and Joseph 1995). In the shown instruction phase, the targets were presented in the six different sequences LUR, RLU, URL, LRU, ULR, and RUL. During fixation of a middle fixation point (FP, line 1 of each subfigure), the three visual stimuli L (left target), U (upper target), and R (right target) were illuminated in predictable order (line 2 of each subplot). Each cell discharge is indicated by a dot and the 6-9 successive trials per neuron are shown on successive lines (middle of each subplot). On each of these line, two large vertical bars indicate onset and offset of the fixation point. The time histogram was computed from the sum of the individual discharges (bottom of each subplot).
Description of the Adaptive Internal Model Approach

The proposed model is closely related to the TD model (Sutton & Barto 1990) and to an internal model approach (Sutton & Pinette 1985). The model consists of a component computing the time dependent prediction vector $p(t)$ (solid arrows, fig. 2A) and a component that uses prediction errors for gradually adapting the weight matrix $V$ (broken arrows, fig. 2A). The time dependent input vector $u(t)$ reports the presence or absence of events which can be stimuli or reinforcers. Predictions and prediction errors are computed as in the TD model (Sutton & Barto 1990), but rather than computing predictions only for reinforcement, predictions are computed for all presented events $u(t)$. The algorithm is designed to learn prediction signals for each presented event. The optimally learned “desired prediction signals” for an event leads to prediction error signals of value zero for all time steps. These desired prediction signals increase gradually until the onset of the predicted event according to a rate determined by the discount factor $\gamma$ and then decrease to zero during event presentation. The prediction signals with this time course correspond to the discounted sum of future event durations, if events occurring in the far future are less weighted than events in the near future. The expected future duration of an event can formally be expressed as the event signal $u(t)$ summed over all future time steps, as the event signal is assumed to be one when the event is present and zero when it is absent. The discount factor $\gamma$ is used in this sum to successively decrease the weight of an event the farther in the future it is predicted to occur (Appendix, eq. 1). The value of the discount factor $\gamma$ was estimated from the time course of measured anticipatory neural activity. We used usually the standard value $\gamma = 0.99$ per time step (1 time step = 100 msec), which led to an increase in the prediction signal of 1% each 100 msec (see Results). Previously, $\gamma = 0.98$ per 100 msec had been estimated from dopamine neuron activity (Suri and Schultz, 1999).

The model is applied to situations in which current predictions depend not only on current events but also on earlier events. To represent events over time, each event is mapped to a fixed temporal pattern of signals which we refer to as “temporal representation of the event”. As the predictions for each time step depend exclusively and linearly on the temporal event representation for this time step, this temporal pattern has to be characteristic for each time step. Therefore, the representation of each event consists of a large number of signals which has been chosen in our model equal to the number of time steps in a trial (70 signals). These single components have been proposed to be phasic (Sutton & Barto 1990), sustained (Desmond & Moore 1988, 1991), or phasic immediately after the event onset and then progressively more sustained (Grossberg & Schmajuk 1989; Fiala et al. 1996; Suri & Schultz 1999). In some models the representation of an event depends on successive events (Domainey et al. 1995; Suri & Schultz 1999). We implemented the temporal event representation as a series of phasic signals $x(t)$ following event onset with varying delays (fig. 2B; Appendix, eq. 2). The adaptive weights in the square matrix $V$ represent the “long term memory” of the model (fig. 2A). This matrix is thought to compute from a current temporal event representation $x(t)$ the expected temporal event representation in the next time step. By repeated multiplications with matrix $V$, the expected event representation could theoretically be estimated for any future time. Using the internal feedback loop over the weight matrix $V$, event representations some time steps ahead are estimated and novel associative chains are formed. In addition, this loop takes the sum of the estimated future event representations to compute the prediction $p(t)$ (Sutton & Pinette 1985; Appendix, eq. 3).

The desired prediction signal increases before a predicted event according to the discount factor $\gamma$ and decreases during its presentation. Therefore, the prediction error $e(t)$ is computed from discounted temporal differences in the prediction (differencer D in fig. 2A) and from the event vector $u(t)$ (Appendix, eq. 4). In order to minimize these prediction errors, the elements of the weight matrix $V$ are incrementally adapted according to the product of the prediction error with
eligibility traces of the temporal event representation (Appendix, eq. 5). These traces are defined as slowly decaying versions of the representation components (Appendix, eq. 6). Such stimulus traces were originally introduced to explain learning for situations with a delay between the stimulus and the reinforcer, as they bridge the time interval between the predictive stimulus and the reinforcer (Hull, 1943). Although TD models with complete temporal event representation learn without representation traces (Montague et al., 1996), the proposed model uses traces to accelerate learning (Sutton & Barto, 1998).

