Neuron, Vol. 28, 329–337, April 24, 2003, Copyright ©2003 by Cell Press

Temporal Difference Models and Reward-Related Learning in the Human Brain

**John P. O'Doherty,[1,*] Peter Dayan,[2] Karl Friston,[1]
Hugo Critchley,[1] and Raymond J. Dolan[1]**
[1]Wellcome Department of Imaging Neuroscience
Institute of Neurology
[2]Gatsby Computational Neuroscience Unit
University College London
London WC1N 3BG
United Kingdom

Summary

Temporal difference learning has been proposed as a model for Pavlovian conditioning, in which an animal learns to predict delivery of reward following presentation of a conditioned stimulus (CS). A key component of this model is a prediction error signal, which, before learning, responds at the time of presentation of reward but, after learning, shifts its response to the time of onset of the CS. In order to test for regions manifesting this signal profile, subjects were scanned using event-related fMRI while undergoing appetitive conditioning with a pleasant taste reward. Regression analyses revealed that responses in ventral striatum and orbitofrontal cortex were significantly correlated with this error signal, suggesting that, during appetitive conditioning, computations described by temporal difference learning are expressed in the human brain.

Introduction

Central to a number of established classical conditioning theories is a prediction error (PE), which signals a discrepancy between expected outcome or conditioned stimulus (CS) and actual outcome or unconditioned stimulus (UCS) (Rescorla and Wagner, 1972; Pearce and Hall, 1980). Learning occurs through updating expectations about the outcome in proportion to prediction error, so that across trials, the expected outcome converges to the actual outcome. A variant of this theory makes use of an algorithm known as temporal difference (TD) learning (Sutton, 1988; Sutton and Barto, 1990; Friston et al., 1994; Schultz et al., 1997). In this model, as in previous theories, a PE is computed as the difference between the expected and actual outcome.

However, in TD learning (Schultz et al., 1997), timing within a trial is taken into account. The goal of TD learning is to provide a prediction, for each time t in the trial during which a CS is presented, of the total future reward that will be gained in the trial from time t to the end of the trial. This is accomplished by means of a particular PE signal that compares the predicted value at time $t + 1$ to the predicted value at time t. At the beginning of learning, the predicted reward $V(t)$ is zero for each time t until the time at which the reward or UCS is delivered (t_{UCS}). On the next learning trial, a comparison between $V(t_{UCS})$ and $V(t_{UCS} - 1)$ generates a positive prediction error that, in the simplest form of TD learning, is used to increment the value at time t_{UCS} **– 1** (in proportion to an arbitrary learning rate). On subsequent learning trials, $V(t)$ is updated for each time t ranging from t_{UCS} back to t_{CS} (the earliest time at which the CS is presented). Learning is complete when $V(t)$ for each time t is equal to the total reward available in the trial. It follows that in the absence of any change of the reward contingencies from trial to trial, once $V(t)$ has converged to the total available reward, the PE signal or δt at that time point is zero.

In this study, we consider appetitive conditioning in humans, for which the outcome is a taste reward. Single-unit recording of reward learning in nonhuman primates indicates that dopamine neurons have a response profile consistent with a TD-related PE (Schultz, 1998). Prominent targets of the mesolimbic dopamine system are the ventral striatum and orbitofrontal cortex (Oades and Halliday, 1987). Ventral striatum is a subdivision of the basal ganglia that includes the nucleus accumbens, parts of olfactory tubercle, as well as ventral and medial portions of the putamen and caudate nucleus (Holt et al., 1997; Karachi et al., 2002). Recent neuroimaging studies have explored the response profiles in these regions for evidence of PE-related effects in humans. A block-design fMRI study showed that temporally unpredictable delivery of reward produced activation of ventral striatum and orbitofrontal cortex (Berns et al., 2001). A further study of trial-based learning showed a positive BOLD response in ventral striatum when a reward was delivered later than expected (Pagnoni et al., 2002). Although these findings are consistent with TD learning, they are restricted to one component of a putative PE response, namely, what happens when reward delivery is unexpected. Whether patterns of neuronal responses during reward learning in humans can be comprehensively described by TD learning has yet to be determined.

TD learning provides specific predictions about the characteristics of a PE response. These are: before learning, a positive PE response should occur to reward presentation (UCS), but during learning, this response should shift to the CS. Furthermore, after learning, an unexpected reward should lead to a positive PE response at the time of reward delivery, whereas an unexpected reward omission should lead to a negative PE response at the expected time of reward delivery.

In this study, we determined whether responses in human ventral striatum, orbitofrontal cortex (OFC), or other brain areas were consistent with a temporal difference prediction error signal during appetitive learning. To achieve this, we used the actual output of a TD learning algorithm to generate a PE (or δ) response at two main time points in a conditioning trial: the time of presentation of the CS and the time of presentation of the reward. The output of this algorithm was then entered into a regression model of fMRI measurements from subjects who underwent appetitive Pavlovian conditioning. This enabled us to test for brain regions that

*Correspondence: j.odoherty@fil.ion.ucl.ac.uk

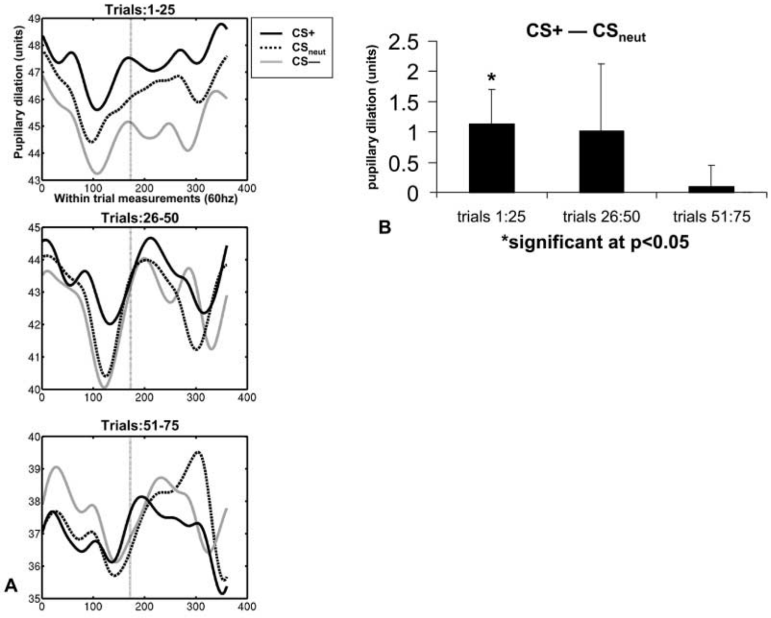

Figure 1. Pupillary Data

(A) Pupillary responses over the course of the 6 s of a trial epoch, measured at 60 hz. The responses are averaged over three blocks of trials, 1–25, 26–50, and 51–75, and are shown separately during each block for the CS+, CS−, and CSneut trials. In each trial epoch, the CS is presented at 0, and in the case of the CS+, the reward is presented at measurement 180 (3 s into the trial). The time at which the reward is presented is shown as a vertical gray line. It can be seen from this figure that a discriminatory anticipatory pupillary response occurs in the first block of trials.

