available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Short-term memory traces for action bias in human reinforcement learning

Rafal Bogacz[a,b,*], *Samuel M. McClure*[a,c,d], *Jian Li*[d], *Jonathan D. Cohen*[a,c], *P. Read Montague*[d]

[a]*Center for the Study of Brain, Mind and Behavior, Princeton University, Princeton, NJ 08544, USA*
[b]*Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA*
[c]*Department of Psychology, Princeton University, Princeton, NJ 08544, USA*
[d]*Department of Neuroscience, Human Neuroimaging Lab, Baylor College of Medicine, Houston, TX 77030, USA*

ARTICLE INFO

Article history:
Accepted 15 March 2007
Available online 24 March 2007

Keywords:
Reinforcement learning
Eligibility traces
Dopamine

ABSTRACT

Recent experimental and theoretical work on reinforcement learning has shed light on the neural bases of learning from rewards and punishments. One fundamental problem in reinforcement learning is the credit assignment problem, or how to properly assign credit to actions that lead to reward or punishment following a delay. Temporal difference learning solves this problem, but its efficiency can be significantly improved by the addition of eligibility traces (ET). In essence, ETs function as decaying memories of previous choices that are used to scale synaptic weight changes. It has been shown in theoretical studies that ETs spanning a number of actions may improve the performance of reinforcement learning. However, it remains an open question whether including ETs that persist over sequences of actions allows reinforcement learning models to better fit empirical data regarding the behaviors of humans and other animals. Here, we report an experiment in which human subjects performed a sequential economic decision game in which the long-term optimal strategy differed from the strategy that leads to the greatest short-term return. We demonstrate that human subjects' performance in the task is significantly affected by the time between choices in a surprising and seemingly counterintuitive way. However, this behavior is naturally explained by a temporal difference learning model which includes ETs persisting across actions. Furthermore, we review recent findings that suggest that short-term synaptic plasticity in dopamine neurons may provide a realistic biophysical mechanism for producing ETs that persist on a timescale consistent with behavioral observations.

1. Introduction

Rewards and punishments generally signify classes of stimuli that enhance and reduce reproductive fitness, respectively. Insofar as this is true, learning to maximize rewards and minimize punishments is a critical challenge faced by the nervous system. Recent experimental and theoretical work indicates that midbrain dopamine neurons may play a key role in this reinforcement learning problem. In particular, the firing rate of dopamine neurons has been proposed to encode

* *Corresponding author*. Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK. Fax: +44 117 954 5208.
E-mail address: R.Bogacz@bristol.ac.uk (R. Bogacz).

0006-8993/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2007.03.057

the difference between experienced reward and long-term predicted reward to generate a *prediction error* used for *temporal difference (TD)* learning (Dayan et al., 2000; Egelman et al., 1998; Montague and Berns, 2002; Montague et al., 1994, 1996; Montague and Sejnowski, 1994; Schultz et al., 1997). In order to learn which stimuli or actions may be associated with reward, dopamine-mediated prediction error signals are further suggested to guide changes in synaptic strength (Montague et al., 1996, 2006; Reynolds et al., 2001). Specifically, the synaptic weights associating an environmental situation with the value of an action are increased if dopamine level is above baseline levels (signaling positive reinforcement) and decreased if it is below baseline levels.

One fundamental problem in reinforcement learning is known as the credit assignment problem, which refers to the challenge of properly assigning credit to actions that lead to reward or punishment when these occur at varying times prior to the reinforcing event. A simple example arises in the game of chess, in which a particularly good or bad move may ensure victory of defeat several moves later. In order to properly value such a critical move, a reinforcement learning algorithm must properly assign credit backwards through time. With sufficient training experience, temporal difference learning can solve this problem, but its efficiency is substantially improved by the addition of *eligibility traces* (ET) (Barto et al., 1981; Sutton and Barto, 1998). In essence, eligibility traces function as decaying memories of previous choices that are used to scale synaptic weight changes. It has been suggested that ETs may be implemented in the brain by elevated levels of calcium that persist in dendritic spines (at synapses subject to learning) (Wickens and Kotter, 1995) or a relatively slow process in synapses triggered by a coincidence of pre-synaptic and post-synaptic spikes (Izhikevich, in press).

In the neuroscience literature, ETs have been proposed as a mechanism for dealing with small delays in reinforcement signal (Raymond and Lisberger, 1998; Wickens and Kotter, 1995). By contrast, computer algorithms implementing artificial reinforcement learning (e.g. for robot control) sometimes employ much longer-lasting ETs which persist across many actions and reinforcements. This allows the reinforcement to adjust the weights corresponding not only to the most recent action but also to previous actions in a recency-weighted manner (Sutton and Barto, 1998). Traces that span a number of actions may improve the performance of reinforcement algorithms if the task demands that a particular *sequence* of actions be performed to maximize overall rewards (e.g. Singh and Sutton, 1996; Sutton and Barto, 1998). However, to our knowledge, the question of whether reinforcement learning in humans and other animals involves ETs persisting over sequences of actions has not yet been investigated. Here we consider this question.

To test for ET-like mechanisms in human reinforcement learning, we examine the performance of human subjects in a sequential economic decision game named the *rising optimum* task (Egelman et al., 1998; Montague and Berns, 2002). Briefly, in the rising optimum task, subjects choose sequentially between two available actions by pressing one of two buttons: A or B. After each choice, a scale bar is updated to reflect the reward earned for that choice (Fig. 1A). Subjects are instructed to keep and maintain the bar at the highest possible level over

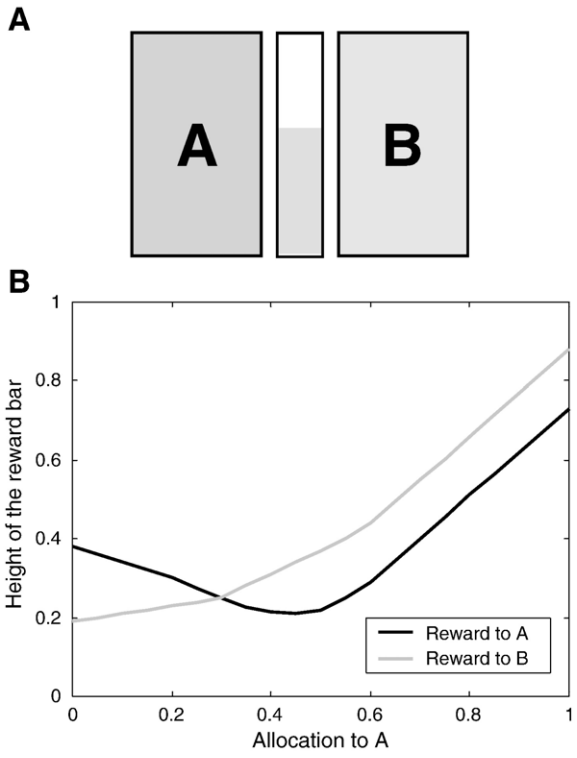

Fig. 1 – Rising optimum task. (A) Sample screen during experiment. The height of bar between the buttons indicates performance. (B) Dependence of the height of the reward bar on the last and previous decisions. *x*-axis shows the proportions of choices A in the last 20 choices, i.e., the allocation to A. *y*-axis shows the height of the reward bar following the decision. The black curve shows the height of reward bar after pressing button A for different allocations to A during the last 20 trials, and the gray curve shows the height of the reward bar after pressing button B. For example, if the subject had pressed equal number of A and B within the last 20 trials and the last choice was A, then the resulting height of the bar is 0.22—it can be read from panel B by looking at black line (because subject's last choice was A) for allocation to A of 0.5 (because the subject pressed A 50% of times within the last 20 trials).