Fig. 2 (A) Model schematic. The component drawn with solid arrows computes the prediction \( p(t) \), and the component drawn with dashed arrows computes the prediction error \( e(t) \) that is used to adapt the weights \( V \). The event signals \( u(t) \) are transformed to a richer temporal representation \( x(t) \) that codes each event as a series of phasic signals covering trial duration. Multiplication of the representation \( x(t) \) with the weight matrix \( V \) estimates the representation \( x(t) \) for the next time step (discounted with a factor \( \gamma < 1 \)). Repeated computations of the loop over the weight matrix \( V \) mimic the development of the environment and lead to an estimate of the prediction signals \( p(t) \). The prediction error signals \( e(t) \) report deviations to the desired prediction signals. This error is minimized by incrementally adapting the elements of the weight matrix \( V \) using the prediction error \( e(t) \). (B) Temporal event representation. The event signal \( u(t) \) reports presence or absence of events (line 1). The representation \( x(t) \) of an event consists of a series of phasic signals covering trial duration. The first component of the representation \( x(t) \) peaks with amplitude one (line 2), the second with amplitude \( \delta \) (line 3), the third with amplitude \( \delta^2 \) (line 4), and so on. Representation \( x(t) \) computed with the standard value \( \delta = 1 \) is shown (without decay of the representation).
Model Simulations

**Pretraining with events alone.** For the experimental situations, animals were typically familiar with the single events occurring in the subsequent experiments. Therefore, the model was pretrained with 20 presentations of the single events alone. Single events were presented during one second. Event presentations were separated by sufficiently long interstimulus intervals to prevent learning of associations between events.

**Sensory preconditioning.** Each learning phase of the sensory preconditioning paradigm, which typically consists of several trials for animal experiments, was simulated with one trial. Events were presented during one second and separated by interstimulus intervals of 5 sec. To compare the model performance with animal behavior, events A and B were referred to as “stimuli” and event X as “reinforcer.” (The model equations do not distinguish between stimuli and reinforcers.) The stimuli A and B and their prediction signals were thought to be behaviorally neutral. As in the original TD model for Pavlovian learning (Sutton and Barto, 1990), reinforcer X and prediction of reinforcer X were compared with the unconditioned behavioral response and with the conditioned anticipatory behavior, respectively.

**Delayed Response Task.** We did not intend to reproduce anticipatory neural activity for all events in the delayed response task but restricted the simulation to a first event A, compared with the instruction stimulus, followed after a delay by a second event B, compared with the reward. The duration of the instructions stimulus (1 sec) and the duration of the intratrial interval (5 sec) corresponded to similar durations in the monkey experiments. Trigger stimulus and the animal’s movements were not modelled. Intertrial intervals were long enough to avoid associations between trials. If only one reward was used in the animal experiment, the model was trained with two different events separated by a delay (stimulus -> reward). If three different instruction stimuli preceded delivery of two different rewards, the model was trained with the corresponding three pairs of events (stimulus A -> reward X, stimulus B -> reward X, and stimulus C -> reward Y). The model was trained with 20 trials for each pair. Tonic and phasic anticipatory activity was compared with prediction signals and prediction error signals, respectively. The temporal discount factor $\gamma$ was chosen to approximate the time course of the anticipatory neural activity.

The chosen event representation covered equally the whole interval between stimuli and rewards without “forgetting” the stimulus presentation. However, not all neurons may have access to such a complete temporal event representation. We therefore examined in a further simulation the influence of an incomplete event representation that decayed rapidly after presentation of events. This was achieved by setting the value of the decay rate $\delta$ to 0.8 per 100 msec. Using this parameter value, the peaks of the event representation components decreased 20% for each additional 100 msec stimulus-peak interval (see legend to fig. 2B). This model with incomplete event representation was trained according to the schedule with two rewards.
Fig. 3 (A) Presentations of a single event. For the first presentation of a novel event $u_1(t)$ of one second duration (left, line 1) the prediction signals $p_1(t)$, $p_2(t)$, and $p_3(t)$ (left, line 2-4) were equal to the components of the temporal event representation since the weight matrix $V$ was initialized with zeros (Appendix eq. 3). Only the first three predictions signals $p_1(t)$, $p_2(t)$, and $p_3(t)$ of the 70 computed prediction signals $p_1(t)$, ..., $p_{70}(t)$ are shown. The prediction error signals $e_1(t)$, $e_2(t)$, and $e_3(t)$ (left, line 5-7) were computed from the event $u_1(t)$ and from temporal changes in the event prediction signals $p_1(t)$, $p_2(t)$, and $p_3(t)$ (Appendix eq. 4). After 20 presentations of this event (right side), learning was completed and the duration of the event $u_1(t)$ (right, line 1) was correctly predicted. All 70 prediction signals were almost identical (all are shown superimposed) and decreased during the presentation of the event, as they reflected the remaining future event duration (right side, line 2). All 70 prediction error signals were almost identical (all are shown) and increased phasically at the event onset, as the time of event onset was unpredictable (line 3).