(B) Average discriminatory pupillary responses for CS+ − CSneut trials, sampled between 0 to 3 s after trial onset, averaged across the eight subjects whose data were included in the analysis for the three trial blocks. In the first block of trials, CS+ pupillary responses are significantly larger than CSneut responses (at $p < 0.05$), but in the latter trial blocks this difference is no longer significant.

manifested a full range of TD error-related PE responses: responding initially at the time of reward presentation but over learning transferring responses to a time at which the CS is presented, as well as responding at the time of the UCS following the unexpected delivery or omission of reward.

Given that we did not have an a priori hypothesis as to whether a putative neuronal deactivation from a negative PE response corresponded to a BOLD deactivation or activation, we modeled the positive and negative components of the PE response at the time of presentation of the UCS separately (see Experimental Procedures). This enabled us to test for areas in which, in the context of the full TD response profile, a negative PE response led to a negative BOLD response (signed TD-related PE) and, in addition, to test for areas in which a negative PE response led to a positive BOLD response (absolute value TD-related PE). In order to establish the effects of different learning rates from the TD model on the results obtained, we tested for the effects of a relatively low (0.2) and high (0.7) learning rate and report results for both. In addition to the fMRI data, we also simultaneously acquired pupillary dilation measurements from each subject as an objective index of learning.

Results

Pupillary Dilation

Pupillary dilation responses for three consecutive blocks of trials are shown averaged across subjects in Figure 1A. It can be seen that in CS+ trials (paired with the reward) there are increased anticipatory pupillary responses in the period following the presentation of the CS stimulus but before the delivery of the reward relative to both the CS− (paired with nothing) and CSneut (paired with affectively neutral taste) trials. The subject averaged mean differences in anticipatory pupillary responses between the CS+ and CSneut trials are shown in Figure 1B. Anticipatory pupillary responses across subjects for the CS+ are significantly greater than both the CSneut ($t = 2.02$, $df = 7$; $p < 0.05$ one-tailed) and CS− ($t = 2.51$, $df = 7$; $p < 0.05$ one-tailed) responses for the first trial block. However, this difference is no longer significant in the latter trial blocks, perhaps reflecting habituation (as evident from inspection of both Figures 1A and 1B). In a separate analysis, the maximum pupillary response in the anticipatory period (using unsmoothed pupillary data) was also found to be significantly greater for the CS+ than for the CS− trials in the first trial block ($t = 1.94$; $p < 0.05$ one-tailed), although a comparison between the CS+ and CSneut trials did not reach significance ($t = 1.35$, $p = 0.11$).

Neuroimaging Results with Learning Rate $\alpha = 0.2$

Signed PE Response, Responding to Both δt_{CS} and Signed(δt_{UCS})

Significant responses were found in ventral striatum, specifically, ventral putamen, to both δt_{CS} and *signed* (δt_{UCS}) components (left: −27, 3, −9; right: 27, −9, −9; Figure 2A). The left ventral putamen activation survived correction at $p < 0.05$ using an ~9.4 cm^3 anatomically delineated mask defined over the ventral striatum (extending from caudal putamen to nucleus accumbens). Significant effects were also evident in ventral globus pallidum, left orbitofrontal cortex (Figure 2B), as well as in dorsal prefrontal cortex, including inferior and middle frontal gyrus (coordinates listed in Table 1). Effects were also found in bilateral cerebellum, significant at $p < 0.05$ corrected for small volume in a 20 mm sphere centered on coordinates reported by Ploghaus et al. (2000), in a study of prediction error in relation to aversive conditioning with thermal pain (Figure 2B). Fitted effects for each regressor component (δt_{CS} and positive and negative components of the δt_{UCS} regressor) are plotted for ventral striatum, orbitofrontal cortex, and cerebellum in Figure 3A). It should be noted that due to the extra sum of squares principle in the general linear model, the effects of this nonstationary TD-related response profile accounts for variance over and above the effects of a stationary response profile (also included in the fMRI

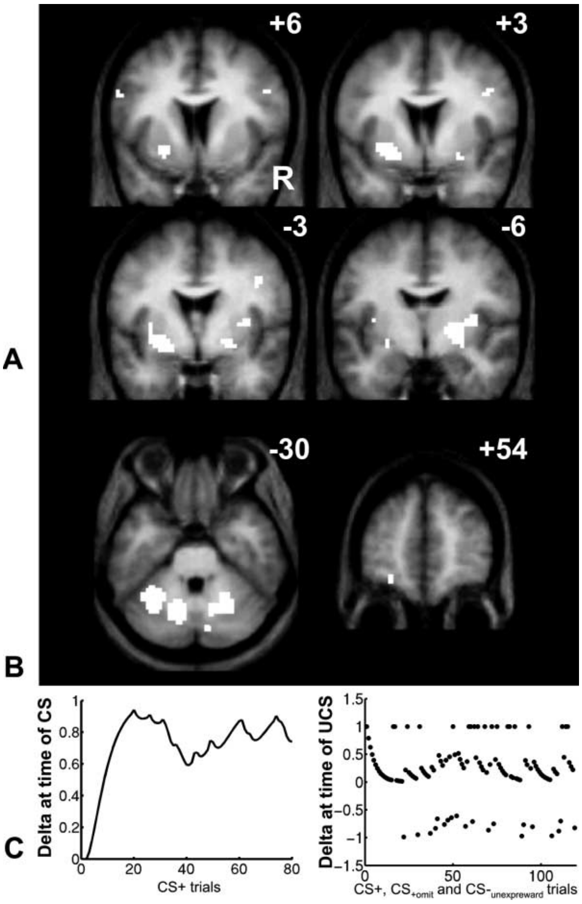

Figure 2. Neuroimaging Results

(A) Regions of striatum (ventral putamen), showing significant effects to δt_{CS} masked inclusively by δt_{UCS} at $p < 0.001$.