the course of the experiment. The bar height following a choice depends on which button was selected and on the proportion of choices to button A over the previous 20 selections, as shown in Fig. 1B (see legend for details). We refer to the proportion of A choices in the last 20 decisions as *allocation to A*.

The optimal strategy in this task is to press button A on every choice (as shown in Fig. 1B: reward is greatest at the far right end of the curve). However, this strategy is not obvious to the subjects because for allocations to A higher than 0.3 (the point of the intersection of the curves in Fig. 1B), selecting B results in greater immediate increase in reward than selecting A. Continuing to select B will produce progressively lesser rewards, but these will still be greater than selecting A until the allocation to A falls below 0.3. At that point, selecting A will produce greater immediate reward. As a consequence, if subjects are driven primarily by concerns

about immediate reward, they will cycle around the crossing point in Fig. 1B.

In this task, the strategy that leads to the greatest short-term gains is a form of matching (Herrnstein, 1990) and conforms to predictions of standard reinforcement learning algorithms (i.e., lacking an eligibility trace). However, this is not the optimal strategy for maximizing long-term gains. The optimal strategy, which involves a sequence of locally sub-optimal moves, is unstable because of subjects' tendency for melioration (Herrnstein, 1982). Here, we show that the use of these two strategies – and in particular the discovery of the optimal one – is significantly affected by the time between choices in a surprising and counterintuitive way. This finding runs counter to predictions of standard reinforcement learning models. We demonstrate, however, that these findings can be explained naturally by augmenting the reinforcement learning algorithm with ETs that persist across actions. Finally, we review recent evidence and modeling work addressing the dynamics of dopamine release that support the hypothesis that this system may provide a realistic biophysical mechanism for producing persisting ETs.

2. Results

2.1. *Behavioral results*

Subjects were divided into 3 groups that differed only in the length of the forced delay between a choice and the start of the next trial (0 ms, 750 ms, 2000 ms), as described in Section 4. Mean inter-choice intervals (including response time and forced experimental delay) in the three delay conditions were 766 ms, 1761 ms, and 3240 ms respectively. Fig. 2A demonstrates the dependence of individual subjects' mean allocation to A on their mean inter-choice interval (in each of the three conditions). Linear regression analysis indicates a significant negative correlation between inter-choice interval and mean allocation to button A ($r \approx -0.23$, $p \approx 0.006$).

This pattern of behavior runs counter to observations about performance in most other behavioral settings. Typically, faster performance is accompanied by decrements in accuracy, in accord with a ubiquitously observed tradeoff between speed and accuracy. In this experiment, however, subjects performed *worse* when they were forced to respond slower. Specifically, subjects in the 2000 ms delay condition, who were given the greatest amount of time for decision and analysis of the task, rarely discovered the optimal strategy.

To more precisely quantify subjects' behavior with regard to strategy type, and how this evolved over the course of the experiment, we calculated the fraction of trials that followed the optimal and the matching strategies for each subject. Since the optimal strategy is to always choose A, we define the *optimizing fraction* simply as the fraction of choices A. The matching strategy involves choosing A when allocation to A is below 0.3 and choosing B if the allocation to A is above 0.3. We define the *matching fraction* as the proportion of choices that satisfy this strategy among the trials in which the allocation to A was not equal to 0.3 (under which condition the matching strategy is undefined). For each subject we determined the strategy that was followed on the greatest number of trials.

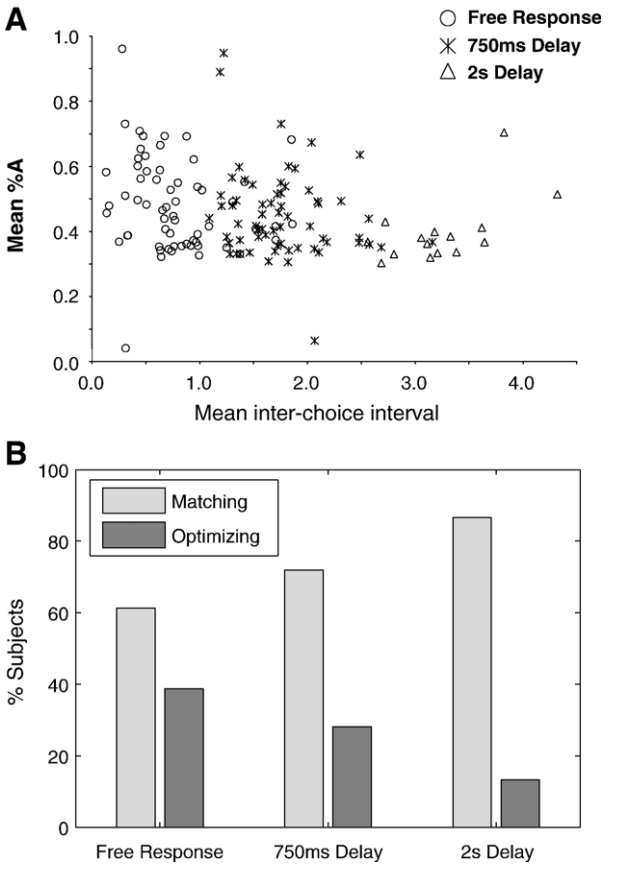

Fig. 2 – Dependence of mean allocation to A on the mean interval between choices in the rising optimum task. (A) Dependence of proportion of choices A during the whole experiment (*y*-axis) on mean interval between choices (*x*-axis). Each point corresponds to one subject, and shape of the point denotes the delay condition under which the subject ran the experiment (see legend in top–right corner of panel A). (B) Proportion of subjects with matching and optimal behavior on a greater proportion of trials in three different experimental conditions; the minimal delay between choices in each condition is indicated below the bars.

Fig. 2B shows the percentages of subjects preferring each of the strategies in the three delay conditions. As it demonstrates, the matching strategy starts to dominate as the length of the delay between continuous choices gets longer.