(B) Sensory preconditioning. In the first trial ($A\rightarrow B$), stimulus $A$ (line 1) preceded stimulus $B$ (line 2). Signals reflecting prediction of reinforcer $X$ (line 4) were not affected. In the second trial ($B\rightarrow X$), stimulus $B$ preceded reinforcer $X$ (line 3). Since the model had been pretrained with the events alone, presentation of reinforcer $X$ increased the signals reflecting prediction of $X$ (signal equal to fig. 3A, right, line 2). Signals reflecting prediction of reinforcer $X$ (line 4) were not affected. In the second trial ($B\rightarrow X$), stimulus $B$ preceded reinforcer $X$ (line 3). Since the model had been pretrained with the events alone, presentation of reinforcer $X$ increased the signals reflecting prediction of $X$ (signal equal to fig. 3A, right, line 2). Signals reflecting prediction of reinforcer $X$ (line 4) were not affected. In the second trial ($B\rightarrow X$), stimulus $B$ preceded reinforcer $X$ (line 3). Since the model had been pretrained with the events alone, presentation of reinforcer $X$ increased the signals reflecting prediction of $X$ (signal equal to fig. 3A, right, line 2). Signals reflecting prediction of reinforcer $X$ (line 4) were not affected. In the second trial ($B\rightarrow X$), stimulus $B$ preceded reinforcer $X$ (line 3).

(C) After 20 presentations of event $A$ (line 1) followed by event $B$ (line 2). The model was trained with the standard value of 0.99 for the discount factor $\gamma$ (left side) and with the value of 0.85 (right side). All 70 signals reflecting prediction of event $B$ are shown superimposed and are almost identical (line 3). The increases in the predictions signals are close to the desired rates of 1% per 100 msec (left side) and 15% per 100 msec (right side). All 70 signals reflecting errors in prediction of event $B$ are shown superimposed and are almost identical (bottom, left and right). For $\gamma = 0.99$ (left), the prediction error signals increased phasically at onset of event $A$, because event $A$ was unpredictable. For $\gamma = 0.85$ (right side), the prediction error signals were small, because the simulated predictions signals were close to the desired prediction signals for all time.
steps. (D) Learning curves for 30 pairings of event A with event B ($\gamma = 0.99$, training as in fig. 3C, left). 100 weights of matrix $V$ (out of 210 x 210 weights for 3 stimuli) were randomly selected. Most weights remained zero during learning. 19 weights changed their values during learning and usually converged in a few trials to a stable value. Simulations shown in (A), (B), (C, left) and (D) with standard value of discount factor $\gamma = 0.99$ per time step, 1 time step = 100 msec, and without representation decay as $\delta = 1$.

Results
The model was pretrained with presentation of all events alone. In this pretraining, the time of event onset was unpredictable but the event duration remained constant. After pretraining, the prediction signals correctly decreased during event presentation, and the prediction error signals increased phasically at the event onset (fig. 3A).

The pretrained model was tested with an analogue of the sensory preconditioning paradigm (Fig. 3B). In the first trial, the neutral stimulus A preceded the neutral stimulus B (A -> B). Both stimuli did not increase prediction of reinforcer X. In the second trial, stimulus B preceded reinforcer X (B -> X). Presentation of reinforcer X increased prediction of X, as the model had been pretrained with presentations of the reinforcer X. In the third trial, presentation of stimulus A alone increased prediction of reinforcer X, because the model had formed the novel associative chain “if stimulus A, then stimulus B, then reinforcer X” (A -> B -> X).

The model was trained with an event A followed after 5 seconds by an event B (Fig. 3C). When the discount factor $\gamma$ was set to the standard value of 0.99 per 100 msec (left), the prediction signals increased at onset of event A and then progressively increased with a rate similar to the desired rate of 1% per 100 msec (left, line 3). When the model was trained with $\gamma = 0.85$, the prediction signals increased with a rate similar to the desired rate of 15% per 100 msec (right, line 3). As the interstimulus interval was long enough to learn the desired prediction signals for $\gamma = 0.85$, but too short for $\gamma = 0.99$, the prediction error was phasically increased at onset of event A for $\gamma = 0.99$ (bottom, left side) but small for $\gamma = 0.85$ (bottom, right side).

To examine the values of the adaptive weights in the matrix $V$ during learning, 100 weights were randomly selected (fig. 3D). Most weights remained zero during training in 30 trials. 19 weights increased and usually approached a stable value.

