EJN fens EUROPEAN JOURNAL OF NEUROSCIENCE

European Journal of Neuroscience, Vol. 35, pp. 1152–1168, 2012 doi:10.1111/j.1460-9568.2012.08026.x

Instrumental vigour in punishment and reward

Peter Dayan
Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London WC1N 3AR, UK

Keywords: dopamine, reinforcement learning, safety, serotonin, two-factor theory

Abstract

Recent notions about the vigour of responding in operant conditioning suggest that the long-run average rate of reward should control the alacrity of action in cases in which the actual cost of speed is balanced against the opportunity cost of sloth. The average reward rate is suggested as being reported by tonic activity in the dopamine system and thereby influencing all actions, including ones that do not themselves lead directly to the rewards. This idea is syntactically problematical for the case of punishment. Here, we broaden the scope of the original suggestion, providing a two-factor analysis of obviated punishment in a variety of operant circumstances. We also consider the effects of stochastically successful actions, which turn out to differ rather markedly between appetitive and aversive cases. Finally, we study how to fit these ideas into nascent treatments that extend concepts of opponency between dopamine and serotonin from valence to invigoration.

Introduction

Until recently, neural reinforcement learning (Daw & Doya, 2006) focused its efforts on understanding which of a collection of actions would be performed, rather than when or how vigorously this should happen. Niv *et al.* (2005, 2007); Niv (2008) pointed out that this overlooked a wealth of standard experimental paradigms in which a measure of vigour such as the rate of lever pressing is assessed. They captured the essence of the control of vigour by considering a decision problem in which subjects choose the latency of their actions to optimize the balance between a direct cost of acting too quickly and an opportunity cost of acting too slowly, with the latter arising from the delay to the next and all subsequent rewards. In this model, the optimal latency for an action is inversely related to the square root of the overall reward rate. Thus, higher reward rates (stemming, for a motivational example, from the higher value that a thirsty subject might accord water reinforcements) lead to faster responding.

However, consider experiments on active avoidance. At best the reward rate is zero; at worst, it could be significantly negative. Thus, it is clear that the mathematical relationship suggested in the work on appetitive vigour cannot be valid. Here, we generalize the underlying decision problem to consider reinforcers that might be delivered even if the subject fails to perform an action in time, rather than succeeding in doing so (Cools *et al.*, 2011). It turns out that this can also model cases where an actual reward is only delivered if a particular action is executed sufficiently quickly.

For the case of active avoidance, a standard treatment involves what are known as 'two factors' (Mowrer, 1947; Johnson *et al.*, 2002; Moutoussis *et al.*, 2008; Maia, 2010). In Reinforcement Learning (RL) terms, subjects are assumed to learn to predict and thus fear the looming shock (one, purportedly Pavlovian, factor), so that a transition from an unsafe to a safe state provides an appetitive prediction error that can reinforce the associated action (the other, instrumental, factor). We construct a form of two-factor theory of active avoidance, in which the benefits of safety act as surrogate rewards, and study the parallels and differences between forms of punishment avoidance and reward collection.

Based partly on evidence from selective lesions to regions of the dopamine system (Salamone & Correa, 2002), (Niv *et al.*, 2007) suggested that it was the tonic activity of dopamine neurons that would report the long-run average reward rate and duly exert an effect on vigour. This would complement the effect of phasic dopamine activity as a temporal difference prediction error for reward, which has been suggested as controlling the learning of predictions and actions (Montague *et al.*, 1996; Schultz *et al.*, 1997). Further, appetitive Pavlovian–instrumental transfer (Estes, 1943; Rescorla & Solomon, 1967; Lovibond, 1981, 1983; Dickinson & Balleine, 1994, 2002), in which cues predicting future rewards boost the vigour of instrumentally controlled actions (even ones aimed at different rewards), is also known to be influenced by dopamine (Smith & Dickinson, 1998; Ikemoto & Panksepp, 1999; Dickinson *et al.*, 2000; Wyvell & Berridge, 2001; Lex & Hauber, 2008). Obversely, there is evidence that serotonin suppresses vigour (i.e. boosts inhibition) in the context of predictions of punishment (Deakin, 1983; Soubrié, 1986; Deakin & Graeff, 1991; Graeff, 2002; Cools *et al.*, 2008; Crockett *et al.*, 2009; Tops *et al.*, 2009).

Thus, recent work looking at the influence of appetitive and aversive Pavlovian predictions on instrumental choice (Williams & Williams, 1969; Dayan *et al.*, 2006a; Crockett *et al.*, 2009; Guitart-Masip *et al.*, 2011b) has been undertaken in the terms of functional opponency between dopamine and serotonin (Daw *et al.*, 2002; Dayan & Huys, 2009) controlling dual dimensions of valence and vigour (Boureau & Dayan, 2011; Cools *et al.*, 2011). Our two-factor model of instrumental vigour in the case of punishment was constructed in the context of these emerging analyses.

Correspondence: Peter Dayan, as above.
E-mail: dayan@gatsby.ucl.ac.uk

Received 21 October 2011, revised 23 December 2011, accepted 5 June 2012

Some of the paradigms that we model resemble fixed and variable interval schedule tasks. However, the quantitative predictions about the times at which the lever will be pressed are probably not correct – we do not consider, for instance, the possibility that subjects are choosing the time at which they switch between a slow and a fast rate of lever pressing (Gallistel *et al.*, 2004; Daw & Courville, 2008) or other latent states of pressing (Eldar *et al.*, 2011). Furthermore, we study the properties of optimal behaviour, without considering the effect of approximations that might emerge from limitations to neurobiological mechanisms of representation and control. However, the relationships between costs and optimal latencies may still be revealing about both behavioural and neural findings. We construct the prediction errors that would arise from optimal choices of vigour, and then ascribe these errors to opponent neuromodulatory systems.

Methods

Figure 1A and B depicts two semi-Markov decision processes (SMDPs) (Puterman, 2005), associated, respectively, with reward and punishment (generically described as 'outcomes'). Of course, rewards and punishments can also be combined – we discuss the implications of this later.

Reward

The characteristic reward SMDP (Fig. 1A) involves a process that arms a lever at time T drawn from density $p_T(T)$. The first lever press after time T causes an outcome of (positive) utility z to be delivered. For clarity, we refer to utilities throughout: the convention is that rewards have positive utility, and both punishments and costs negative utility. Earlier lever presses lead to no return, but merely negative utility associated with their latencies τ. Following acquisition of the outcome, there is a fixed inter-trial interval τ_I. Figure 1A shows two separate trials: in the first, the subject pressed twice to get the outcome, as the first press came before the lever had been armed; in the second, only a single press was necessary. This is an SMDP, as choices are only made at discrete points, but time evolves continuously. Thus when the subject chooses a latency τ, this much time elapses before the process arrives at the next state.