Fig. 3 illustrates how the subjects' strategies changed during the course of the experiment. Figs. 3A–C shows the optimizing and matching fractions for each subject in early (first 40 choices), middle (middle 40 choices) and late (final 45 choices) phases of the experiment. In all panels, the fractions are strongly anti-correlated, which results from the fact that for allocations to A above 0.3 the two strategies predict different choices. Subjects' strategies between the phases of the experiment are quite subtle. There is no evident tendency to switch from one strategy to another. Although the mean of optimizing fractions across subjects increases slightly from the early to the late phase (45.5% to 47.6% respectively), this difference is far from significant (Wilcoxon test, $p \approx 0.72$). How-

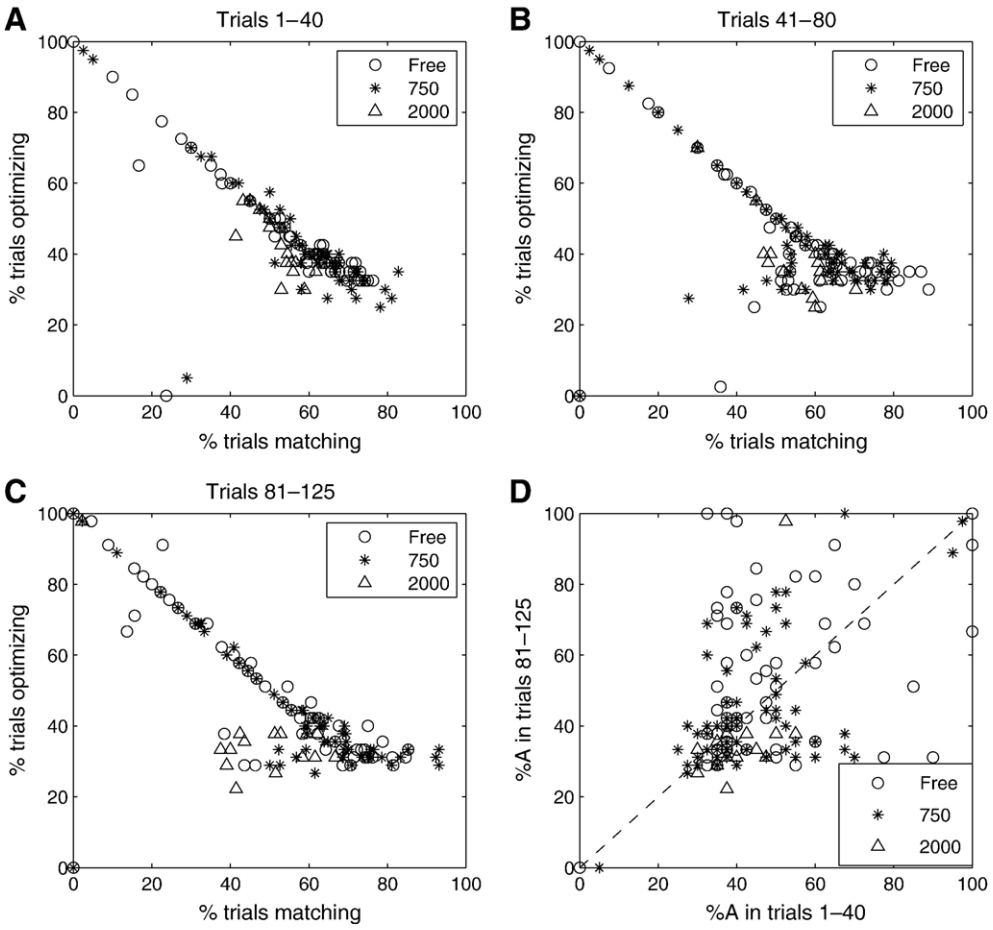

Fig. 3 – The changes in strategies employed by the subjects. Subjects studied in different delay conditions are shown by different symbols as indicated in the figure legend. (A–C) The panels correspond to the following phases of the experiment: trials 1–40, trials 41–80, and trials 81–125, respectively. Each panel plots the proportion of trials in a given phase in which a subject follows optimal strategy (*y*-axes) against the proportion of trials it follows the matching strategy (*x*-axes). (D) The proportion of trials a subject followed the optimal strategy in trials 81–125 plotted against the proportion of trials it followed this strategy in trials 1–41. The dashed line indicates the positions in the figure corresponding to equal proportions.

ever, the data do seem to suggest that subjects become more committed to particular strategies since, in the later phase (Fig. 3C), there are more subjects with high proportions (e.g. above 80% or 90%) of trials under a particular strategy than in the early phase (Fig. 3A). The commitment to strategies can be quantified by the variance of the optimizing fractions across subjects for the following reasons: the subjects committed to the optimal strategy will have a high value of the optimizing fraction, while the subjects committed to the matching strategy will have a low value of the optimizing fraction (due to the anti-correlation of the optimizing and matching fractions). Hence the more subjects are committed to their strategies, the more extreme values of optimizing fractions will be observed, and thus the higher the variance of the optimizing fraction across the subjects will be. As expected, the variance in the optimizing fractions in the late phase is significantly higher than in the early phase (*F*-test, $p<0.005$).

Figs. 3A–C also shows that there are a number of subjects who do not follow either the matching or optimal strategy, corresponding to the points in the bottom left quadrants of the panels. These subjects sometimes choose B also on trials in which allocation to A is below 0.3. Two factors influence the presence of this behavior: first, more subjects studied in 2 s delay condition show this behavior (the mean of the sum of matching and optimizing fractions based on all trials is significantly lower for subjects in 2 s condition than in the other conditions, *t*-test, $p<0.01$). Second, more subjects show this behavior in the late and middle than in the early phases on the experiment (the mean of the sum of matching and optimizing fractions based on trials 41–125 is lower than the one based on trials 1–40, paired *t*-test, $p<0.01$). We would like to come back to the interpretation of this result in the Discussion section.

Fig. 3D illustrates changes in the strategies of individual subjects between early and late phases of the experiment. There is a significant correlation between optimizing fractions in the two phases ($r\approx0.46$, $p<0.000001$), which implies that the subjects are more likely to adopt the same strategies in the two phases than to change them.

Finally, we tested two simple hypotheses regarding the reasons for why the subjects choose particular strategies. First, at shorter intervals subjects may simply have a higher

tendency to choose the same key on two consecutive trials, and thus some would have a higher allocation to A. This, in turn, may enable more frequent discovery of the optimal strategy. However, closer examination of the data suggests that this is not the case. There is no correlation between subjects' inter-choice interval and their proportion of repeated choices ($r=-0.005$; $p=0.95$).

Second, it might be reasoned that subjects who started out with a high allocation to A more frequently discovered the optimal strategy (as we describe in Section 4, allocation to A was initialized arbitrarily for each subject by randomly generating a 20-choice history). However, we found no evidence for this. The correlation between the optimizing fraction and the initial allocation to A across subjects is very weak ($r\approx0.11$) and is not significant ($p\approx0.18$). We further investigated whether the strategy evident at the beginning of the experiment depends on the initial allocation to A. Again, we found no evidence for this as, for any number of initial trials, the correlation between optimizing fraction and initial allocation to A is not significant.