Simulated signals of the proposed model were compared with anticipatory neural activity measured in putamen (part of striatum) and midbrain dopamine neurons (fig. 4). When the stimulus and the reward were presented in temporal succession for the first time (fig. 4A), the reward prediction signal was not affected by presentation of the stimulus as the stimulus was not associated with reward. At the time of the unpredicted reward, the reward prediction signal was increased (fig. 4A, line 3). Similar to this simulated signal, activity of a subset of putamen neurons was not affected by presentation of the stimulus and was increased by the reward (line 4). The simulated reward prediction error was phasically increased at the onset of the reward as the reward was unpredicted (line 5). This signal was comparable to the activity of midbrain dopamine neurons (bottom).

Simulated signals were then compared with neural activities after learning (fig. 4B). The simulated reward prediction signal was already increased at stimulus onset and increased gradually until reward onset (line 3) as the stimulus predicted the reward. This signal increased between the stimulus and the reward 1% for each 100 msec because the discount factor $\gamma$ was set to 0.99 per 100 msec (standard value). The simulated reward prediction signal was comparable to reward-anticipatory activity of a subset of striatal neurons (line 4). The signal representing the reward prediction error was phasically increased by the unpredicted stimulus but not affected by the predicted reward (line 5). This response to the conditioned stimulus was smaller than the response to the unpredicted reward before learning (fig. 4A, line 5) as the prediction signal (fig. 4B, line 3) increased gradually according to the discount factor. The reward prediction error was comparable to
dopamine neuron activity (fig. 4B, bottom). Taken together, fig. 4 shows that the simulated reward prediction resembles the activity of a subset of putamen neurons and that the simulated reward prediction error resembles activity of dopamine neurons.

**Fig. 4** Comparable time courses of predictive signals and activity histograms for a stimulus (line 1) preceding a reward (line 2). The same time scale applies to simulated trajectories and histograms of neural activity. Activity histograms were selected from delayed response tasks with the simulated stimulus corresponding to the instruction stimulus. Simulated signals were selected from fig. 3A (right) and fig. 3C (left). (A) Before learning. A typical reward prediction signal was increased when the reward was presented (line 3). This signal was comparable to the activity histogram of a set of putamen neurons aligned to an unpredictable reward (line 4) (Schultz et al. 1993b). A typical signal reflecting reward prediction errors was phasically increased at onset of the reward (line 5). This signal was comparable to dopamine neuron activity (bottom). These neurons do not respond to a small instruction light (Ljungberg et al. 1992) but rather to an unpredictable drop of liquid reward delivered to the mouth of the animal (Mirenowicz & Schultz 1994). (B) After learning (20 stimulus-reward pairings). A typical reward prediction signal increased already when the stimulus was presented and then increased gradually until occurrence of the reward (line 3). This reward prediction signal was comparable to anticipatory activity of a putamen neuron (line 4; from Apicella et al. 1992, fig. 15). This neural activity seems to anticipates the future reward, because it was also increased before the reward regardless of the instruction stimulus which indicated reward delivery. Out of 1173 studied neurons, 6 striatal neurons showed similar sustained reward-anticipatory activity lasting over the whole task duration (compare fig. 1B). A typical signal reflecting reward prediction errors was already phasically increased at the stimulus onset and on baseline level when the reward was presented (line 5). This signal was comparable to the activity of dopamine neurons which respond after learning to a small instruction light but not to a predictable drop of liquid reward (bottom) (Ljungberg et al. 1992).

Model performance was also comparable with reward-specific anticipatory activity recorded in orbitofrontal cortex (fig. 5). Monkeys had been trained in a delayed response task with three instruction stimuli A, B, and C followed by two different rewards X and Y. The model had been trained with corresponding pairs of events. In trials without reward Y, prediction of reward Y was not affected (fig. 5A, top, left and middle). This prediction signal was activated when stimulus C was presented and then gradually increased until reward Y was presented (fig. 5A, top, right), because reward Y was completely predicted by stimulus C. Prediction of reward Y was comparable to reward-specific anticipatory activity of a subset of orbitofrontal neurons anticipating reward Y but not reward X (fig. 5A, bottom).
When the model was trained with the value of 0.95 per 100 msec for the temporal discount factor $\gamma$, prediction signals increased more rapidly (fig. 5B, top). Therefore, prediction of reward X was only slightly active at the onset of stimuli A and B and then increased rapidly until reward X (top, left and middle), because reward X was completely predicted by the stimuli A and B. Prediction of reward X was comparable to the activity of a subset of orbitofrontal neurons anticipating reward X (fig. 5B, bottom). This finding indicates that variations in the discount factor $\gamma$ correspond to variations in the time course of anticipatory activity. Furthermore, comparison of fig. 5B with fig. 5A shows that simulated prediction signals and anticipatory neural activities discriminate between specific rewards.