(B) Responses to δt_{CS} and δt_{UCS} in cerebellum (left of figure) and left orbitofrontal cortex (right of figure), again at $p < 0.001$.

(C) Illustration of δt_{CS} and δt_{UCS} regressors. On the left of the figure, the values of δt_{CS} for each CS+ trial in the experiment as determined by the TD learning model with $\alpha = 0.2$. The nonzero responses of δt_{UCS} for CS+, CS$+_{omit}$, and CS$-_{unexpreward}$ trials are shown to the right of the figure.

analysis), where there is no change in CS responses over time.

A similar pattern of activations was also observed in striatum and elsewhere when including the additional four subjects who did not evaluate the glucose as pleasant at the end of the experiment, except that left orbitofrontal cortex no longer showed significant effects. A direct comparison between TD responses from the nine subjects who did report the glucose as pleasant and TD responses from the four subjects who did not revealed a significantly greater response in left OFC in the nine subjects who did report the glucose as pleasant (at $p < 0.001$). However, due to the small number of subjects in the groups, this analysis lacks power and is thus reported only descriptively.

Absolute Valued PE Response, Responding to Both δt_{CS} and Absolute(δt_{UCS})

The contrast of δt_{CS} masked by the absolute value of δt_{UCS} revealed significant effects in inferior frontal cortex, right inferior frontal sulcus, and adjoining inferior frontal gyrus. No other brain regions showed significant effects at the selected threshold. Fitted effects in these regions for each regressor component are plotted in Figure 3B. It can be seen from the figure that, in the case of inferior frontal gyrus, the negative δ component is in fact negative, and indeed this region was also identified in the test for signed PE responses. The reason that this region was identified in both analyses is that the contrasts of (δt_{UCS} *positive* − δt_{UCS} *negative*) and (δt_{UCS} *positive* + δt_{UCS} *negative*) are both significantly above baseline. Thus, the only region showing a truly positive BOLD response following a negative PE signal is inferior frontal sulcus.

Neuroimaging Results with Learning Rate $\alpha = 0.7$

Signed PE Response, Responding to Both δt_{CS} and Signed(δt_{UCS})

This analysis revealed effects in the same brain areas reported above, except that the significance levels of activations in regions of interest differed. In particular, responses in left ventral striatum were less significant than with $\alpha = 0.2$ (with a peak z score of 4.02 instead of 4.25 reported above). On the other hand, left orbitofrontal cortex and right ventral striatum showed stronger effects (in left OFC, peak z = 3.58, and in right striatum, peak z = 4.59, respectively). In addition, right orbitofrontal cortex now showed significant effects (15, 45, −18; peak z = 3.1).

Absolute Valued PE Response, Responding to Both δt_{CS} and Absolute(δt_{UCS})

No regions showed significant effects in this comparison for $\alpha = 0.7$.

Formal Comparison between Different Learning Rates

We formally tested for a difference in responses between the two learning rates reported in the above analysis by incorporating both learning rates in the same model and performing a linear contrast between them (see Experimental Procedures). There were no significant differences between the learning rates in the ventral striatal or orbitofrontal cortex regions reported above at $p < 0.001$ uncorrected, although, in a part of left cerebellum, the slower learning rate was a significantly better fit (−18, −60, −30; $p < 0.001$). In an analysis restricted to voxels that showed significant effects to the slower learning rate ($\alpha = 0.2$) (reported above), a part of left striatum showed stronger responses to $\alpha = 0.2$ than $\alpha = 0.7$, at $p < 0.05$ uncorrected (−33, 0, −6). No regions that showed effects to the higher learning rate at $p < 0.001$ uncorrected were found to respond significantly more to the higher learning rate than the lower learning rate, even at $p < 0.05$ uncorrected.

Time Course Analysis

We plotted the averaged event-related evoked responses to the different trial types as a function of time. In Figure 4, the time course of the BOLD signal is shown for each trial type averaged across subjects from ventral putamen. Also shown are the predicted hemodynamic response functions (HRF) that would result from convolving a pair of stick functions representing the putative δ signal at the time of the CS and the time of the presentation of the reward (UCS) with a canonical HRF. Plotted in Figure 4A is the predicted time course that would

Table 1. Regions with BOLD Responses Conforming to δt_{CS} and δt_{UCS} Components, with $\alpha = 0.2$

		Tal X	*Tal Y*	*Tal Z*	*Z Score*[a]	
δt_{CS} inclusively masked with signed δt_{UCS}						
Striatum (ventral putamen)	Left	−27	3	−9	4.25	$p < 0.05$svc
	Right	27	−9	−9	3.74	
	Left	−27	15	−3	4.13	
Orbitofrontal cortex	Left	−21	45	−9	3.38	
	Left	−24	54	−18	3.17	
Globus pallidus	Left	−12	−21	−6	3.34	
	Right	24	−15	−6	4.1	
Cerebellum	Left	−30	−51	−30	4.05	$p < 0.05$svc
	Right	24	−60	−30	3.83	$p < 0.05$svc
Precuneus		0	−54	51	4.23	
Left parietal cortex	Left	−42	−33	33	3.91	
Inferior frontal gyrus	Left	−57	9	27	3.55	
	Right	54	12	18	3.44	
	Right	48	48	15	3.72	
Middle frontal gyrus	Left	−33	21	57	3.23	
	Right	39	18	54	3.25	
Para-cingulate cortex	Right	3	9	51	3.34	
δt_{CS} inclusively masked with absolute value of δt_{UCS}						
Inferior frontal sulcus	Right	42	3	30	3.6	
		45	15	30	3.36	
Inferior frontal gyrus	Right	54	15	18	3.37	

[a] Significance level corresponds to z scores of voxels from δt_{CS} regressor.

occur in the early stages of learning in which the neuronal response representing the δ signal occurs mostly at the time of the presentation of the reward. Plotted alongside this is the observed BOLD signal averaged over the first ten trials of the experiment (early CS+). Shown in Figure 4B is the predicted time course that would occur after learning in which the neuronal response representing the δ signal has transferred to the time of the presentation of the CS, 3 s earlier in the trial. Also shown is the observed BOLD signal averaged over the remaining CS+ trials in the experiment: mid (trials 21–40) and late CS+ (trials 41–80) (Figure 4C). These results show that, consistent with the model predictions, there is a transfer in the time to peak of the actual hemodynamic response in this region as learning progresses from 8 s (early CS+) to 4–6 s (mid CS+). Responses to $CS-_{unexpreward}$ and $CS+_{omit}$ trials are shown for the striatum in Figures 4D and 4E.