In summary, we observed a surprising pattern of performance which decreases as the intervals between choices increase, as illustrated in Fig. 2. The performance did not change significantly between phases of experiment and was not influenced by the initial allocation to A. In the reminder of this section we focus on explaining the results of Fig. 2.

2.2. *Reinforcement learning models*

Here we consider how reinforcement learning models can explain the pattern of performance reported above. We begin by reviewing a previous model of learning in the rising optimum task (Egelman et al., 1998; Montague and Berns, 2002) that we refer to as the *standard* model. We then introduce an augmented form of this model that includes ETs and use this to address our empirical observations.

2.2.1. *Standard model*

Egelman et al. (1998) developed a reinforcement learning model of performance in the rising optimum task that was then used by Montague and Berns (2002) to account for matching behavior in the task. In this model the two possible actions are associated with weights w_A and w_B. Choices are made stochastically using a "softmax" decision rule with the probability of choosing A given by:

$$P(A)=\frac{1}{1+e^{\mu(w_B-w_A)}} \quad (1)$$

This decision rule is formally equivalent to the drift diffusion model (Ratcliff, 1978; Stone, 1960) in which the drift is equal to w_B-w_A, noise is equal to 2, and the threshold is equal to μ. The drift diffusion model implements the continuous version of the sequential probability ratio test (Laming, 1968), and hence is the optimal decision maker (Wald and Wolfowitz, 1948). This model has been used successfully to describe a wide corpus of empirical data regarding human performance in two alternative forced choice decision making tasks (Bogacz et al., 2006; Ratcliff and Smith, 2004; Ratcliff et al., 1999) and is consistent with recent observations of the neural dynamics associated with performance in such tasks (Gold and Shadlen, 2001, 2002; Ratcliff, 2006; Ratcliff et al., 2003; Schall, 2001; Shadlen and Newsome, 1996, 2001).

After each decision an error (δ) is calculated that is equal to the difference between the new height of the bar and the expected height of the bar. The expected height of the bar is taken simply as the weight of the decision just made w^*:

$$\delta=r-w^* \quad (2)$$

Subsequently the weight associated with the last choice is updated as follows:

$$w^*\leftarrow w^*+\lambda\delta \quad (3)$$

where λ denotes the learning rate.

Montague and Berns (2002) showed that the standard model predicts that only matching behavior should occur on the task (with $\lambda=0.93$) since weight modifications are based strictly on the value of the reward associated with the last choice. Hence, when allocation to A is greater than 0.3 and choosing B yields a higher reward than choosing A (see Fig. 1B), the model reinforces choosing B and the allocation to A decreases. Conversely, when allocation to A is less than 0.3 and choosing A yields higher reward than choosing B, the model reinforces choosing A so the allocation to A increases. Therefore, allocation to A converges to around 0.3 in the standard model. This reproduces the matching behavior predicted by the process of melioration proposed by Herrnstein (1982, 1990) to govern human decision making behavior. However, the standard model cannot readily explain the behavior of subjects who discover the optimal strategy[1], nor the dependence of this behavior on inter-choice delay, since there is no notion of time in the model.

2.2.2. *Eligibility trace model*

The effects of delay can be introduced into the standard model by including a mechanism for ETs (Sutton and Barto, 1998). To do so, we modified the model of Montague and Berns (2002) by allowing rewards to depend both on the last decision and on previous decisions weighted by an ET. The model stores a separate ET for each action (choosing A or choosing B) that we denote as e_A and e_B. The ET for each choice is updated each time a decision is made favoring that choice, and their values decay with time constant τ. Hence, e_A and e_B provide an exponentially decaying average of the frequency with which each decision is made, which are used to scale the amplitude of weight modification. The algorithm for updating choice weights is given as:

Initialize: $e_A=0$; $e_B=0$
Repeat
  Make choice and denote chosen button by *
  Observe reward r, and compute: $\delta=r-w^*$

[1] The standard model can also produce the optimal behavior, but only for particular parameter values: very high threshold μ (e.g. $\mu>100$) or a learning rate equal to $\lambda=2$. However, for these parameter values, the model generates sequential patterns of choice that differ substantially from those observed for human subjects.

Update ETs

$$e_A = e_A \exp(-\tau^{-1}T);\ e_B = e_B \exp(-\tau^{-1}T)$$
$$e^* = e^* + 1 \quad (4)$$

Update both weights

$$w_A \leftarrow w_A + \lambda\delta e_A \quad (5)$$

$$w_B \leftarrow w_B + \lambda\delta e_B \quad (6)$$

In Eq. (4), T denotes the time from the last update of the ET (i.e., from the previous choice). Note that if the decay rate of the ET is very rapid (τ is small), or the time between choices T is large, then the ET model reduces to the standard model. This is because, according to Eq. (4), both ETs decay to zero during the interval T. Accordingly, the weight modification is determined strictly by w^* since e^* is incremented to 1 before the weights are updated. Thus, for sufficiently long intervals between choices T, the ET model reduces to the standard model and produces matching behavior. This is consistent with the observation that subjects in our experiment exhibit a greater propensity for matching behavior as the delay is increased (i.e., for longer inter-choice intervals; see Fig. 2).

Conversely, when the inter-choice interval is short, then ETs provide a (fading) memory of recent choices. For example, if the last choice was B, but preceding choices were predominantly A, then the ET e_A may be larger than e_B. In this case, the positive reward received for choosing B will reinforce not only that decision, but also A (since e_A is elevated due to the previous decisions). Therefore if choosing A generates sufficiently consistent (and increasing) benefits, this will be progressively reinforced. This will be particularly true when the inter-choice time is short and more decisions contribute to the eligibility trace. This condition of progressively increasing reward occurs in the rising optimum task for allocations to A of greater than 0.3. Accordingly, the addition of ETs to the reinforcement learning algorithm can explain the observation that subjects can discover the optimal strategy and that they tend to do so more often in the shorter delay conditions.

2.2.3. *Simulation of the model*

To formally demonstrate the relationships described above, we simulated the ET model and compared it to empirical observations of performance in the rising optimum task. The model has 3 parameters: time constant for decay of eligibility trace τ, learning rate λ, and decision threshold μ. Although subject performance would likely be fit best by different parameter values of the model for different subjects, the available data did not provide sufficient power to fit the model individually for each subject. Therefore, we sought to fit a single set of parameter values to the distribution of subjects' behavior across delay conditions. Parameter values were fit using the maximum likelihood method. The behavior of each subject k was described by the mean allocation to A, A_k, and mean inter-choice interval, T_k, in each of the three delay conditions (i.e., the data shown in Fig. 2B). We sought parameter values that maximized the likelihood of the experimental data, namely:

$$\prod_{k=1}^{n} P_{\tau,\lambda,\mu}(A_k|T_k) \quad (7)$$

The prior probability $P_{\tau,\lambda,\mu}(A|T)$ in Eq. (7) was obtained from simulation as described in the legend for Fig. 3A. Parameter values maximizing this likelihood were found using the Simplex optimization algorithm (Nedler and Mead, 1965).