We investigated the influence of a rapidly decaying event representation on the prediction signal. After 20 presentations of stimulus-reward pairs, prediction of reward X was activated by stimuli A and B and by reward X (fig. 5C, top). Since learning was not completed, prediction of reward X was associated with the correct predictive stimuli A and B but did not bridge the time between the stimuli and the reward. The responses to the predictive stimuli A and B were learned with the representation traces, because the traces still bridged the time gap between the stimuli and the rewards. Prediction of reward X was comparable to activity of a subset of orbitofrontal neurons anticipating reward X but not reward Y (fig. 5C, bottom). This finding suggests that neurons with anticipatory neural activity that is on base line levels during the intratrial interval do not have access to the complete temporal event representation.

![Fig. 5](image-url) Comparable time courses of prediction signals and orbitofrontal activities after learning pairings between different stimuli and rewards. The simulated stimulus is compared with the instruction stimulus in a delayed response task. (Histograms reconstructed from Tremblay & Schultz, in preparation; see section “Description of Anticipatory Neural Activity”. Durations of simulated stimuli and rewards as in fig. 4.) (A) Prediction of reward Y. In trials without reward Y, all signals reflecting prediction of reward Y were zero (top, left and middle). When stimulus C preceded reward Y, a typical prediction signal was activated when stimulus C was presented and then increased gradually until reward Y. Prediction of reward Y was comparable to the activity of an orbitofrontal neuron anticipating reward Y but not reward X (bottom). (In the histogram at bottom, right, neural activity before the task was larger than after the task, because the
previous task predicted already reward Y.) (B) Prediction of reward X with discount factor $\gamma = 0.95$ per 100 msec. A typical signal reflecting prediction of reward X slightly increased when stimuli A and B were presented and then increased rapidly until reward X (top, left and middle). This signal was zero in trials without reward X (top, right). The prediction of reward X was comparable to the activity of an orbitofrontal neuron anticipating reward X but not reward Y (bottom). The two neurons shown in (A) and (B) belong to a subset with sustained activations preceding specific rewards which included 9% of tested neurons (Tremblay & Schultz 1999). (C) Prediction of reward X with representation decay ($\delta = 0.8$ per 100 msec). In trials with presentations of reward X, a typical signal reflecting the prediction of reward X increased when the stimuli A and B or the reward X were presented (top, left and middle). Prediction of reward X did not bridge the intratrial interval, as the event representation decayed rapidly. In the trial without presentation of reward X, prediction of reward X was zero (right). Prediction of reward X was comparable to the activity of an orbitofrontal neuron. This neuron responded to the stimuli A and B and to the reward X (bottom, left and middle). The activity did not bridge the intratrial interval. Activity was at baseline levels in trials without reward X (bottom, right). This neuron belongs to a subset with reward-dependent responses to the instruction (15% of tested neurons) and to a subset with reward-dependent responses following the reward (8%) (Tremblay & Schultz 1999). (Activities were aligned to rewards. Horizontal broken lines below histograms indicate stimulus onsets.)

Discussion

The current study demonstrates that the proposed adaptive internal model approach reproduces formation of novel associative chains in an analogue of a sensory preconditioning experiment and reproduces anticipatory neural activity. The model learned tonic event-specific prediction signals using phasic event-specific prediction error signals.

The model was shown to form novel associative chains in an analogue of the sensory preconditioning paradigm (fig. 3D). The signal representing prediction of reinforcer X acquired a response to stimulus A in the third phase of this paradigm. This prediction signal, corresponding to the animal’s anticipatory behavior, reproduces the animal’s acquired response in the third phase of the sensory preconditioning paradigm (Mackintosh 1974; Dickinson 1980). As anticipatory neural activities resemble predictions signals, this suggests that novel associative chains may be formed by chains of neurons with tonic anticipatory neural activities.

Before and after learning the contingency between a stimulus and a reward, simulated signals representing errors in the prediction of rewards were comparable to typical time courses of phasic activities of midbrain dopamine neurons anticipating rewards (fig. 4). In addition, simulated reward-specific prediction signals were comparable to reward-specific anticipatory activity of subsets of cortical and striatal neurons (fig. 4 and fig. 5). Variation of two model parameters, the discount factor and the decay rate of temporal event representation, reproduced variations in time courses of anticipatory neural activity (fig. 5). Although the shown model simulations do not include learning between more than two event, they can be applied to more complex tasks. For the delayed response task, anticipatory neural activity in the striatum could be reproduced with separate models learning the pairs reward-instruction (fig. 1B, top), instruction-trigger (fig. 1B, middle), instruction-reward (fig. 1B, bottom), and trigger-reward (fig. 1B, bottom). For the sequence reproduction task (fig. 1C; Kermadi & Joseph, 1995), an elaborated and biologically plausible internal representation has been proposed which reproduces order- and stimulus-dependent neural activities (Dominey et al.). A model with an elaborated internal representation and TD learning rules may be able to reproduce these neural activities as their time course resembles the time course of predictions signals simulated in the present study (compare fig. 3C, right). Taken together, these findings suggest that phasic anticipatory activities can reflect prediction error signals and that tonic anticipatory activities can reflect prediction signals.