The state structure of the task turns out to be a little complicated. We start from the case that pressing the lever always leads to the outcome if it is armed (i.e. success probability $v = 1$). There are two macroscopic states, which we label 1 and 2, corresponding respectively to pre-outcome and post-outcome. Depending on the arming distribution $p_T(T)$, it may also be important to capture the time s since state 1 was entered. Thus, we label the unarmed state as the pair $\{1, s\}$. Importantly, we assume that subjects know when state {1,0} is entered (say, by a visual cue) and when they enter state 2, by receiving a reward (or, in the case of punishment, a negative outcome or a signal that they are now safe); but they do not necessarily know the arming time T.

We follow (Niv *et al.*, 2007)'s hyperbolic cost structure for lever pressing. A press at latency τ has utility $\frac{a}{\tau} + b$, where $b \leq 0$ is a unit utility for the press, and $a \leq 0$ is a factor that determines the magnitude of the hyperbolic dependence on τ. Note that (Niv *et al.*, 2007) used $-a$. Their proposal was originally made for convenience; it was not based on an assessment of, for instance, the true economic cost of acting quickly.

The case that (Niv *et al.*, 2007) studied is roughly equivalent to setting $T = 0$ (i.e. arming the lever immediately), and then choosing the latency of the first lever press τ_1 to optimize the average rate of utility (i.e. reward minus the sum of punishment and cost) per unit time. This choice is sometimes called a policy. The whole task can then be characterized as a Markov decision problem, where the only choice of action (i.e. a choice of latency) is at state {1,0}, without any further choice at the post-outcome state (state 2). The theory of average utility rate optimization for Markov decision problems (Puterman, 2005) tells us to choose the latency to maximize a quantity known as the optimal differential Q-value of that latency at the start state, so $\tau_1^* = \operatorname{argmax}_\tau\{Q*(\{1,0\},\tau)\}$, where

$$Q^*(\{1,0\},\tau) = \frac{a}{\tau} + b + z - \rho^*\tau + V^*(2). \qquad (1)$$

Here, $V^*(2)$ is the differential value of state 2 given the optimal overall policy (the asterisks are used to imply optimality). We use $V^*(2)$ rather than $Q^*(2)$, as there is no choice of action at state 2. Differential Q-values differ from conventional Q-values (Watkins, 1989) in that they are only meaningful relative to each other [so one can arbitrarily be set to 0; we choose to set $V^*(2) = 0$ to satisfy some technical conditions] and that they take explicit account of the passage of time (via the term $\rho^*\tau$), rather than have this be only implicit in a discount factor. Further, ρ^* is the optimal average rate of utility per unit time, defined as

$$\rho^* = \lim_{t\to\infty} \varepsilon\left[\frac{1}{t}\left(n_t z + \sum_{i=1}^{\omega_t}\left(\frac{a}{\tau_i^*} + b\right)\right)\right] \qquad (2)$$

where n_t is the number of outcomes accrued in time t, ω_t is the number of presses attempted, and τ_i^* are the latencies of all those presses. That ρ^* is optimal depends on the latencies being chosen optimally. ρ^* thus depends on the policy as well as the utility z; it obviously also depends on the costs of acting.

The two forces controlling the optimal latency in Eqn 1 are the explicit contribution to the utility of acting quickly (a/τ) and the opportunity cost of acting slowly ($-\rho^*\tau$). The rationale for the latter is that, by dedicating time τ to the press, an average total utility worth $\rho^*\tau$ is missed. Optimizing $Q^*(\{1,0\},\tau)$ with respect to τ (Niv *et al.*, 2007) revealed that

$$\tau^* = \sqrt{\frac{-a}{\rho^*}} \qquad (3)$$

showing, as mentioned above, that the optimal latency decreases as the average utility rate ρ^* goes up. If, for instance, the subject is made thirsty, the immediate utility z of a drop of juice increases, so ρ^* increases, and the latency of the press decreases.

Of course, ρ^* also depends on τ^*, in this case as

$$\rho^* = \frac{z + \frac{a}{\tau^*} + b}{\tau^* + \tau_I} \qquad (4)$$

making the problem recursive. There are various techniques for finding the optimal solution (Mahadevan, 1996; Puterman, 2005). For this article, we concentrate on properties of the optimal solution; this is not to say that realistic methods for solving for the optimal policy, or its near approximations are not also of great interest. In order to compute the solutions, we discretize continuous variables such as the latency τ, and thus solve a finite state approximation to the full, continuous, problem (and find maxima rather than suprema).

In the case (as in variable interval schedules) that the arming time T is stochastic, drawn from a density $p_T(T)$, Eqn 1 changes. First, it is necessary to take account of the possibility that more than one lever press will be necessary (as in the left trial of Fig. 1A). This means that

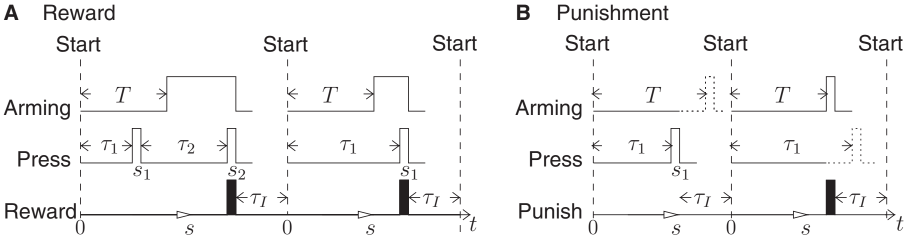

FIG. 1. SMDPs associated with reward (A) and punishment avoidance (B). (A) A lever is armed according to distribution $P_T(T)$ such that the first successful lever press after time T leads to a reward (an outcome with positive utility; bottom black bar). There is then a fixed inter-trial interval of length τ_I, after which the next trial starts. Both pre-outcome (state 1) and post-outcome (state 2) states are explicitly signalled. Here, the first lever press in the first trial did not lead to the outcome, as the lever was not armed; the second press was productive. In the second trial, the subject achieved the outcome from the first time that the lever was pressed. Subjects choose the latency τ of each lever press (which need not be the same, and can even depend on the time s within each trial). The overall passage of time is captured by t. Lever pressing may be only stochastically successful – with probability v. (B) A punishment (i.e. an outcome with negative utility) is scheduled for time T [drawn from $P_T(T)$] and will be delivered unless a lever press precedes it. The first punishment is obviated (dotted line on the arming process) by the earlier lever press; the second punishment is delivered because the lever press latency τ exceeded the arming time T. There is again a fixed inter-trial interval τ_I.

there are differential Q-values $Q(\{1, s\}, \tau)$ for latencies τ starting at states $\{1, s\}$ with $s > 0$. Second, such values become expectations over the future. There are two possibilities – either $T > (s + \tau)$, which means that the press will fail (as the lever is not armed by the time the press finishes), and the problem will then continue from state $\{1, s + \tau\}$; or $T \leq (s + \tau)$, in which case the press will succeed, the reward will be delivered, and a transition to state 2 will occur. In both cases, the probabilities of these happening are conditional on $T > s$, as considering $Q(\{1, s\},\tau)$ at state $\{1, s\}$ implies that pressing prior to s has been unsuccessful. Note that, for suitably regular distributions, the slope of this conditional distribution at $\tau = 0$, namely $\lim_{d\tau\to 0}\left\{\frac{P_T(d\tau+s>T|T>s)}{d\tau}\right\}$, is known as the hazard function. Thus

$$Q^*(\{1,s\},\tau) = \frac{a}{\tau} + b - \rho^*\tau + P_T(T > (s+\tau)|T > s)V^*(\{1,s+\tau\}) + P_T(T \leq (s+\tau)|T > s)(z + V^*(2)) \qquad (5)$$

where$V^*(\{1,s+\tau\}) = \max_{\tau'}\{Q^*(\{1,s+\tau\},\tau')\}$ quantifies the value of state $\{1, s + \tau\}$ on the basis of the best possible latency that could be chosen there. The presence of the term involving this quantity makes the problem as a whole recursive, but again solvable by the various dynamic programming methods mentioned above.