Fig. 4A compares the prior probability distribution obtained for best fitting parameters (maximizing likelihood) and the experimental distribution. Although a few subjects had mean allocation to A unlikely under the ET model, the model provides a good description of the main features of subjects' behavior. Specifically, matching behavior is most likely in the long delay condition, while for shorter delays both optimal and matching strategies are probable. The time constant of decay of ET maximizing likelihood was estimated as τ=4.89 s[2]. Fig. 4B compares histograms of the allocation to A of subjects in the three delay conditions with the corresponding simulations of the ET model. This also shows that, for longer delays, the matching behavior of the ET model is much more likely than for shorter delays.

2.3. *Neural bases for the ET*

Several neural structures have been shown to be sensitive to rewards (Breiter and Rosen, 1999; Olds, 1962; Rolls, 2000; Shizgal, 1999). We focus here on the midbrain dopamine system insofar as changes in the firing rate of these neurons parallel computationally derived reinforcement learning prediction error signals (Bayer and Glimcher, 2005; Montague et al., 1996; Schultz et al., 1997). In addition, dopamine is thought to modify synaptic weights in a manner necessary for the encoded prediction error signal to drive reward learning (Reynolds et al., 2001). Finally, dopaminergic function is critical for generating reward-directed actions (Berridge and Robinson, 1998; McClure et al., 2003). While the dopamine system may play only a part in generating reward-directed decisions, we review findings below that support the hypothesis that short-term plasticity in the synaptic terminals releasing dopamine provides a parsimonious and biophysically plausible account of decision-making and learning which closely matches the parameters of the model described above.

With the advance of electrochemical techniques, it is now possible to measure extracellular dopamine concentration with time resolution on the order of 100 ms. These measurements have shown that dopamine release from pre-synaptic terminals is subject to multiple forms of plasticity (Michael et al., 1987). These observations have been formalized in a series of dynamic equations that account for the fluctuations in dopamine release with three independent factors (Montague et al., 2004). Each factor changes with kinetics similar to the ET (Eq. (4)), eliciting a "kick" at the time of each action potential and exponential decay back to unity (as in Abbott and Nelson, 2000). Two of these factors served to depress subsequent dopamine release, while the third was facilitating. For the timescale corresponding to the delays between choices in the

[2] The optimizations have been started from 10 different starting points (chosen randomly), and in 6 cases converged to almost the same highest likelihood, and very similar parameters, suggesting the existence of global maximum; in particular, decay time constant of ET ranged $\tau \in$ [4.41, 5.52].

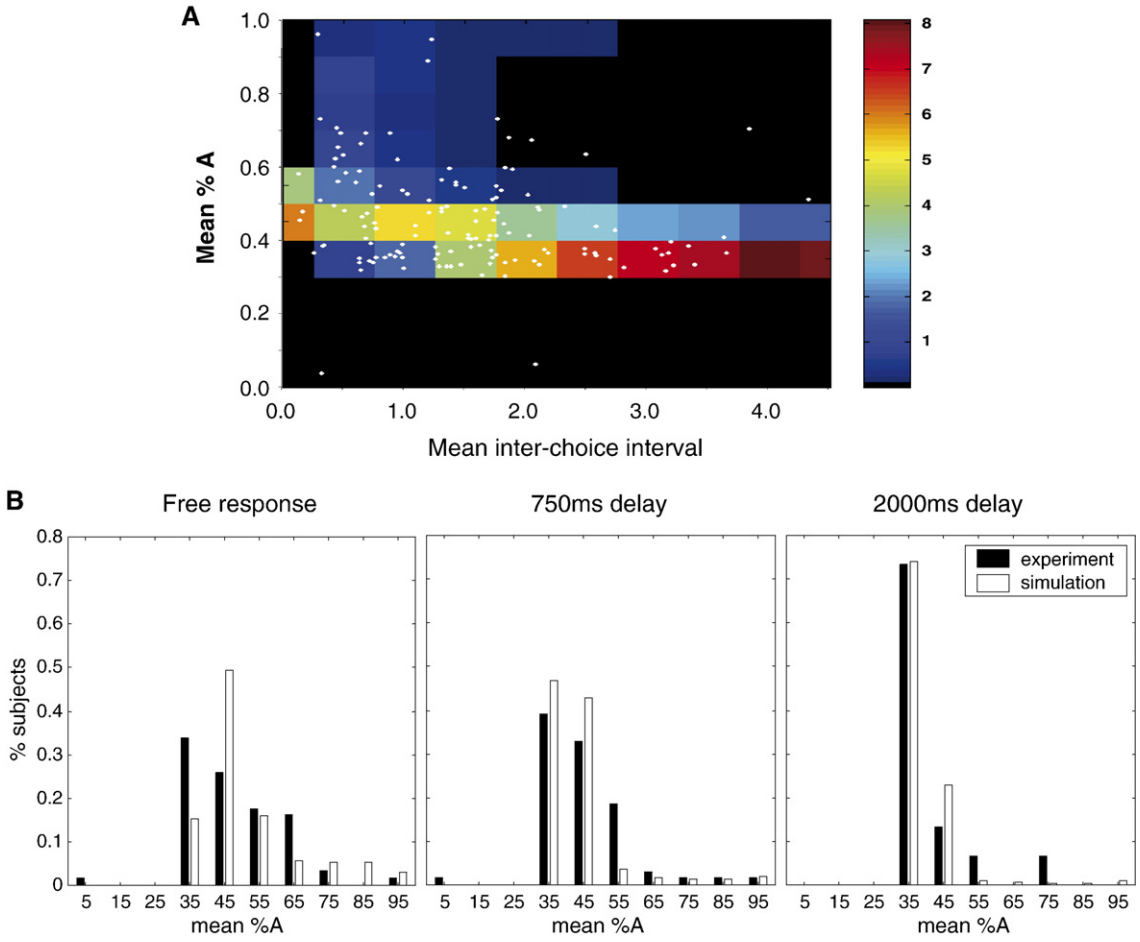

Fig. 4 – The comparison between allocation to A in the model and experiment. (A) Subjects' mean allocation to A (white dots) and the probability distribution of the mean allocation to A of the ET model (colors) for the parameters resulting in the maximum likelihood, namely: τ=4.89, λ=0.1675, μ=12.2. The probability distribution of ET model was found in the simulations in the following way. The inter-choice interval was discretized by dividing it into bins 0.5 s long. For a given set of model parameters, the probability distribution of mean allocation to A was estimated for the inter-choice intervals T corresponding to all centers of the bins. For each T, the ET model was simulated 1000 times performing the rising optimum task for 125 trials, and for each simulation the mean allocation to A was computed. The distribution of mean allocation to A was then obtained by building the histogram with 10 bins. Such procedure results in some bins of the histogram being empty and thus $P(A|T)=0$ for T and A corresponding to these bins. If any of the subjects' T_k and A_k fall into these bins, then the whole likelihood of Eq. (7) would be equal to 0. To avoid this problems the probability density $P(A|T)$ for empty bins is set up 0.01. (B) Histograms of allocation to A in three experimental conditions concerning minimal interval between choices for the subjects (black bars) and for the model (while bars). The model was simulated with the same parameters τ, λ, and μ as above. For each panel (i.e., simulated condition) the model was simulated (performing the rising optimum task for 125 trials) for the following number of times: one thousand multiplied by the number of subjects run in this condition. In each thousand of trials, T was equal to mean inter-choice interval of one of the subjects from the given condition.