TD Analysis of CSneut Condition

We also tested whether TD-related PE responses occurred in trials in which the neutral taste was presented (CSneut). Although the neutral condition was not completely balanced with the reward condition in that there were no unexpected omissions or unexpected receipt trials, we were at least able to test for areas in which a shift in the PE response occurred over learning from the time of the presentation of the neutral taste to time of presentation of the CS. No significant PE-related responses were found in ventral striatum, orbitofrontal cortex, or elsewhere in the CSneut condition at the same stringent threshold used to detect reward-related re-

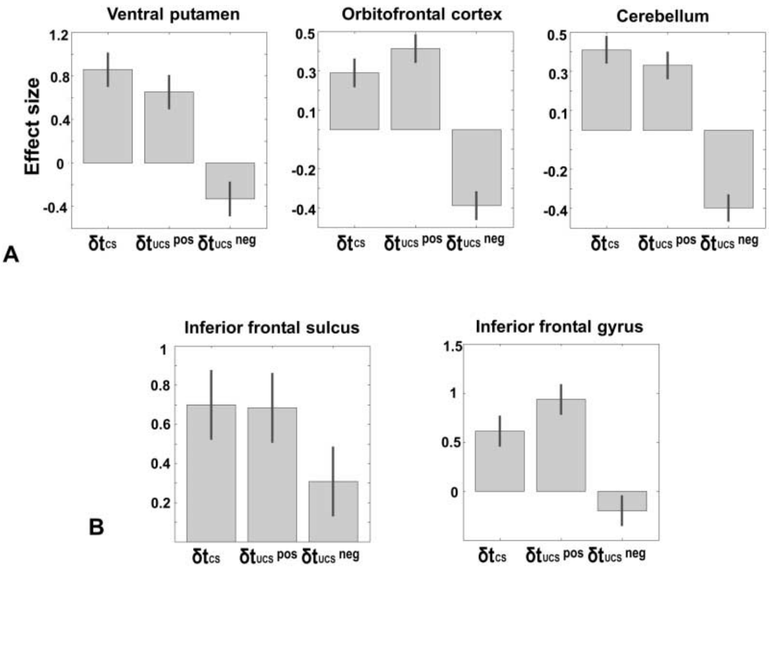

Figure 3. Plots of Effect Sizes from Second-Level SPM Analysis, Shown for Regions of Interest

Effect size (parameter estimates) from regressors for δt_{CS} as well as positive and negative components of δt_{UCS}.

(A) Regions showing responses consistent with a signed PE as identified by a contrast of δt_{CS} *and* δt_{UCS}(*positive*) – δt_{UCS}(*negative*).

(B) Regions identified by a contrast to test for absolute PE responses: δt_{CS} *and* δt_{UCS}(*positive*) + δt_{UCS}(*negative*).

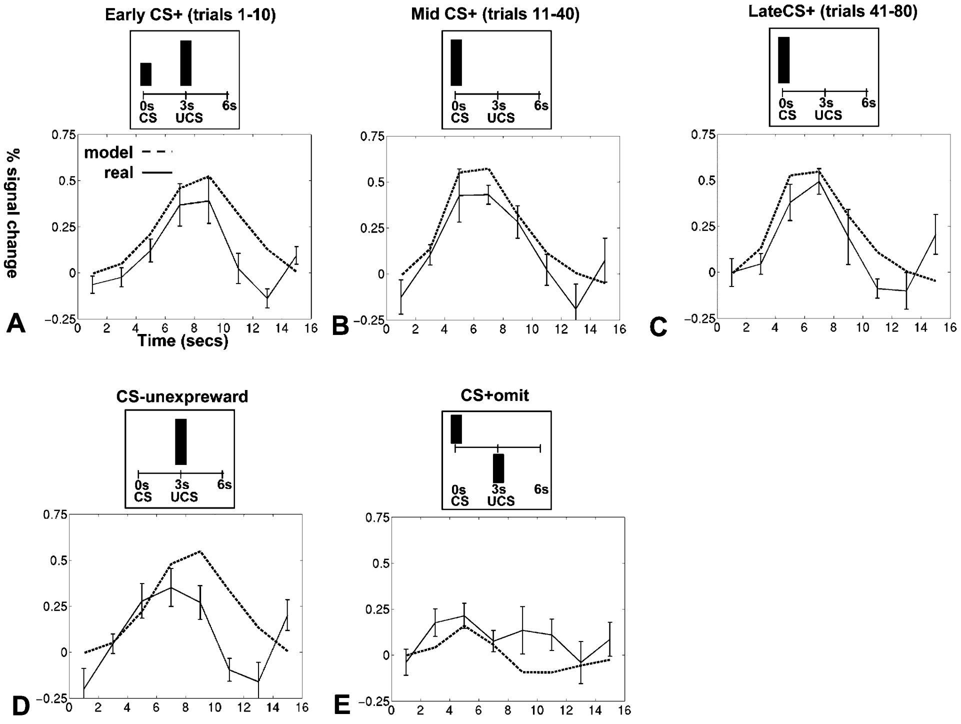

Figure 4. Time Course of Evoked Responses in Left Ventral Striatum

All averaged evoked responses are shown relative to the baseline condition of CS−. Also shown are predicted average responses scaled to the observed BOLD responses (by equating the maximum – minimum values).

(A) Averaged evoked responses during the early stages of learning (first ten trials of the CS+ condition). Also plotted is the predicted average response over the first ten trials in which, on the first trial, the prediction error signal responds only at the time of the reward, but over the ten trials, transfers back to the time of the presentation of the CS. By the tenth trial, this transfer has not occurred completely. Note that the time to peak of the HRF occurs later on in the response profile than for the mid to late CS+ trials plotted in (B) and (C).

(B) Averaged evoked responses during trials 11 to 40 of the CS+ condition. Also plotted is the predicted response after learning in which a neuronal response occurs only at the time of the presentation of the CS. This transfer of neuronal response leads to a shift in the time to peak of the estimated HRF. This transfer is evident in the averaged evoked responses.