The last case for reward concerns the effect of setting the probability that a press succeeds even if the lever is armed, to be $v < 1$. The problem changes character rather dramatically, as the state in the world (i.e. whether or not the lever is armed) is not known by the subject whereas, by contrast, for $v = 1$, the subject always knows that the lever is not armed when choosing τ. The subject just has information about arming from the passage of time s, and from observations of past unsuccessful attempts to press (the lack of success is discovered at times $s_1 = \tau_1$, $s_2 = \tau_1 + \tau_2$, $s_3 = \tau_1 + \tau_2 + \tau_3$,...). Formally, the problem becomes a partially observable SMDP; see (Daw *et al.*, 2006) for a similar issue in the case of prediction without control.

In brief, the solution to this partially observable SMDP involves augmenting state 1 with a belief $\beta \in [0, 1]$ that the lever is armed. The evolution of β goes as follows. First, β starts off at 0 at the moment of transition from the post-outcome state (2) back to the pre-outcome state (1), as the subject knows that the lever is not initially armed. Consider the first lever press τ_1. If this succeeds, then the outcome is delivered, and the state makes a transition to the post-outcome state (2), at which β is irrelevant. If, however, the lever press does not lead to reward, then this is either because the lever has not yet been armed, which has probability $P_T(T > \tau_1)$, or because the lever has been armed [which has probability $P_T(T \leq \tau_1)$], but the press failed (which has probability $1-v$). Thus, the belief in the latter option [$\beta(\tau_1)$] becomes

$$\beta(\tau_1) = \frac{(1-v)P_T(T \leq \tau_1)}{(1-v)P_T(T \leq \tau_1) + P_T(T > \tau_1)} \qquad (6)$$

Now, imagine that the subject is at state $\beta(s)$ after some time s without receiving an outcome in a trial, then executes a press with latency τ, and still does not receive the outcome. Then, by a similar process of reasoning, its new belief that the lever is armed, which, for convenience, we write as $\beta(s, \tau)$, becomes

$$\beta(s,\tau) = \beta(s) + (1-\beta(s))\frac{(1-v)P_T(s < T \leq s+\tau)}{(1-v)P_T(s < T \leq s+\tau) + P_T(T > s+\tau)} \qquad (7)$$

With use of this quantity, the differential Q-values become functions of three continuous variables s, τ, β rather than just two, but satisfy a similar, although slightly more complex, recursive relationship. If we temporarily define $P^+(s,\tau) = v(\beta(s) + (1-\beta(s))P_T(T \leq (s+\tau)| T > s))$ as the total probability of success of the lever press of latency τ from state $\{1, s\}$, then we have:

$$Q^*(\{1,s,\beta(s)\},\tau) = \frac{a}{\tau} + b - \rho^*\tau + (1-P^+(s,\tau))V^*(\{1,s+\tau,\beta(s,\tau)\}) + P^+(s,\tau)(z+V^*(2)) \qquad (8)$$

where, as above,

$$V^*(\{1,s+\tau,\beta(s,\tau)\}) = \max_{\tau'}(\{Q^*(\{1,s+\tau,\beta(s,\tau)\},\tau')\}$$

It is convenient to note that, if we describe any deterministic policy π as the latency $\tau^\pi(s)$ of the lever press starting at state $\{s, \beta(s)\}$ within a trial (assuming that this exists), then we can evaluate the utility rate ρ^π as $\bar{r}^\pi(0,0)/\bar{t}^\pi(0,0)$, where

$$\bar{t}^\pi(s,\beta(s)) = \tau^\pi(s) + (1-P^+(s,\tau))\bar{t}^\pi(s+\tau^\pi(s),\beta(s,\tau^\pi(s))) + P^+(s,\tau)\tau_I \qquad (9)$$

is the average recurrence time between transitions from post-outcome to pre-outcome, and

$$\bar{r}^{\pi}(s,\beta(s)) = \frac{a}{\tau^{\pi}(s)} + b + (1 - P^{+}(s,\tau))\bar{r}^{\pi}(s + \tau^{\pi}(s), \beta(s,\tau^{\pi}(s))) + P^{+}(s,\tau)z \tag{10}$$

is the average net utility during this recurrent episode. Similar formulæ hold for the special cases above with $T = 0$ and $v = 1$.

Punishment

The problem of punishment is illustrated in Fig. 1B. In this case, a punishing outcome (of utility z, albeit with $z < 0$) is delivered unless a successful lever press is emitted before the time T, drawn from $p_T(T)$, at which the outcome is programmed. As before, the probability that a lever press is successful is v. Again, there are two characteristic states (both of which are signalled): 1, pre-outcome, and 2, post-outcome, and the transition from state 2 to state 1 involves the inter-trial interval τ_I. However, now, the transition from state 1 to state 2 can happen through the delivery of the punishment at time T, as well as through a successful lever press. Furthermore, if the subject is in state 1, then it will receive the outcome as soon as it is scheduled – so there is never any call for partial observability, even if the probability of success is $v < 1$. The process is still semi-Markov, with the punishment also able to change the discrete state.

One question is the net utility when the lever press is chosen to terminate at time τ (starting at time s) but the outcome arrives at time T, with $s < T \leq (s + \tau)$. We consider the simple case that the time-dependent component of the utility is proportional to the length of time that the press has been in progress before the punishment is scheduled. That is, we make the simplification that cost accrues at a uniform rate throughout the whole latency of the pressing action (in fact, in just the same way as the opportunity cost), and so less has been expended if the punishment comes early in the latency than if it comes late. This makes the total contribution to the utility, including the outcome itself, be

$$\frac{a}{\tau}\frac{T-s}{\tau} + b + z. \tag{11}$$

To put this another way, there is a utility per unit time of execution of $\frac{a}{\tau^2}$ so that the total utility if the press completes without interruption is $\frac{a}{\tau}$.