rising optimum task (<1–2 s), facilitation is known to dominate changes in dopamine release (Montague et al., 2004). Furthermore, the time constant of this facilitation process was found to be roughly 4.5 s, which corresponds closely to the time constant of ET decay that produced the maximum likelihood fit to subjects' behavioral data.

To further explore how the dynamics of dopamine release may relate to ETs, we incorporated detailed mechanisms of dopamine release (using best fit parameters from Montague et al., 2004) into our model of the rising optimum task, and compared the dynamics of this dopamine signal with the ET in the model described above. In particular, dopamine release is proposed to drive the change in weights (w_A and w_B) between sensory inputs and corresponding motor output (Fig. 5A). Separate dopamine terminals modifying w_A and w_B were modeled to be independent but subject to synaptic plasticity with equal parameter values. Kicks to short-term facilitation and depression terms were modeled to occur only after choice of the corresponding action. Biologically, this action dependence of synaptic plasticity requires that the output of dopamine target structures (e.g. action selection in the striatum) be communicated to the dopamine terminals. This effect may be mediated by the production of nitric oxide (e.g. by striatal interneurons) which is known to influence DA neuron activity

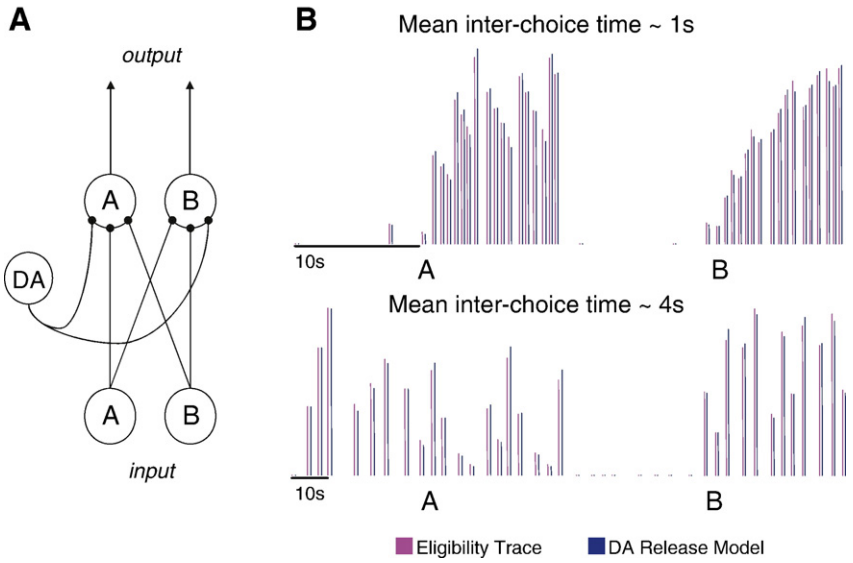

Fig. 5 – Comparison of dopamine release model and eligibility trace dynamics. (A) For each choice, choosing either A or B is determined according to the action values, w_A and w_B. After making a choice, the error δ between the actual reward received and expected reward (given by the corresponding w value) is used to update action values. This error signal is proposed to be carried by midbrain dopamine (DA) neurons. (B) Known short-term synaptic plasticity dynamics in dopamine release are sufficient to account for the model discussed in Fig. 4. Using the model developed by Montague et al. (2004, Eq. (2)), the relative amplitude of expected dopamine release was compared with the value of the eligibility traces at the time of each choice. The plots shown were generated from choices and choice times from one of the subjects in the experiment. Even though the release model incorporated more processes than the eligibility trace, for the time ranges over which subjects performed the task the two models are very similar (compare blue and magenta bars for short ITI, top, and long ITI, bottom). The amplitude of the eligibility trace values and release is in arbitrary units and was normalized to the same maximum value to demonstrate the similar changes in relative value.

(West and Grace, 2000). Driving both the ET model and the dopamine model with choices made by subjects produced signals with nearly equivalent dynamics (compare magenta and blue plots in Fig. 5B). These observations illustrate that the dynamics of dopamine release observed at the synaptic level correspond closely to those of the ET that provide the best account of behavioral observations.

3. Discussion

3.1. *Summary*

We have shown that a reinforcement learning model that incorporates a mechanism for ETs can account for human performance in the rising optimum task, with dynamics for the ET that closely match those observed for dopamine synaptic release in neurophysiological studies.

To achieve optimal performance in the rising optimum task, learning must take account of the recent history of actions, and not just the last one. Correspondingly, the ET model provides a mechanism for reinforcing actions based not only on the last action (as in standard reinforcement learning models), but on a decaying memory of previous actions—that is, an ET. If the intervals between actions are short, then more will fall within this window of memory and be subject to reinforcement learning. Conversely, if actions are more separated in time, then fewer will fall within the window of opportunity for reinforcement provided by the ETs and in the limit learning behavior will regress to that of the standard model, in which only the most recent action is reinforced. This dependence on timing provides an account for empirical observations: at longer delays, subjects were less likely to discover the optimal strategy, which depends on learning the greater value of a sequence of A choices over the most recent choice of B. Furthermore, fitting the model to the data provided an estimate of the time constant of decay on the ET – approximately 5 s – which corresponds closely with the time constant for facilitation of dopamine release estimated directly from measurements of synaptic concentrations of dopamine.