Associative weights involved in the computation of tonic prediction signals were adapted according to phasic signals reporting prediction errors. Therefore, the model suggests that phasic anticipatory activities induce long-term adaptations of neurons with tonic anticipatory activities.
Consistent with evidence for dopamine-dependent long-term adaptation of corticostriatal transmission (Calabresi et al. 1992, 1997; Wickens et al. 1996), the model suggests that activity of midbrain dopamine neurons leads to long-term adaptations of cortical or striatal neurons with tonic reward-anticipating activity. Furthermore, the model postulates the existence of a category of neurons that are phasically active, report errors in predictions of specific events, and induce long-term adaptations of neurons with tonic event-specific anticipatory activity. Although phasic context dependent neural activities in striatum and prefrontal cortex have been reported (see Schultz and Romo 1992), it has not been investigated if these activities anticipate events.

When the model was trained using an incomplete temporal event representation, the prediction signal did not increase gradually before the predicted event but instead decreased to zero in the intratrial interval. This time course was similar to some time courses of anticipatory activity (fig. 5C). If a series of distinguishable stimuli were presented during the trial, these stimuli would serve as a complete temporal event representation, and the model would learn the gradually increasing prediction signals. This suggests that anticipatory neural activity could reveal their optimal time course when measured in an experiment with a series of stimuli preceding the anticipated event.

The proposed model is a combination of two earlier TD algorithms which reproduce findings in animal learning research (Sutton & Pinette 1985; Sutton & Barto 1990). Prediction error, learning rule, temporal event representation, and representation traces were computed according to the TD model (Sutton & Barto 1990). The proposed model performs very similar to the TD model for tasks with one predicted event. In particular, the reward prediction error of the proposed model should reproduce dopamine neuron activity in the same situations as the TD model (Montague et al. 1996; Schultz et al. 1997). Unfortunately, both models fail to reproduce dopamine neuron activity when a reward is delivered earlier than expected (Hollerman and Schultz 1998, Suri and Schultz 1999). For situations with delayed or omitted rewards, prediction signals decrease for both models to zero at the time of the expected reward. These prediction signals resemble some, but not all, tonic anticipatory activity for delayed reward presentation (Hikosaka et al. 1989). These subtle inconsistencies between simulated signals and measured anticipatory activity suggest that some components of the temporal event representation are influenced by subsequent events (Suri and Schultz 1999).

In contrast to the proposed model, the TD model does not reproduce the finding that animals form novel associative chains in latent learning experiments. To reproduce these findings, Sutton & Pinette (1985) proposed an internal model approach which learned with TD methods event-specific prediction signals. We extended this approach with the temporal event representation of the TD model (Sutton & Barto 1990) in order to reproduce the time course of anticipatory neural activity.

The proposed model can be partially related to networks of biological neurons. The model learns reward predictions and stimulus predictions, and forms novel associative chains. It has been suggested that reward predictions are learned in limbic parts of pathways from cortex via striatum to midbrain dopamine neurons (Houk et al. 1995; Montague et al. 1996; Suri & Schultz 1999). Stimulus predictions may be learned predominantly in the cortex. For visual stimuli, it has been proposed that pyramidal neurons in higher cortical areas learn to predict neural activity of pyramidal neurons in lower cortical areas (Rao & Ballard, 1997; Rao & Ballard, 1999). According to this view, feedforward connections to higher cortical areas carry the prediction errors, whereas the feedback connections carry the prediction signals. Novel associative chains are formed by the proposed model in a positive feedback loop in which predicted events serve as representation for further event predictions. Formation of novel associative chains often requires cortical areas (Bunsey & Eichenbaum 1996; Balleine & Dickinson 1998) and has been simulated in a model of the hippocampus with reciprocal connections between hippocampal pyramidal cells within stratum radiatum of CA3 (Wallenstein, et al. 1998; personal communication with Wallenstein). Taken
together, stimulus predictions and novel associative chains may be computed in reciprocal projections between cortical areas.