Using the same reasoning as before, the optimal Q values satisfy the recursive relationship

$$\begin{aligned} Q^*(\{1,s\},\tau) = {} & P_T(T > (s+\tau)|T > s)\Big(\frac{a}{\tau} + b + vV^*(2) \\ & + (1-v)V^*(\{1,s+\tau\}) - \rho^*\tau\Big) \\ & + \int_{u=0}^{\tau} du\, p_T(T = s+u|T > s) \\ & \times \Big(\frac{a}{\tau}\frac{u}{\tau} + b + z + V^*(2) - \rho^* u\Big) \end{aligned} \tag{12}$$

taking account also that the opportunity cost (or, as we will see, benefit) of time associated with the press is also restricted to the time $u = T{-}s$, if the press is terminated early by receipt of the outcome. The first term in Eqn 12 comes from the case that the punishment is not programmed during the latency τ, and includes the immediate cost for the lever press and the future costs associated with the transition to either state 2 (if the lever press is successful) or to state $\{1,s+\tau\}$ (if it is not). The second term evaluates the utility that accrues when the punishment arrives before the lever press is complete (at time $s < u \leq s + \tau$). This always leads to state 2 and outcome z, but only the fraction in Eqn 11 of the latency-dependent cost. Note that whereas for the case of reward, the only transition to state 2 comes from a successful lever press, here it comes either from that or from the delivery of punishment.

As for Eqns 9 and 10, for any policy π that defines $\tau^{\pi}(s)$, we can write the average rate of utility as $\bar{r}^{\pi}(0)/\bar{t}^{\pi}(0)$, where

$$\begin{aligned} \bar{t}^{\pi}(s) = {} & P_T(T > (s+\tau^{\pi}(s))|T > s)(\tau^{\pi}(s) + v\tau_I + (1-v)\bar{t}^{\pi}(s+\tau^{\pi}(s))) + \\ & P_T(T \leq (s+\tau^{\pi}(s))|T > s)\tau_I + \int_{u=0}^{\tau^{\pi(s)}} du\, p_T(T = s+u|T > s)u, \text{ and} \\ \bar{r}^{\pi}(s) = {} & P_T(T > (s+\tau)|T > s)\left(\frac{a}{\tau^{\pi}(s)} + b + (1-v)\bar{r}^{\pi}(s+\tau^{\pi}(s))\right) + \\ & \int_{u=0}^{\tau^{\pi}(s)} du\, p_T(T = s+u|T > s)\left(\frac{a}{\tau^{\pi}(s)}\frac{u}{\tau^{\pi}(s)} + b + z\right) \end{aligned}$$

Finally, note that we can set the parameters of this version of the problem to accommodate a different form of rewarding task, in which the subject only gets a reward (of utility R, say) if the lever is successfully pressed before a stochastic deadline T. This then resembles the task used by (Guitart-Masip *et al.*, 2011a) in order to encourage quick pressing. We do this by adding a constant R to the value of state 2, and making $z = -R$ resemble a punishment. In this case, the subject will always garner a utility of R unless the punishment happens before a lever press, in which case the net utility associated with the outcome will be 0. We can formally still set $V^*(2) = 0$, but now the formula governing the average utility per recurrent cycle becomes

$$\begin{aligned} \bar{r}^{\pi}(s) = {} & P_T(T > (s+\tau)|T > s)\left(\frac{a}{\tau^{\pi}(s)} + b + (1-v)\bar{r}^{\pi}(s+\tau^{\pi}(s)) + vR\right) \\ & + \int_{u=0}^{\tau^{\pi}(s)} du\, p_T(T = s+u|T > s)\left(\frac{a}{\tau^{\pi}(s)}\frac{u}{\tau^{\pi}(s)} + b + z + R\right) \end{aligned}$$

Distributions

The last thing we need to specify is the distribution governing T. This could be seen as modelling two factors – the actual arming distribution that the experiment might involve (for instance, in a variable interval schedule); and the uncertainty about internal timing associated with the subjects' own estimates (Gibbon, 1977). We impute all the uncertainty in timing into the external variable T, and thus ignore certain subtle issues associated with interval timing, such as whether it is the objective or subjective (and hence stochastic) rate of utility that we should model the animal as optimizing.

For convenience, we treat T as coming from a gamma distribution $T \sim \gamma(T;k,\theta) = \frac{T^{k-1}e^{-T/\theta}}{\Gamma(k)\theta^k}$. Parameter k is called the shape, and θ is the scale. Γ(k) is the gamma function, which generalizes the factorial function. This distribution has mean $k\theta$ and variance $k\theta^2/2$.

The upper row of Fig. 2 shows three characteristic gamma distributions that we use throughout, referred to as d^{l}, d^{m} and d^{s}. These have $k = 1, 10, 10^3$ and $0 = 1, 1/10, 1/10^3 s$ respectively, so they have the same mean of 1s, but variances of $1/2s^2, 1/20s^2$, $1/2 \times 10^{-3}s^2$. The lower row shows the continuation densities $\mathrm{p}_T(T = (s + u)|T > s)$ for three values of s (shown by the three colours). Distribution d^{l}, which is an exponential distribution, is called memoryless, as the continuation distributions are all identical, exponential distributions. This is evidently not the case for the others. We also consider a trivial distribution d^0 for which $T = 0$ deterministically. This is the distribution associated with (Niv *et al.*, 2007)'s treatment; however, it is not appropriate for the case of punishment, as it would afford no opportunity for successful avoidance.

For the case of reward with stochastic success of the lever pressing (i.e. $v < 1$), Fig. 3 uses these three distributions to illustrate the progression of the belief $\beta(s)$ that the lever has actually been armed. Here, we consider the case that the lever is pressed at $s = \{0.6, 1.2, 1.8\}$s in the trial without leading to the reward. The bars show $\beta(s)$ at these times for $v = 0.1$ (red), $v = 0.5$ (green), and $v = 0.9$ (blue). The three plots are for the three different distributions d^l, d^m, d^s. The bars show the combination of prior probability of arming (this is most clear for d^s, for which the prior probability is near 0 for $s = 0.6$ seconds and near 1 for $s = 1.2$, 1.8 seconds) and the likelihood associated with having not yet obtained the reward.

Example

Figure 4 shows the inner workings of the costs for a case of reward ($z = 4$) for the four distributions and when lever pressing is always successful ($v = 1$). The other parameters take the values $a = -1$, $b = 0$, $\tau_I = 1$s). The upper plots show optimal differential state-action values Q^* $(\{1,s\}, \tau)$ (coloured lines) associated with the pre-reward state for a variety of times $s^k = \{0, 0.6, 1.2, 1.8\}s$ within a trial (red, green, blue, and magenta, respectively), and optimal differential state values $V^*(\{1,s\})$ (dashed black line). The coloured asterisks on the plots show how the maximal Q^* values form the optimal V^* values. The plots use s^k as the origins for each line $Q^*(\{1, s^k\},\tau)$. The curves are truncated for distributions d^s and d^0 when the probabilities $\mathrm{P}_T(T > s)$ are so low that the relevant states are visited with a probability less than around 10^{-10}.

The lower plots in Fig. 4 show the optimal latencies $\tau^*(s)$ for each starting time s. These constitute the optimal policy, and are the key output of the model.

The different distributions shown in Fig. 4 illustrate some of the central factors controlling vigour. Consider first d^0, for which the lever is armed as soon as the pre-outcome state is entered (Niv *et al.*, 2007). In this case, only a single lever press will ever be necessary (as it does not fail), with a latency chosen to balance the actual and opportunity costs of the press. The values $Q^*(\{1, 0\},\tau)$ for these latencies show the two costs rather clearly – for short τ, going from the hyperbolic actual cost for τ near 0, and the linear opportunity cost for τ larger. Below, we show the consequences of increasing and decreasing these costs.