3.2. *Alternative accounts of behavior in the rising optimum task*

Other explanations of the behavioral data are in principle possible. In Fig. 2A, the mean allocation to A of subjects with short inter-choice intervals is close to 0.5, and random responding is expected to produce an optimizing fraction equal to 0.5. Hence we investigated if our results could be explained by the fact that subjects who respond faster simply respond more randomly. We give two arguments against this hypothesis. First, notice that if subjects responded completely randomly for 125 trials, then according to the central limit theorem, the optimizing fraction would have approximately normal distribution with mean 0.5 and variance 0.25/125=0.002. By contrast the optimizing fractions of subjects

studied in the free response condition have variance of 0.021 which is significantly higher (chi-square test, $p<0.0001$). Second, to quantify the randomness in subjects' response more precisely, we investigated how the decision threshold μ (described in Eq. (1); measure of how deterministic the choices are) changes as a function of inter-choice interval. We fitted four separate ET models, each one to the data from a group of subjects with one of the following inter-choice intervals: 0–1 s, 1–2 s, 2–3 s, and 3–4 s, using the method of Section 2.2.3. The estimated values of the decision threshold μ for each interval were 15.8, 13.9, 12.4, and 13.5, respectively, and did not show significant increasing or decreasing trend (correlation between the values of the threshold and centers of the inter-choice intervals for which they were estimated was not significant, $p>0.2$).

Alternatively, it is possible that subjects who discover the optimal strategy rely on more sophisticated cognitive processes than standard reinforcement learning supposes. For example, there have been recent efforts to extend reinforcement learning theories to include model-based mechanisms, thought to rely on higher level cortical systems (such as prefrontal cortex) that can support a richer representation of the state space and can adapt flexibly to changes in the environment (e.g. Daw et al., 2005). However, it seems unlikely that such mechanisms could account for our findings. Typically, model-based mechanisms represent a tradeoff of greater flexibility at the cost of greater computational complexity, which is typically assumed to involve greater time for each decision. This runs counter to our finding that subjects were more likely to discover the optimal strategy when they had *less*, not *more* time for each decision.

In the rising optimum task, the engagement of the model-based mechanisms might still be possible in the 2 s delay condition, when subjects had most time to consider the alternative choices, and in the later phase of the experiment, when they were more familiar with the task and could allocate more attentional resources to modeling the task. Analysis of Figs. 3A–C in Section 2.1 is consistent with a presence of such an engagement. Recall that there are a number of subjects who do not follow either the matching or optimal strategy as they sometimes choose B also on trials in which allocation to A is below 0.3. Such behavior might be a result of testing some higher order models. As shown in Section 2.1, this behavior is present in the same conditions when the engagement of the model-based mechanisms could be expected, i.e., in 2 s delay condition and in the later phases on the experiment.

Nevertheless, to adjudicate satisfactorily between these hypotheses, it is necessary to develop a model of the task involving model-based learning mechanisms and then identify quantitative predictions regarding behavioral and/or neural measurements that differ between this and the ET model. Of course, it is also possible (even likely) that both types of mechanism contribute to human performance.

Experimental evidence indicates that, in other gambling tasks, the prefrontal cortex initiates control process allowing the subjects to make a choice leading to a lower immediate payoff, for example: (i) lesions to prefrontal cortex are known to impair performance in gambling tasks (Manes et al., 2002). (ii) Certain regions of frontal cortex are more activated on the trials in which subjects make exploratory choices often leading to lower immediate reward (Daw et al., 2006). It has been shown also that the prefrontal cortex is activated when a significant change to the policy is required (Li et al., 2006). Thus, one could predict that increased activity in the prefrontal cortex may occur in our experiment when subjects employ model-based decision mechanisms in 2 s delay condition.

3.3. *Individual differences*

While comparing the ET model with the experimental data, we fit a model with a single set of parameters to the data from all subjects. Hence the variability in behavior for shorter inter-choice intervals in simulations shown in Fig. 4 comes only from the stochastic decision rule of Eq. (1). As we mentioned in Section 2.2.3, it is likely that the variability of subjects' behavior also reflects individual differences in the parameters of their reinforcement learning systems. Indication for such an influence has been provided by Montague and Berns (2002). Subjects who performed the rising optimum task were also asked to participate in an experiment in which brain activity was monitored during delivery of sequences of fruit juice and water. In this experiment, the sequences were predictable on some blocks and unpredictable on others. It is known that the striatum is more activated by unpredictable than predictable sequences (Berns et al., 2001). Most importantly, this 'predictability' difference was higher in the subjects who used the optimal strategy in the rising optimum task than subjects who used the matching strategy. This may indicate that individual differences in the characteristics of striatal responses across subjects may influence their choice of strategy in the rising optimum task.

One could also ask if the difference in strategies chosen can be influenced by the subjects' discount function for future reward (independently of the decay rate of the eligibility trace). If this were the case, the subjects heavily discounting future rewards may choose the strategy resulting in higher immediate reward (i.e., matching), while the other subjects may choose the strategy resulting in long-term pay-off (i.e., optimal). It is not possible to test this hypothesis within our model because it does not include the discount factor parameter. To verify this hypothesis, one could ask the subjects who performed the rising optimum task to also perform a task allowing estimation of subjects' discount factor (e.g. Benzion et al., 1989).

3.4. *Conclusions*

Previous modeling work of reward learning and decision-making has made profitable use of the simplest possible reinforcement learning algorithm. While these models accurately predicted behavior in a number of simple economic games (Egelman et al., 1998), they are unable to account for optimal behavior in tasks where matching is not the optimal strategy (e.g. rising optimum task considered here). Here, we have demonstrated how performance depends on inter-choice interval in a principled way (Sutton and Barto, 1998) through the use of ET mechanisms. The success of the current model suggests that in humans reinforcement learning operates over a (decaying) memory of recent actions and not just the last

action performed. Furthermore, the time constant of decay that we estimated for this memory corresponds closely to the time constant of short-term facilitation and depression of dopamine release in rapid electrochemical measurement made in freely moving animals (Montague et al., 2004). This correspondence is made more provocative given the growing body of evidence that dopamine release mediates the reinforcement learning signal at the synaptic level (Pan et al., 2005; Reynolds and Wickens, 2002; Wickens and Kotter, 1995; Wickens et al., 2003). Together, these findings suggest an increasingly detailed and formally explicit view of how reinforcement learning is implemented in the brain—a view that can serve to guide new empirical inquiry at both the neural and behavioral levels.

4. Experimental procedures

The experimental task is briefly described in the Introduction section. A more detailed description of the experiment is given elsewhere (Egelman et al., 1998; Montague and Berns, 2002). A simulator of this task is available at: www.cs.bris.ac.uk/home/rafal/rl/.

Each subject performed 125 choices during the task. Allocation to A was initialized arbitrarily for each subject by randomly generating a 20-choice history. This produced a mean initial allocation to A across subjects of 0.5, with a standard deviation of 0.11. The experiment was conducted under three different conditions which differed only in the pace at which subjects were allowed to execute choices: subjects in the *free response* condition responded as rapidly as they wished ($n=62$ subjects); in the *750 ms delay* condition, a minimum delay of 750 ms was imposed following each choice, before the next could be executed ($n=64$); and in the *2 s delay* condition a 2 s delay was imposed ($n=15$). Delays between choices were enforced after each decision by disabling the buttons, which was indicated to subjects by changing the button color to gray.