(C) Averaged evoked responses during late CS+ trials (from 41 to 80 trials).

(D and E) Plots of observed and predicted hemodynamic responses for $CS-_{unexpreward}$ and $CS+_{omit}$ trials, together with the putative neuronal responses from which the predicted responses are derived.

sponses (in which the δt_{CS} contrast at $p < 0.001$ was inclusively masked by the signed δt_{UCS} contrast at $p < 0.001$ uncorrected). The less stringent test of the conjunction between the two regressors, with the *conjoint* threshold set at $p < 0.001$, revealed effects in left ventrolateral prefrontal cortex (−33, 33, 3; z = 4.27), medial prefrontal cortex (18, 54, 6; z = 4.28), and right cerebellum (12, −45, −18; z = 3.99) but not in ventral striatum or orbitofrontal cortex. A direct comparison between PE responses in the reward condition and in the neutral condition revealed a significantly greater response in left putamen (−18, 0, −9; z = 3.18) and right orbitofrontal cortex (21, 36, −9; z = 4.63) at $p < 0.001$ uncorrected [using a conjunction of one contrast testing for a difference between δt_{CS}(*reward*) and δt_{CS}(*neutral*) responses *and* another contrast testing for a difference between signed δt_{UCS}(*reward*) and signed δt_{UCS}(*neutral*) responses].

Discussion

We demonstrate transient learning-related changes in the brain during appetitive classical conditioning. Most notably, the output of a PE response from a temporal difference model of learning predicts neuroimaging data, producing a significant fit to the observed BOLD responses in ventral striatum, orbitofrontal cortex, and a number of other brain regions. Single-unit neurophysiological recordings demonstrate that dopamine neurons, projecting from VTA and substantia nigra to the striatum and frontal cortex, exhibit a pattern of responses during learning that have the characteristics of a TD learning-related prediction error (Schultz, 1998). One of the critical findings underlying this hypothesis is the observation that dopamine neurons transfer their responses over the course of learning from the time at which the reward is received to the time at which the cue stimulus is presented (Ljungberg et al., 1992; Mirenowicz and Schultz, 1994). In the present study, we observed a similar backward shift in the time to peak of the hemodynamic response in the ventral striatum over the course of learning, a finding that reflects a transfer in evoked neuronal activity at a population level from time of actual receipt of a reward to the time at which a predictive cue stimulus was presented.

A crucial feature of the TD model is that if a reward fails to occur when predicted, a negative prediction error occurs at the time when the reward was expected (Friston et al., 1994). It is known that under these conditions in the nonhuman primate, dopamine neurons depress their firing from baseline at the time when the reward was supposed to occur (Mirenowicz and Schultz, 1996). In the present study, we did not have an a priori hypothesis as to whether a putative neuronal deactivation would lead to an increase or decrease in BOLD signal from baseline. Consequently, we tested both a signed prediction error model, in which a negative prediction error led to a negative BOLD signal, and an absolute prediction error model, in which a negative prediction error led to a positive BOLD signal. In the two main regions of interest, as well as in other parts of the brain, such as cerebellum and other parts of prefrontal cortex, the signed prediction error model showed significant effects. The absolute prediction error model only showed effects in a circumscribed part of inferior prefrontal gyrus.

Our findings can be contrasted with those of Pagnoni and colleagues (Pagnoni et al., 2002). These authors showed that when a reward was delivered later than expected in a trial, a positive BOLD signal was generated relative to trials in which the reward was delivered as expected. However, the direction of the response suggests that in their results a negative prediction error generated a positive BOLD signal. One difference between the two studies is that in the study by Pagnoni and colleagues, a region of interest analysis was performed in which effects were reported exclusively from the nucleus accumbens, a part of ventral striatum in which we were unable to observe significant responses. However, we found significant effects in another part of ventral striatum, corresponding to ventral putamen. One possibility is that the response profile in nucleus accumbens differs from that in ventral putamen and elsewhere. However, in a study using monetary reward, BOLD responses in nucleus accumbens were found to decrease from baseline when an expected reward was not delivered, suggesting that BOLD deactivation for a negative prediction error might under some conditions be expressed in this part of striatum (Knutson et al., 2001a). In any case, it is clear from our data that a negative prediction error leading to a negative BOLD signal is more common throughout the brain than the converse.

An absolute valued PE response may also have some of the characteristics of an attentional signal, in that a positive response to both an unexpected omission as well as an unexpected delivery of reward could reflect the increased salience of those events. Our finding that this type of response is not widespread throughout the brain in reward learning suggests that this specific form of attentional modulation is unlikely to account for the majority of TD PE-related responses. This does not rule out the possibility that other forms of attentional modulation can occur in regions in which we observed TD responses (Nobre et al., 1999), nor does it preclude the possibility that attention plays an important role in conditioning (see Dayan et al., 2000).

A number of previous neuroimaging findings implicate human ventral striatum in reward (Delgado et al., 2000; Elliott et al., 2000) and, specifically, in anticipation for monetary as well as taste reward (Breiter et al., 2001; Knutson et al., 2001b; O'Doherty et al., 2002). The present results indicate that anticipatory responses in ventral striatum and orbitofrontal cortex can be accounted for by temporal difference learning, with afferent dopamine input providing a prediction error. However, single-unit neurophysiology data in nonhuman primates indicate that some neurons in ventral striatum and orbitofrontal cortex show a more complicated pattern of responses during reward anticipation than the phasic activity of dopamine neurons. For example, spiny neurons in striatum have been found that demonstrate sustained responses during the anticipatory period which terminate when the reward is presented (Schultz et al., 1998; Apicella et al., 1992). In addition, neurons have been found throughout the prefrontal cortex and in orbitofrontal cortex in particular that demonstrate sustained responses during the period before a reward is expected (Schoenbaum et al., 1998; Hikosaka and Watanabe, 2000). Thus, prediction error-related responses that we describe are likely to capture only some of the response characteristics of these regions during appetitive learning.