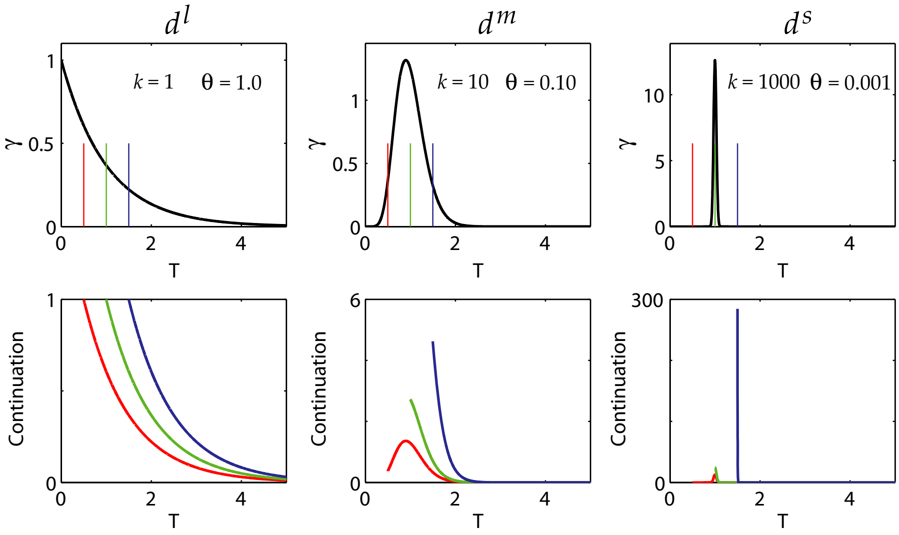

FIG. 2. Distributions. The upper plots show gamma densities of $p_T(T)$ for three values of k and θ. The lower plots show continuation values [i.e. the conditional distributions $\mathrm{p}_T(T = (s + u) \mid T > s)$] for the three values of s shown by the coloured bars in the upper plots. Please consult the on-line version of the paper for colour figures.

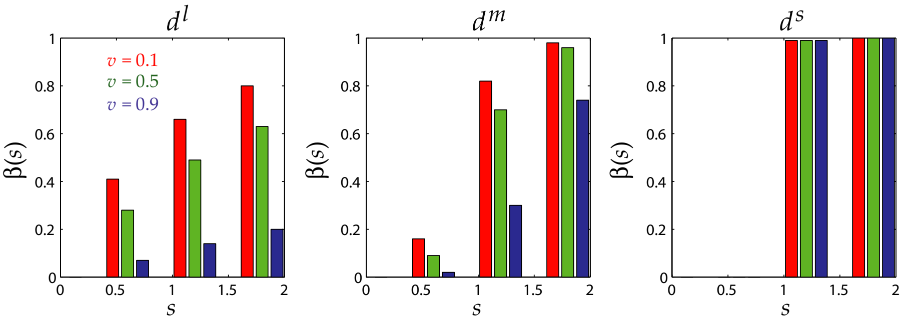

FIG. 3. Belief inference. Each bar plot shows the beliefs that the lever has been armed after three successive presses at $s = \{0.6, 1.2, 1.8\}$ s for $v = \{0.1, 0.5, 0.9\}$ (red, green, and blue, respectively). The three plots are for the three distributions d^l, d^m and d^s.

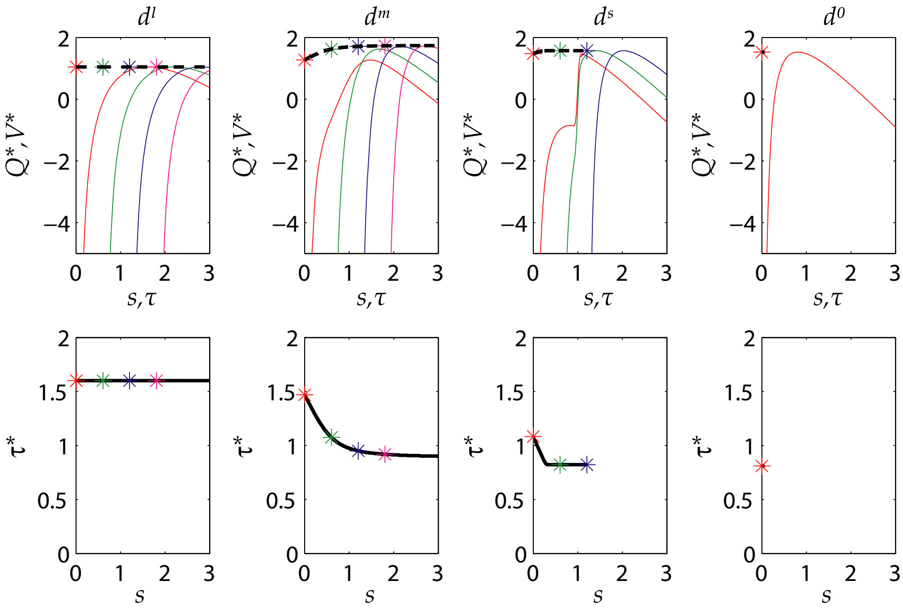

FIG. 4. Optimal Q^*, V^* and τ^* values for the four distributions, for parameter values $a = -1$, $b = 0$, $z = 4$, $\tau_I = 1$s. The upper plots show $Q^*(\{1, s^k\}, \tau)$ as a function of τ (using the s axis but starting from s_i) for $s^1 = 0$ (red), $s^2 = 0.6$ (green), and $s^3 = 1.2$ (blue), $s^4 = 1.8$ (magenta), all in seconds, along with $V^*(\{1,s\})$ (dashed black line) for all s. The coloured asterisks show the values of $V^*(\{1,s^i\})$ that come from the peaks of the Q^* plots. The lower plots show $\tau^*(s)$. The plots are truncated when $P_T(T > s) < 10^{-10}$.

Distribution d^s is close to what we would expect for a fixed interval schedule, as the lever is armed at almost exactly 1 s after the pre-outcome state is entered. Once this happens in a trial, the problem is very much like that for d^0, except that the opportunity cost is a little less, as the delay to the outcome decreases the optimal overall utility rate ρ^*. Thus, the latencies τ^* show that, starting from $s = 0$, it is optimal to wait until just after the lever is very likely to have been armed, and then press. Starting just later in the trial, the actual time at which it is optimal to press increases slightly [i.e. $s + \tau^*(s)$ increases slightly], but there is an asymptotic latency that is marginally longer than for d^0 because of the reduced opportunity cost of delay. The optimal Q^* values for d^s show how these latencies come about. $Q^*(\{1{,}0\},\tau)$ (red) shows the futility of pressing before $\tau = 1$; for $s > 1$, $Q^*(\{1{,}s\},\tau)$ shows a similar pattern to those for d^0.