Acknowledgments

This work was supported by NIMH grant P50 MH62196 (J.D.C), Kane Family Foundation (P.R.M.), NINDS grant NS-045790 (P.R.M.), NIDA grant DA-11723 (P.R.M.), NIMH grant F32 MH072141 (S.M.M.), and EPSRC grant EP/C514416/1 (R.B.).

REFERENCES

Abbott, L.F., Nelson, S.B., 2000. Synaptic plasticity: taming the beast. Nat. Neurosci. Suppl. 3, 1178–1183.

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

Bayer, H.M., Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141.

Benzion, U., Rapoport, A., Yagil, J., 1989. Discount rates inferred from decisions: an experimental study. Manag. Sci. 35, 270–284.

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

Berridge, K.C., Robinson, T.E., 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res. Rev. 28, 309–369.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J.D., 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks. Psychol. Rev. 113, 700–765.

Breiter, H.C., Rosen, B.R., 1999. Functional magnetic resonance imaging of brain reward circuitry in the human. Ann. N. Y. Acad. Sci. 877, 523–547.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8, 1704–1711.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 44, 876–879.

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

Egelman, D.M., Person, C., Montague, P.R., 1998. A computational role for dopamine delivery in human decision-making. J. Cogn. Neurosci. 10, 623–630.

Gold, J.I., Shadlen, M.N., 2001. Neural computations that underlie decisions about sensory stimuli. Trends Cogn. Sci. 5, 10–16.

Gold, J.I., Shadlen, M.N., 2002. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions and reward. Neuron 36, 299–308.

Herrnstein, R.J., 1982. Melioration as behavioral dynamism. In: Commons, M.L., Herrnstein, R.J., Rachlin, H. (Eds.), Quantitative Analyses of Behavior. Matching and Maximizing Accounts, vol. II. Ballinger Publishing Co, Cambridge, MA. p.^pp.

Herrnstein, R.J., 1990. Rational choice theory: necessary but not sufficient. Am. Psychol. 45, 356–367.

Izhikevich, E.M., in press. Solving the distal reward problem through linkage of STDP and dopamine signaling, Cereb. Cortex (doi:10.1093/cercor/bhl152).

Laming, D.R.J., 1968. Information Theory of Choice Reaction Time. Wiley, New York.

Li, J., McClure, S.M., King-Casas, B., Montague, P.R., 2006. Policy adjustment in a dynamic economic game. PLoS ONE 1, e103.

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., Robbins, T., 2002. Decision-making processes following damage to the prefrontal cortex. Brain 125, 624–639.

McClure, S.M., Daw, N.D., Montague, P.R., 2003. A computational substrate for incentive salience. Trends Neurosci. 26, 423–428.

Michael, A.C., Ikeda, M., Justice Jr., J.B., 1987. Mechanisms contributing to the recovery of striatal releasable dopamine following MFB stimulation. Brain Res. 421, 325–335.

Montague, P.R., Berns, G.S., 2002. Neural economics and the biological substrates of valuation. Neuron 36, 265–284.

Montague, P.R., Sejnowski, T.J., 1994. The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. Learn. Mem. 1–13.

Montague, P.R., Dayan, P., Sejnowski, T.J., 1994. Foraging in an uncertain environment using predictive Hebbian learning. Adv. Neural Inf. Process. Syst. 6, 598–605.

Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.

Montague, P.R., McClure, S.M., Baldwin, P.R., Phillips, P.E.M., Budygin, E.A., Stuber, G.D., Kilpatrick, M.R., Wightman, R.M., 2004. Plasticity in neuromodulator release: dynamic control of dopamine delivery in freely moving animals. J. Neurosci. 24, 1754–1759.

Montague, P.R., King-Casas, B., Cohen, J.D., 2006. Imaging valuation models in human choice. Annu. Rev. Neurosci. 29, 417–448.
Nedler, J.A., Mead, R., 1965. A simple method for function minimization. Comput. J. 7, 308–313.
Olds, J., 1962. Hypothalamic substrates of reward. Psychol. Rev. 42, 554–604.
Pan, W.X., Schmidt, R., Wickens, J.R., Hyland, B.I., 2005. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci. 25, 6235–6242.
Ratcliff, R., 1978. A theory of memory retrieval. Psychol. Rev. 83, 59–108.
Ratcliff, R., 2006. Modeling response signal and response time data. Cogn. Psychol. 53, 195–237.
Ratcliff, R., Smith, P.L., 2004. A comparison of sequential sampling models for two-choice reaction time. Psychol. Rev. 111, 333–367.
Ratcliff, R., Van Zandt, T., McKoon, G., 1999. Connectionist and diffusion models of reaction time. Psychol. Rev. 106, 261–300.
Ratcliff, R., Cherian, A., Segraves, M., 2003. A comparison of macaques behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. J. Neurophysiol. 90, 1392–1407.
Raymond, J.L., Lisberger, S.G., 1998. Neural learning rules for vestibulo-ocular reflex. J. Neurosci. 18, 9112–9129.
Reynolds, J.N., Wickens, J., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15, 507–521.
Reynolds, J.N., Hyland, B.I., Wickens, J.R., 2001. A cellular mechanism of reward-related learning. Nature 413, 67–70.
Rolls, E.T., 2000. The orbitofrontal cortex and reward. Cereb. Cortex 10, 284–294.
Schall, J.D., 2001. Neural basis of deciding, choosing and acting. Nat. Rev., Neurosci. 2, 33–42.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Shadlen, M.N., Newsome, W.T., 1996. Motion perception: seeing and deciding. Proc. Natl. Acad. Sci. U. S. A. 93, 628–633.
Shadlen, M.N., Newsome, W.T., 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. 86, 1916–1936.
Shizgal, P., 1999. On the neural computation of utility: implications from studies of brain stimulation reward. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Foundations of Hedonic Psychology: Scientific Perspectives on Enjoyment and Suffering, Russell Sage Foundation.
Singh, S.P., Sutton, R.S., 1996. Reinforcement learning with replacing eligibility traces. Mach. Learn. 22, 123–158.
Stone, M., 1960. Models for choice reaction time. Psychometrika 25, 251–260.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. MIT Press, Cambridge, MA.
Wald, A., Wolfowitz, J., 1948. Optimum character of the sequential probability ratio test. Ann. Math. Stat. 19, 326–333.
West, A.R., Grace, A.A., 2000. Striatal nitric oxide signaling regulates the neuronal activity of midbrain dopamine neurons in vivo. J. Neurophysiol. 83, 1796–1808.
Wickens, J., Kotter, R., 1995. Cellular models of reinforcement. In: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in Basal Ganglia. MIT Press, Cambridge, MA. pp. 187–214.
Wickens, J.R., Reynolds, J.N., Hyland, B.I., 2003. Neural mechanisms of reward-related motor learning. Curr. Opin. Neurobiol. 13, 685–690.