We note that a number of previous imaging studies of associative learning have sought to explain their results in terms of formal learning models (Ploghaus et al., 2000; Fletcher et al., 2001). In particular, a study of aversive conditioning with thermal pain as the UCS reported both signed and absolute valued prediction error effects in cerebellum, hippocampus, and frontal cortex (Ploghaus et al., 2000). These were interpreted as consistent with one of two associative learning theories: Rescorla-Wagner or Pearce-Hall (Rescorla and Wagner, 1972; Pearce and Hall, 1980). In contrast to these earlier learning rules, the temporal difference rule can incorporate interval timing within a trial and provides specific predictions about the nature of the neuronal response during learning at both the time of presentation of the CS and reward. Previous fMRI studies have shown that an unpredicted reward produces a response consistent with a TD-related prediction error (Berns et al., 2001; Pagnoni et al., 2002). The present study extends these findings by showing that the temporal difference rule accounts for initial learning, as well as for both positive and negative errors in prediction that occur following a violation of expectation once learning has occurred.

It is of interest that we did not find significant PE-related effects in the amygdala, given that previous studies have reported reward expectation responses in this region (Knutson et al., 2001a; O'Doherty et al., 2002). Although the amygdala is likely to contribute toward the formation of stimulus-reward associations, our data suggest that it does not represent a TD-related prediction error, at least of the same form as that observed in ventral striatum and OFC.

To conclude, we show that responses in human ventral striatum and orbitofrontal cortex can be described by a theoretical learning model: temporal difference learning. Specifically, the output of a TD learning algorithm accounts for neuronal responses in the human brain, responding initially to the presentation of the reward, and over learning transferring its response to the time of presentation of the CS. On the basis of evidence from nonhuman primates, it is reasonable to infer that a source of TD learning-related activity in these regions is the modulatory influence exerted by phasic responses of afferent dopamine neurons.

Figure 5. Abstract Fractals Used in the Study as the Visual Conditioned Stimuli

One stimulus was allocated to each of the main trial types (CS+, CS−, and CSneut), in a manner that was counterbalanced across subjects. Stimuli were presented on a gray background. On each trial, one stimulus was presented on the screen for 3 s and was then removed and replaced by a fixation cross. On trials in which a taste was delivered, this coincided with the termination of the visual stimulus.

Experimental Procedures

Subjects

Thirteen right-handed healthy normal subjects participated in the experiment (mean age, 28.11; range, 18–42) of which nine were female. The subjects were preassessed to exclude those with a prior history of neurological or psychiatric illness. All subjects gave informed consent, and the study was approved by the Joint Ethics Committee of the National Hospital for Neurology and Neurosurgery. Subjects were asked to refrain from eating or drinking sweet drinks for 4 hr prior to scanning and were thus in a mildly food-deprived state, increasing the likelihood of the glucose taste being perceived as pleasant (mean hunger ratings were 1.38 ± SEM [0.25], using a scale from +2 = very hungry through −2 = not at all hungry).

Imaging Procedure

A 2 Tesla Siemens Vision MRI scanner was used to acquire gradient echo T2* weighted echo-planar images (EPI) images with BOLD (blood oxygenation level dependent) contrast. Each volume comprised 33 axial slices of 3.3 mm thickness and 3 mm in-plane resolution. A total of 685 volumes were acquired continuously every 2.506 s. These parameters produced EPI images in which signal dropout from susceptibility artifact was restricted to far caudal OFC, leaving the remaining sectors of OFC intact. Subjects were placed in a light head restraint within the scanner to limit head movement during acquisition. To detect transient head movements due to swallowing, we attached a 1.5 cm long copper coil with a radius of 0.5 cm to the neck of each subject. Small movements of the coil induced a current in the magnetic field that could be detected when amplified using one channel of an EEG system positioned in the scanner room (National Hospital for Neurology and Neurosurgery, London, UK). This produced a time series over the whole experiment in which signal changes represented transient head movement. Due to technical difficulties, we obtained useable data from the movement detector coil in only 9 out of 13 subjects (6 of which were included in the main analysis reported in this paper). A T1-weighted structural image was also acquired for each subject.

Apparatus

The tastes were contained in four 50 ml syringes (two for the pleasant and two for the neutral taste), which were attached to an SP220I electronic syringe pump (World Precision Instruments Ltd, Stevenage, UK), positioned in the scanner control room, and delivered to the subjects via two separate 6 meter long 3 mm wide polythene tubes. The syringes were also attached to a computer-controlled valve system which enabled the different tastes to be delivered independently along the tubing. The apparatus was controlled by the stimulus presentation computer positioned in the control room, which also received volume trigger pulses from the scanner, and the visual stimuli were presented on a projector screen positioned ~10 cm away from the subject's face. Pupil dilation was recorded by using a model 504 fMRI eyetracking system (Applied Science Laboratories, Bedford, MA).

Experimental Design

Each trial consisted of the presentation of one of three arbitrary visual stimuli (which were abstract fractals, see Figure 5) followed 3 s later by either 0.5 ml of a pleasant sweet taste (1 M glucose) (CS+), a neutral taste (CSneut: consisting of the main ionic components of saliva [Francis et al., 1999; O'Doherty et al., 2001, 2002]), or no taste (CS−). The visual stimuli were presented on a gray background and were removed from the screen after 3 s to coincide with taste delivery. A fixation cross was then presented for the remainder of the trial. After a further 3 s, the next trial was scheduled. The allocation of each stimulus to a given trial type was counterbalanced across subjects. There was a total of 280 trials in the experiment, 100 each of CS+ and CS− and 80 of CSneut. The whole experiment lasted a total of ~28.5 min. After the first ten CS+ stimuli had been presented and paired on each occasion with reward, in 20 out of 90 subsequent CS+ presentations, the reward was omitted (CS+_{omit}). Further, in 20 presentations of the CS−, a reward was unexpectedly delivered ($\text{CS−}_{\text{unexpreward}}$). The CSneut condition was primarily included to provide a low valence rinse for the glucose taste during the experiment. The order of presentation of events was randomized within MATLAB (Mathworks Inc) and presented to subjects using Cogent 2000 stimulus presentation software (Wellcome Department of Imaging Neuroscience, London, UK). At the end of the experiment, all 13 subjects were invited to provide pleasantness ratings for the glucose and neutral tastes, using a scale ranging from −2 (very unpleasant) through 0 (neutral) up to +2 (very pleasant). Of those, eight subjects rated the glucose taste as pleasant (e.g., >0, with a mean pleasantness rating of 1.375 ± 0.18 (SEM), whereas the control taste was given a pleasantness rating of 0.5 ± 0.27 (SEM). A further three subjects rated the glucose as aversive (with ratings < 0), and two subjects rated the glucose as neutral (with ratings = 0). Consequently, we report neuroimaging and behavioral results from 9 out of 13 subjects, including the 8 subjects who rated the glucose as pleasant, as well as an additional subject who rated the glucose as affectively neutral but nevertheless preferred the glucose to the neutral control taste.