The values and policy for distribution d^m resemble smoother versions of those for d^s. Here, T is less deterministic, and thus the possibility that the lever is not armed adds to the actual cost of pressing the lever quickly to balance out the (also increased) opportunity cost. This collectively results in slower lever pressing, and, indeed, the chance of having to press the lever more than once (which is around 8% for these parameters).

Finally, distribution d^l is a standard, exponentially distributed, variable interval schedule. This is memoryless in the sense that the distribution of times that the lever will be armed after time s in the trial if it is not already armed at s is independent of s. Thus, we expect $V^*(\{1{,}s\})$ and $\tau^*(s)$ to be independent of s (as indeed is apparent in Fig. 4). Now, even though the mean arming time is still 1s, as also for d^m and d^s, the optimal latency at $s = 0$ is rather greater, and yet there is still a 20% chance that more than one lever press will be necessary.

Results

We consider a collection of timing tasks with three key characteristics. First, the latency to press a lever determines the cost (falling off in a hyperbolic manner), so that faster pressing is more costly. Second, the time at which the lever is planned to be pressed has a bearing on whether and when a reward or punishment is delivered. Finally, lever pressing is not always successful – there is a chance of $1{-}v$ that it fails to deliver reward or avoid punishment on any given press. We separate our consideration into the cases of reward and punishment; although these can, of course, also be mixed.

The Methods section described these processes in detail, showing the provenance of optimal solutions whose nature is illustrated here. The state and state-action predictions to which this leads will engender a set of tonic, quasi-tonic and phasic prediction error signals; these are also described later, and ascribed to opponent neuromodulatory systems.

Rewards

Figure 1A shows the standard model for reward, in which delivery is conditional on the lever having been ‘armed’. This happens after a time T that can either be 0 (we call this case distribution d^0) or can be stochastic, drawn from a density $p_T(T)$. For convenience, we use the gamma distributions shown in Fig. 2, which we call d^l, d^m, and d^s, with the labels reflecting their variances. If the lever is pressed when it is not armed, it has no effect, and so the subject just has to press it again later in the hope that it has by then been armed. We first consider the case that the lever is deterministically successful.

Figure 5 shows example results for this case, for the four distributions mentioned, for three different utilities for the actual outcome: $z = 2$ (red), $z = 4$ (green), and $z = 8$ (blue), and across different values of parameter a, which determines the utility per inverse second (so that the utility of latency τ is a/τ). Here, we set the unit utility of the press is set to $b = 0$, and the inter-trial interval to $\tau_I = 1$s.

The top row shows the optimal latencies [called $\tau_1^* \equiv \tau^*(0)$, as these are the latencies starting at time $s = 0$ during a trial] for the lever press as a function of a, for the various conditions. The middle row shows the optimal rate of utility (ρ^*) that results. The bottom row shows the relationship between ρ^* and τ_1^*.

For all of the distributions, when the cost of acting quickly is so great that (i) the lever is sure to be armed by the time that the lever is pressed, and (ii) the latency outweighs the inter-trial interval (i.e. $\tau_1^* \gg \tau_I$), then we have $\tau_1^* \simeq \sqrt{(-a)/\rho^*}$ (as in Eqn 3, and from (Niv *et al.*, 2007)) and $\rho^* \simeq z/\tau_1^*$ (because one outcome, worth *z*, is acquired each $\tau_1^* + \tau_I \simeq \tau_1^*$ seconds). Putting these together, $\tau_1^* \simeq -a/z$, as is evident in the linear relationships for large costs in the top row. Furthermore, the bottom row of plots demonstrate the relationship that τ_1^* is consistent with being roughly proportional to $\sqrt{-a/\rho^*}$.

When the utility per inverse latency *a* is nearer 0, the different characteristics of the distributions become apparent. For d^0, it is possible to press and collect rewards very frequently (although the utility rate cannot be greater than z/τ_I; for d^s, there is little point in pressing before about 1s, as is apparent even for the most valuable outcome *z* = 8; blue). The precise dependence of τ_1^* on *a* depends on the exact characteristics of the distributions. The bottom row shows the relationship in a slightly different way – unlike for the case of d^0, for distributions d^l, d^m and d^s, even as the (optimal) utility rate grows (as *z* gets bigger and *a* gets nearer 0), the optimal latency τ_1^* is distant from 0.

Figure 6 shows the effect of making lever pressing only partially successful, that is, setting *v* < 1. The stacked bar plots show the optimal latencies of each of four lever presses (from bottom to top), assuming that all previous ones had been unsuccessful, for *v* = {0.1, 0.5, 0.9} (and, for comparison, for *v* = 1, for which there is only a single latency). The plots exhibit a competition between two effects: when *v* is low, the overall rate of utility is small, whence the opportunity cost of acting slowly is also limited. Thus, long latencies are favoured. However, when pressing is only stochastically successful, the subjects will be unsure whether or not the lever has been armed (formally, the decision problem becomes partially observable). Then, as seen in Fig. 3, the smaller the probability of success, the more sure the subject will be that the lever is armed, despite an unsuccessful press. This tends to hasten lever pressing. The 'V'-shape in the latencies for d^l and d^m show this well. With respect to d^l, even though arming is memoryless, the belief inference process is not – it is cumulative – and so the latencies of successive lever presses are not identical.

Figure 7 generalizes this result to a range of values of *a* and v. It has the same format as Fig. 5, showing the latency τ_1^* of the first lever

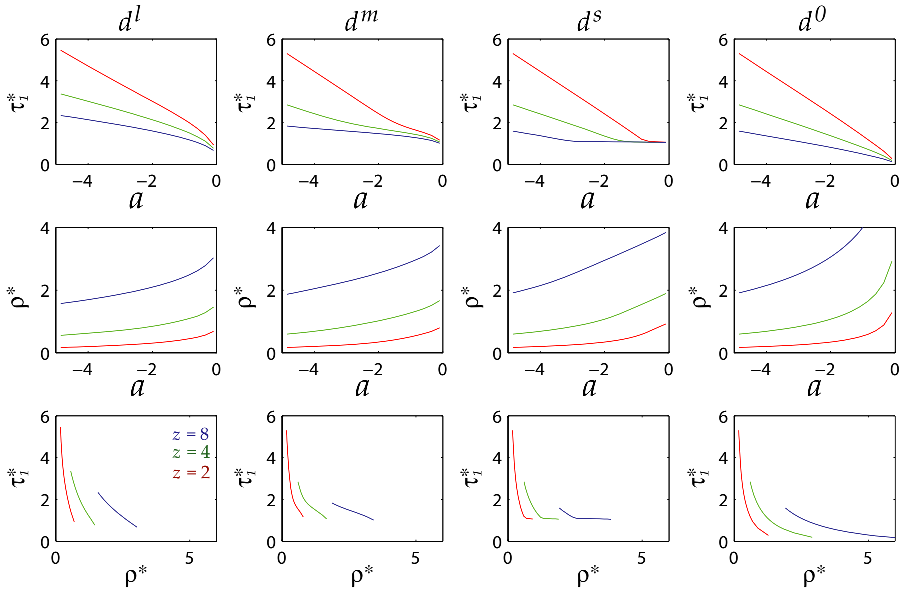

FIG. 5. Optimal latencies $\tau_1^* \equiv \tau^*(0)$ and utility rates ρ^* as a function of the parameter *a* governing the hyperbolic cost of latencies, for outcome values of *z* = 2 (red), *z* = 4 (green), and *z* = 8 (blue), and for the four distributions d^l, d^m, d^s and d^0 described in the Methods section. The top row shows τ^* as a function of *a* (note the approximate linear dependence for large |*a*|), the middle row shows ρ^* as a function of *a*, and the bottom row shows τ_1^* as a function of ρ^* when *a* is varied.