The proposed model computes prediction signals but does not select the actions that optimize reinforcement. Two approaches could overcome this limitation. First, instead of learning sensorimotor associations with the reinforcement given by the environment, dopamine-like reinforcement prediction errors can serve as an internal predictive reinforcement. Simulation studies suggest that sensorimotor learning with a dopamine-like predictive reinforcement signal avoids perseveration errors (Suri & Schultz 1999) and is advantageous in situations with delayed reinforcement (Sutton & Barto 1998; Suri & Schultz 1998). Second, predicted chains of stimuli, reinforcers and actions computed by internal models can also be used to select the actions that predict the optimal outcome (Sutton & Barto 1981; Suri, Bargas, & Arbib, in preparation; Suri, Marmol, & Arbib, http://latte.usc.edu/~bmw, in preparation). Such action choices on the basis of novel associative chains were suggested to underlie cognitive processes (Craik 1943). This approach using nonadaptive internal models is termed Internal Model Control or Model Predictive Control and has been used for control of real world processes (reviewed by Garcia et al. 1989). These successful applications in engineering sciences demonstrate the computational advantages of internal models.

We thank Leon Tremblay for the permission to use neural activities he recorded in orbitofrontal cortex (Tremblay & Schultz in prep.). This study was supported by the James S. McDonnell Foundation grant 94-39.

References


Appendix A: Model Equations

The event vector $u(t)$ reports the presence ($u_k(t) = 1$) and absence ($u_k(t) = 0$) of events. The desired prediction signal $p_d(t)$ is defined as the discounted sum of future events $u(t)$

$$p_d(t) = u(t) + \gamma u(t+1) + \gamma^2 u(t+2) + ...$$  \hspace{1cm} (eq. 1)

(standard value of discount factor $\gamma = 0.99$ per time step, 1 time step = 100 msec). The proposed model consists of the equations below (eq. 2 to eq. 6) which allow the estimation of the desired prediction signal $p_d(t)$.

The dimension of the event vector $u(t)$ is increased in order to construct vectors of equal dimension in the whole model. The same signal reporting presence or absence of an event (fig. 1C, top) is copied $N$ times in the event vector. Therefore, the $N$ identical signals $u_k(t)$,  

$$k = (l - 1)N + 1,(l - 1)N + 2, ..., (l - 1)N + N$$

report the presence or absence of the event with number $l$. According to this definition, $u(t)$ and $x(t)$ are both time dependent vectors of dimension “$N$ times number of events”. (For example, if processing of three events is modelled, only the second event is present at time $t$, and assumed that $N = 2$, then $u(t) = [0,0,1,1,0,0]$).

Events $u(t)$ are represented with the temporal event representation $x(t)$ as it is shown in fig. 2B. Event number 1 was represented in the event representation components $x_1(t)$, $x_2(t)$, ..., $x_N(t)$. Event number 2 in the event representation components $x_{N+1}(t)$, $x_{N+2}(t)$, ..., $x_{2N}(t)$, and so on. (In the above example with $N = 2$, $\delta = 1$, and $u(t) = [0,0,1,1,0,0]$, it is $x(t) = [0,0,1,0,0,0]$.
\[ x(t+1) = [0,0,0,1,0,0]. \] The function which computes the temporal event representation \( x(t) \) from the events \( u(t) \) is referred to as “transformation \( T \)”. This is written as

\[ x = T[u] \quad \text{(eq. 2)} \]

In order to cover trial durations of 7 seconds with the event representation \( x(t) \), each event is represented with 70 phasic representation components (\( N = 70 \)).

A basic concept of the internal model approach is to estimate a weight matrix \( V \) (size 280 x 280 for four events) that computes the temporal representation for the next time step (\( Vx(t) \equiv x(t+1) \)) (Sutton & Pinette 1985). With this weight matrix \( V \), the desired prediction signal defined in eq. 1 could be estimated with \( p(t+1) = T^{-1}(x(t+1) + Vx(t+1) + V^2x(t+1) + ...) \) (\( T^{-1} \) denotes the inverse of the transformation \( T \) and computes the event vector from a given temporal representation.) Instead of implementing this equation, it was simplified using two assumptions. First, terms of third and higher order in \( V \) were omitted as their values should decrease with the corresponding powers in \( \gamma \). Second, the inverse transformation \( T^{-1} \) was omitted to avoid the difficult matrix inversion. Therefore, the prediction signal is estimated with

\[ p(t+1) = x(t+1) + Vx(t+1) + V^2x(t+1) \quad \text{(eq. 3)} \]

This expansion in matrix \( V \) is shown as feedback loop over matrix \( V \) in fig. 2A. Additional higher order terms in eq. 3 did not affect the results of this paper, but the model failed to form the novel associative chain in the sensory preconditioning experiment when the term \( V^2x(t+1) \) was omitted (results not shown).