Analysis of Pupillary Data

Analysis of pupillary data was restricted to the nine subjects who were included in the neuroimaging analysis. Of those, the pupillary data from one subject had to be subsequently excluded due to significant instability in the signal during the session. Two analyses were performed on the pupillary data. In the first analysis, the data were smoothed with a low-pass filter of 1.5 s, and then binned into 6 s trial epochs, separately for each trial type. Mean pupillary responses in the anticipatory period (between 0 to 3 s into the trial) were then calculated for three trial types (CS+, CS−, and CSneut), separately for three consecutive blocks of 25 trials. In the second analysis, the data were not smoothed but were again binned into 6 s trial epochs. The maximum pupillary response in the anticipatory period was then calculated for each trial type, separately for each consecutive block of 25 trials. The motivation for this second analysis is that this approach is less susceptible to the possibility that variable blink rates between trials could lead to apparent differences in pupillary diameter.

Temporal Difference Learning Model

The TD learning model used in this study is that described by Schultz et al. (1997). On each trial, the predicted value (V) at any time t within a trial is calculated as a linear product of the weights w_i and the presence or absence of a CS stimulus at time t, coded in the stimulus representation vector $x_i(t)$:

$$\hat{V}(t) = \sum_i w_i x_i(t).$$

Learning occurs by updating the predicted value of each time point t in the trial by comparing the value at time $t + 1$ to that at time t, leading to a prediction error or $\delta(t)$:

$$\delta(t) = r(t) + \gamma \hat{V}(t + 1) - \hat{V}(t)$$

where $r(t) =$ reward at time t.

The parameter γ is a discount factor, which determines the extent to which rewards that arrive earlier are more important than rewards that arrive later on. In the present study, we set $\gamma = 0.99$. The weights w_i are then updated on a trial-by-trial basis according to the correlation between prediction error and the stimulus representation:

$$\Delta w_i - \alpha \sum_t x_i(t)\delta(t)$$

where $\alpha =$ learning rate.

In the TD model, we assigned six time points to each trial and used each subject's individual event history as input. On each trial, the CS was taken to be delivered at time point 1, and the reward was delivered at time point 3. The stimuli x_i corresponding to the presence of the CS were represented as vectors in which the ith component was = 1 and 0 elsewhere (as used by Schultz et al., 1997). We note that other stimulus representations are possible, and these could lead to a different output from the model. Given that the learning rate of the model (α) was not known a priori, we chose two values for α: a lower learning rate ($\alpha = 0.2$) and a higher learning rate ($\alpha = 0.7$). The rationale for choosing these specific learning rates was that, with the specific stimulus representations used, learning rates outside the range of 0.1 and 0.9 produced a learning profile that was not plausible (i.e., did not converge until after the end of the experiment or converged in a single trial). Thus, the two learning rates used were deemed to be a reasonable sample of a relatively low and high learning rate within the range of "biologically" plausible learning rates.

To generate regressors corresponding to PE responses separately for CS and UCS trial components for the SPM analysis (see below), δt_{CS} was sampled at time point 1 in the trial, and δt_{UCS} was sampled at time point 3.

Image Analysis

Image analysis was performed using SPM99 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). To correct for subject motion, the images were realigned to the first volume, spatially normalized to a standard T2* template with a resampled voxel size of 3 mm^3, and spatial smoothing was applied using a gaussian kernel with a full width at half maximum (FWHM) of 10 mm (Friston et al., 1995). Intensity normalization and high-pass temporal filtering (using a filter width of 160 s) were also applied to the data. The subject-specific δt_{CS} and δt_{UCS} components were then convolved with the hemodynamic response function and fitted to the data for each single subject. It is possible for δt_{UCS} to be either positive or negative, depending on whether an expected reward is omitted or whether a reward is unexpectedly received. Given that it is not known whether a negative value for δt_{UCS} would correspond to a BOLD deactivation or activation, the positive and negative components of δt_{UCS} were modeled separately. It was then possible to test separately for areas with BOLD signal conforming to δt_{CS} and the absolute value of δt_{UCS} (δt_{UCS} *positive* + δt_{UCS} *negative*) or δt_{CS} and the real value of δt_{UCS} (δt_{UCS} *positive* − δt_{UCS} *negative*). Also included in the SPM model was the temporal derivative of the hemodynamic response for each trial type. Non-learning related stationary responses to the CS+, CS−, or CSneut stimuli were also modeled as regressors of no interest. In addition, the six scan-to-scan motion parameters produced during realignment were included to account for residual effects of scan-to-scan motion. To take into account transient head motion effects produced by, for example, swallowing, we also included an additional motion regressor that featured the output of the motion detector coil, band-pass filtered appropriately and subsampled to the number of scans in the experiment.

The results from each subject were taken to the random effects level by including the β images for the δt_{CS} and two δt_{UCS} regressors from each single subject in a one-way analysis of variance with no mean term. Sphericity correction was applied to correct for possible violation of the independence assumption for the three regressors. In the main analysis of TD responses reported in this paper, the statistical thresholds used are $p < 0.001$ uncorrected separately for both the δt_{CS} contrast and δt_{UCS} contrast (real or absolute). To identify regions responding to a TD prediction error at both the time of the CS and the UCS, we report regions that are conjointly significant at $p < 0.001$ in both of these contrasts. This was achieved by inclusively masking the δt_{CS} t map (with threshold set at $p < 0.001$) by the δt_{UCS} t map (with threshold set at $p < 0.001$). The p values we report correspond to the significance levels for the δt_{CS} contrast alone, but it should be noted that, as each region we reported is also significant at $p < 0.001$ uncorrected for the δt_{UCS} contrast, the actual conjoint level of significance (i.e., the combined probability of a voxel being significant at that level in both contrasts) is much more stringent. For regions predicted a priori, we report responses that survive small volume correction using a spherical region of interest centered on coordinates derived from a previous study or else using a binary mask defined over the extent of the anatomical region.