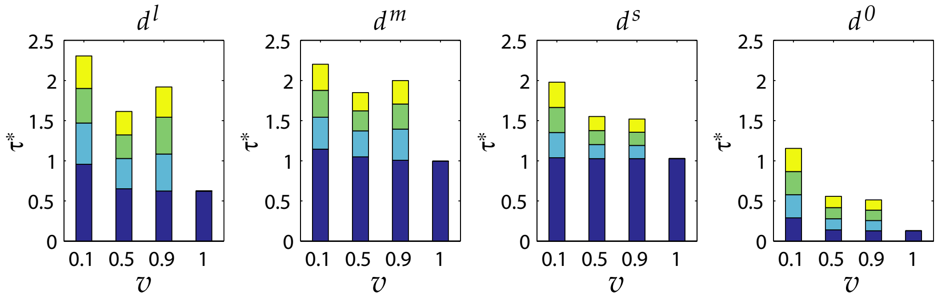

FIG. 6. Optimal latencies for the first four lever presses, assuming that all previous presses were unsuccessful. Each bar shows the latencies as a stack (so the second latency is shown by the height of the lower two bars, and so forth), for *v* = {0.1, 0.5, 0.9} with the (single) optimal latency for *v* = 1 shown at the far right. Here *a* = −0.05; *b* = 0; *z* = 4.

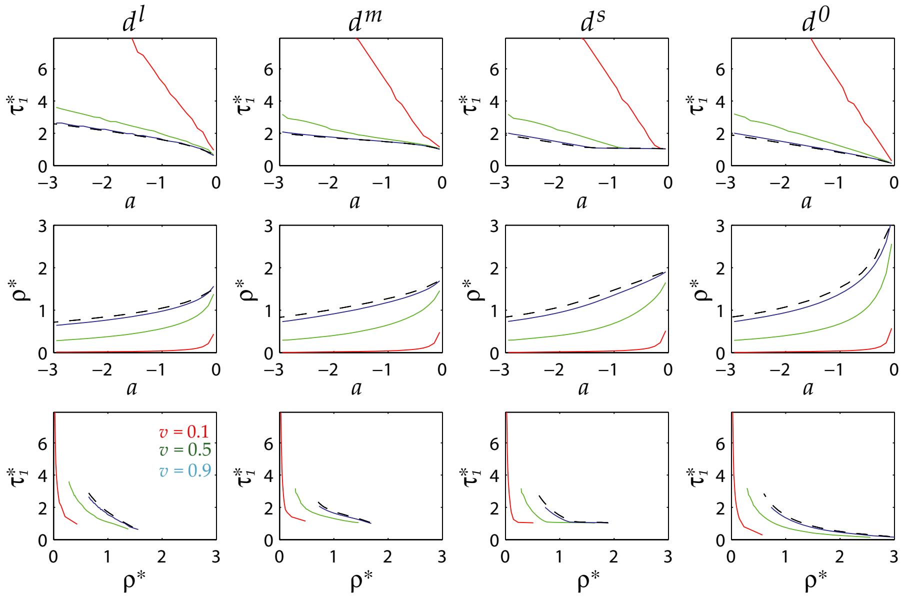

FIG. 7. Optimal first press latencies τ_1^* and utility rates ρ^* for stochastically successful lever pressing. This plot has the same format as Fig. 5, except that $z = 4$ and the colours represent different values of $v = 0.1$ (red), $v = 0.5$ (green), and $v = 0.9$ (blue). The black dashed line is for $v = 1$, from Fig. 5. The top row shows the optimal latency of the first lever press τ_1^* as a function of a; the middle row shows the optimal utility rate ρ^* as a function of a; and the bottom row shows the latency τ_1^* as a function of ρ^*. Other parameters and distributions are as in Fig. 5.

press, except that all of the curves are for $z = 4$ (which was the green trace in Fig. 5, repeated here as the black dashed lines). In Fig. 7, the colours signify different values of v, with $v = 0.1$ in red, $v = 0.5$ in green, and $v = 0.9$ in blue. For small values of v, the utility associated with collecting the reward gets very large, and so it becomes barely worthwhile – the latencies increase dramatically. If there was also a unit utility of each press (i.e. parameter $b < 0$), then, depending on z, the rewards could be avoided altogether. Again, for latencies that are sufficiently large that the lever is sure to be armed, the optimum is roughly

$$\tau_1^* \simeq \frac{1-v}{v^2}\frac{-a}{z + b/v^2}$$

The term $z + b/\mathrm{v}^2$ is the average net (latency-independent) utility per cycle. If this is negative, then it is not worth trying to press the lever at all. Otherwise, if $b = 0$, the linear dependence that we remarked upon earlier changes from $\tau_1^* \simeq -\frac{a}{z}$ to $\tau_1^* \simeq \left(-\frac{a}{z}\right) \times \frac{1-v}{v^2}$, which becomes an extremely steep function of $-a/z$ as $v \rightarrow 0$.

Punishments

We assume that the utility of lever pressing itself is the same for the case of punishments; however, the structure of the task is quite different. In particular, a punishment is delivered if a successful lever press is not emitted before the arming time, whereas above, rewards could only be collected by a successful lever press after the given time. As discussed by (Cools *et al.*, 2011), this provides a potentially strong instrumental incentive for haste – delaying pressing the lever too long brings the possibility of acquiring the large punishment. In some cases, this is balanced by the sort of opportunity benefit of delay or sloth discussed by (Dayan & Huys, 2009; Boureau & Dayan, 2011; Cools *et al.*, 2011), in that, even though acting quickly might safely avoid the current punishment, it also brings forward the advent of the next and subsequent possible punishments – therefore, delaying can make perfect sense. More formulaically, the term $-\rho^*\tau$ in the definition of the state-action values (see Methods) is actually positive, encouraging slow pressing. This is exactly opposite to its effect for the case of reward, for which it encourages fast pressing. Note additionally that the subjects are never in doubt as to whether the lever is 'armed' (as they will receive the punishment as soon as it is), and so the problem does not suffer from the sort of partial observability that we saw for rewards.

Figure 8 shows how these factors play out for various different probabilities of success of lever pressing, using distributions d^l, d^m, d^s and success probabilities as for Fig. 6. Distribution d^0 is not meaningful in the punishment context, as the subject would receive the punishment before having had the chance to act to avoid it. For illustrative purposes, the punishment is moderately unpleasant ($z = -4$) and the utility per inverse time is rather modest ($a = -0.05$).