As eq. 1 is equivalent to \( p_d(t) = u(t) + \gamma p_d(t+1) \), the error \( e(t) \) between the estimated prediction \( p(t) \) and the desired prediction \( p_d(t) \) can be computed from discounted temporal differences between successive predictions (differencer D in fig. 2A) and from the event vector \( u(t) \) with

\[ e(t) = u(t) + \gamma p(t+1) - p(t) \quad \text{(eq. 4)} \]

The weight matrix \( V \) is initiated with zeros and then adapted with the two factor learning rule

\[ V = V + \beta e(t) \cdot \bar{x}^T(t), \quad \text{(eq. 5)} \]

where \( \bar{x}^T(t) \) denotes the transpose of the trace \( \bar{x}(t) \). Usually the learning rate \( \beta \) was set to the value of 10. Only for the result shown in fig. 3C, \( \beta \) was set to the value of 3 to ensure that the algorithm converged. The trace \( \bar{x}(t) \) is a slowly decaying version of the event representation \( x(t) \)

\[ \bar{x}(t) = \lambda \bar{x}(t-1) + (1-\lambda) x(t), \quad \bar{x}(0) = 0 \quad \text{(eq. 6)} \]

The traces decrease 0.3% each time step (1 time step = 100 msec, \( \lambda = 0.997 \)), as this produced fast learning. The intertrial interval was long enough for all eligibility traces to decrease to zero, which was simulated by setting the traces to zero. The vectors \( u(t), x(t), \bar{x}(t), p(t) \) and \( e(t) \) are composed of 70 time dependent components per event. (Programs in Matlab code at ftp://ftp.usc.edu/pub/bsl/Suri/Suri_Schultz)

We cannot offer a convergence proof for the proposed model but give some reasons why the model usually converges. If \( \gamma = 0 \), eqs. 4 to 6 adapt the weights of the matrix \( V \) in the correct direction as the weights increase if the prediction is too small and decrease if the prediction is too large. Otherwise, if \( 0 < \gamma < 1 \), the prediction \( p(t+1) \) should be more accurate than the prediction \( p(t) \), because more information is available at time \( t+1 \) than at time \( t \). Therefore, \( p(t+1) \) influences the adaptation of the weights which have been used to compute \( p(t) \) in the previous time step (eq. 2). Learning terminates when the error signals \( e(t) \) has the only nonzero value at onset of the first event of a trial, because the error signals \( e(t) \) and the trace \( \bar{x}(t) \) have to be nonzero for learning (eq. 5). For the shown simulations, the error signals \( e(t) \) converge to the optimal values. In other simulations, large values of the learning rate \( \beta \) or of the discount factor \( \gamma \) abolished convergence as the weights and the prediction signals increased to incorrect large values. This
problem occurs when intertrial intervals are short and persists even when only one event is presented each trial. In this case, event presentation in the current trial predicts progressively more event presentations in future trials. This seems to cause that the algorithm diverges.

Appendix B: Comparison to Other Models
To justify our choice of the proposed algorithm, we compare our model with similar and alternative algorithms. Our model is a compromise between the TD model (Sutton & Barto 1990), the model for latent learning of Sutton and Pinette (1985), and the successor representation of Dayan (1993).

Eq. 2 and eq. 4 - eq. 6 result from replacing the scalar reinforcement prediction of the TD model with a vectorial event prediction. Instead of using eq. 3, the scalar reinforcement prediction \( p(t) \) was computed in the TD model with \( p(t+1) = v x(t+1) \) with the weight vector \( v \) and the representation vector \( x(t) \). As in the TD model, the temporal event representation \( x(t) \) was used to compute the predictions for the events \( u(t) \) (eq. 3).

Eq. 3 was proposed by Sutton and Pinette (1985) who used a more precise higher order expansion in the matrix \( V \). We are not aware of a convergence proof for their model. Dayan (1993) argued that such a high order expansion is very sensitive to errors in the estimate of \( V \) and proposed to use the equation \( p(t+1) = V x(t+1) \) instead of eq. 3. However, Dayan’s algorithm lost the capability to form novel associative chains. In both studies it was not distinguished between events and their temporal representation (\( T = 1 \)).

An further alternative model has been considered. If the error is computed with \( e(t) = \gamma x(t+1) - V x(t) \) instead of eq. 5, the future temporal event representation components could be predicted with the matrix \( V (\gamma x(t+1) = V x(t)) \). The prediction signal could then be computed with \( p(t+1) = T^{-1} \left[ x(t+1) + V x(t+1) + V^2 x(t+1) + \ldots \right] \) instead of eq. 3, and the dopamine-like reward prediction errors could be computed with eq. 4. To avoid computation of the inverse of transformation \( T \), we may assume \( T^{-1} = 1 \) and compute the prediction with \( p(t+1) = x(t+1) + V x(t+1) + V^2 x(t+1) + \ldots \). According to this equation, each correctly learned prediction signal component would increase before its corresponding component of the temporal event representation. As the temporal event representation covers the intratrial interval, this alternative algorithm computes prediction signals with their peak in the intratrial interval. However, these prediction signals would not be consistent with anticipatory neural activities which peak typically at the beginning of sensory events.