An additional analysis was also conducted in which, in a separate model, TD PE responses to the neutral taste condition were included alongside TD PE responses to the reward condition. Further in this model, TD responses corresponding to both high and low learning rates were included as separate regressors (for both the reward and neutral conditions). The results of linear contrasts from each subject were taken to the random effects level as described above. This enabled a comparison of reward PE TD responses (averaged over the two learning rates) to neutral PE TD responses (similarly averaged), as well as a direct comparison between the different learning rates.

In order to obtain event-related plots, fMRI time courses were extracted from peak voxels (to the δt_{CS} regressor) at the individual subject level in the striatum. These were then binned into events and averaged across subjects.

The structural T1 images were coregistered to the mean functional EPI images for each subject and normalized using the parameters derived from the EPI images. Anatomical localization was carried out by overlaying the t maps on a normalized structural image averaged across subjects and with reference to the anatomical atlases of Duvernoy (1995, 1999).

Acknowledgments

R.J.D. and K.F. are supported by Wellcome Trust program grants. P.D. is supported by the Gatsby Charitable Foundation. P.D. would like to thank Read Montague and Sam McClure for discussions. The authors would also like to thank Eric Featherstone and Oliver Josephs for help with technical aspects of taste delivery and motion/swallowing detection.

Received: August 8, 2002
Revised: February 6, 2003
Accepted: March 7, 2003
Published: April 23, 2003

References

Apicella, P., Scarnati, E., Ljungberg, T., and Schultz, W. (1992). Neuronal activity in monkey striatum related to the expectation of predictable environmental events. J. Neurophysiol. *68*, 945–960.

Berns, G.S., McClure, S.M., Pagnoni, G., and Montague, P.R. (2001). Predictability modulates human brain response to reward. J. Neurosci. *21*, 2793–2798.

Breiter, H.C., Aharon, I., Kahneman, D., Dale, A., and Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron *30*, 619–639.

Dayan, P., Kakade, S., and Montague, P.R. (2000). Learning and selective attention. Nat. Neurosci. Suppl. *3*, 1218–1223.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., and Fiez, J.A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. J. Neurophysiol. *84*, 3072–3077.

Duvernoy, H.M. (1995). The Human Brain Stem and Cerebellum (Vienna: Springer-Verlag).

Duvernoy, H.M. (1999). The Human Brain (Vienna: Springer-Verlag).

Elliott, R., Friston, K.J., and Dolan, R.J. (2000). Dissociable neural responses in human reward systems. J. Neurosci. *20*, 6159–6165.

Fletcher, P.C., Anderson, J.M., Shanks, D.R., Honey, R., Carpenter, T.A., Donovan, T., Papadakis, N., and Bullmore, E.T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nat. Neurosci. *4*, 1043–1048.

Francis, S., Rolls, E.T., Bowtell, R., McGlone, F., O'Doherty, J., Browning, A., Clare, S., and Smith, E. (1999). The representation of the pleasantness of touch in the human brain, and its relation to taste and olfactory areas. Neuroreport *10*, 453–459.

Friston, K.J., Tononi, G., Reeke G.N., Jr, Sporns, O., and Edelman, G.M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience *59*, 229–243.

Friston, K.J., Ashburner, J., Poline, J.B., Frith, C.D., Heather, J.D., and Frackowiak, R.S. (1995). Spatial registration and normalisation of images. Hum. Brain Mapp. *2*, 165–189.

Hikosaka, K., and Watanabe, M. (2000). Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. Cereb. Cortex *10*, 263–271.

Holt, D.J., Graybiel, A.M., and Saper, C.B. (1997). Neurochemical architecture of the human striatum. J. Comp. Neurol. *384*, 1–25.

Karachi, C., Francois, C., Parain, K., Bardinet, E., Tande, D., Hirsch, E., and Yelnik, J. (2002). Three-dimensional cartography of functional territories in the human striatopallidal complex by using calbindin immunoreactivity. J. Comp. Neurol. *450*, 122–134.

Knutson, B., Fong, G.W., Adams, C.M., Varner, J.L., and Hommer, D. (2001a). Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport *12*, 3683–3687.

Knutson, B., Adams, C.M., Fong, G.W., and Hommer, D. (2001b). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J. Neurosci. *21*, RC159.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. J. Neurophysiol. *67*, 145–163.

Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. J. Neurophysiol. *72*, 1024–1027.

Mirenowicz, J., and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. Nature *379*, 449–451.

Nobre, A.C., Coull, J.T., Frith, C.D., and Mesulam, M.M. (1999). Orbitofrontal cortex is activated during breaches of expectation in tasks of visual attention. Nat. Neurosci. *2*, 11–12.

O'Doherty, J., Rolls, E.T., Francis, S., Bowtell, R., and McGlone, F. (2001). Representation of pleasant and aversive taste in the human brain. J. Neurophysiol. *85*, 1315–1321.

O'Doherty, J., Deichmann, R., Crtichley, H.D., and Dolan, R.J. (2002). Neural responses during anticipation of a primary taste reward. Neuron *33*, 815–826.

Oades, R.D., and Halliday, G.M. (1987). Ventral tegmental (A10) system: neurobiology. 1. Anatomy and connectivity. Brain Res. *434*, 117–165.

Pagnoni, G., Zink, C.F., Montague, P.R., and Berns, G.S. (2002). Activity in human ventral striatum locked to errors of reward prediction. Nat. Neurosci. *5*, 97–98.

Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. *87*, 532–552.

Ploghaus, A., Tracey, I., Clare, S., Gati, J.S., Rawlins, J.N., and Matthews, P.M. (2000). Learning about pain: the neural substrate of the prediction error for aversive events. Proc. Natl. Acad. Sci. USA *97*, 9281–9286.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Classical Conditioning II: Current Research and Theory, A.H. Black and W. F. Prokasy, eds. (New York: Appleton Century Crofts), pp. 64–99.

Schoenbaum, G., Chiba, A.A., and Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat. Neurosci. *1*, 155–159.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J. Neurophysiol. *80*, 1–27.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Schultz, W., Tremblay, L., and Hollerman, J.R. (1998). Reward prediction in primate basal ganglia and frontal cortex. Neuropharmacology *37*, 421–429.

Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. Machine Learning *3*, 9–44.

Sutton, R.S., and Barto, A.G. (1990). Time derivative models of Pavlovian reinforcement. In Learning and Computational Neuroscience: Foundations of Adaptive Networks, M. Gabriel and J. Moore, eds. (Cambridge, MA: MIT Press), pp. 539–602.