Consider first d^s, for which the time of the punishment is rather circumscribed. For low v, that is, rather uncontrollable punishment, the latency of the first lever press is actually shorter than for larger v – this is to allow for the possibility of pressing again relatively cheaply if this first press fails. Conversely, as the time of the punishment nears, for low v, it is not economic to press the lever quickly enough to have the chance of beating the punishment, as the lever press is likely to be unsuccessful. By contrast, for high v, lever pressing gets more frenetic as the punishment time approaches, as it is worth trying to avoid. However, again, after some time, the cost of the lever press that is sufficiently fast to avoid the punishment gets so great that it is better just to wait and suffer z.

The optimal latencies associated with distribution d^l are quite different. This distribution is memoryless, so that the policy is the same no matter when the choice is made (this is different from the case of reward). Thus, the latencies are all the same for a given probability v. However, they increase as the probability of success decreases. This is because the cost of acting quickly is balanced against the benefit in terms of reducing the chance of being shocked. As v decreases, the benefit decreases too, and so the balance point is moved in the direction of slower responding. Indeed, if we choose the time τ_l to cancel out the effect of the opportunity cost (by making it decrease if

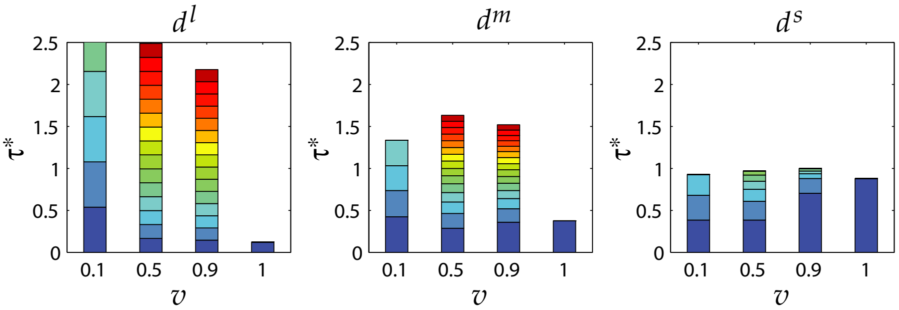

FIG. 8. Optimal latencies for up to the first 15 lever presses for the case of punishment, assuming that all past presses were unsuccessful. Each bar shows the latencies as a stack (so the second latency is shown by the height of the lower two bars, and so forth), for $v = \{0.1, 0.5, 0.9\}$ with the (single) optimal latency for $v = 1$ shown at the far right. Here, $z = -4$; $a = -0.05$; $b = 0$.

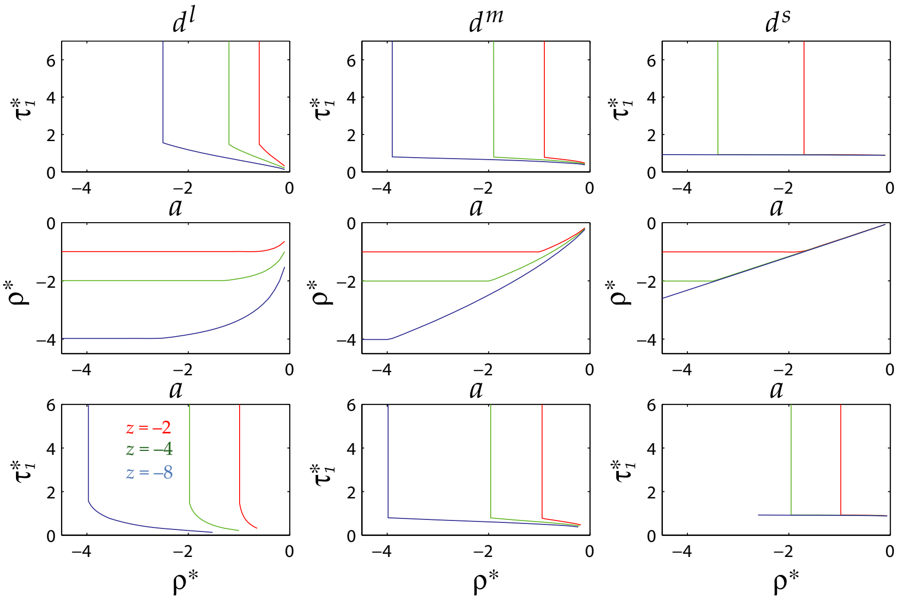

FIG. 9. Optimal first press latencies τ_1^* and utility rates ρ^* as a function of the parameter a governing the hyperbolic utility of latencies, for punishment values of $z = -2$ (red), $z = -4$ (green), and $z = -8$ (blue), and for the three standard distributions d^l, d^m, d^s. The figure has the same structure as Fig. 5. The top row shows τ_1^* as a function of a (note the approximate linear dependence for large $-a$), the middle row shows ρ^* as a function of a, and the bottom row shows τ_1^* as a function of ρ^* when a is varied.

the post-outcome state is entered more quickly, and increase if the post-outcome state is entered more slowly), lower probabilities v can still lead to longer latencies because of this effect. The latencies for distribution d^m are intermediate between those for d^s and d^l.

Figure 9 shows the optimal first latencies and utility rates as a function of a for three values of the punishment: $z = -2$ (red); $z = -4$ (green), and $z = -8$ (blue) – for the three distributions d^l, d^m, d^s. The comparison with the case of reward (Fig. 5) is rather stark. Whereas for rewards, when $b = 0$, it is always worth pressing the lever in the end provided that $z > 0$, this is not true for punishments. If the punishment is weak such that the cure is worse than the disease – that is, the cost of pressing the lever sufficiently quickly is greater than the benefit in terms of punishment foregone – then the latency becomes effectively infinite (shown by the vertical lines), and the subject just waits for the punishment. The value of a at which this happens scales with z; for d^s, it happens when $a = z$ [although the plots in the figure are derived from a density $\mathrm{p}_T(T)$ that is slightly smoother than a delta function and so the critical value is millimetrically displaced]. This behaviour is also readily apparent in the plot of τ_1^* against ρ^*. As $\tau_I = 1$s, and all of the distributions specify that the punishment comes after 1s on average, the rate of utility given infinite sloth is $\frac{z}{2}$ utils/s. Thus, it is only worth pressing the lever more quickly if the net cost leads to a less negative utility rate than this.

Prediction errors and neuromodulation

One point of the modelling in this article is to make predictions for the activity of neuromodulatory systems. In view of results on the phasic and tonic activity of dopamine cells from (Montague *et al.*, 1996; Schultz *et al.*, 1997) and (Niv *et al.*, 2007), we consider the total prediction errors associated with the various possible outcomes. Although it would be interesting to study the course of learning, we confine ourselves to considering optimal choices, for which the same prediction errors would arise from conventional temporal difference (TD) learning (Sutton, 1988; Montague *et al.*, 1996), and both Q−learning (Watkins, 1989; Roesch *et al.*, 2007) and the so-called SARSA rule (Rummery & Niranjan, 1994; Morris *et al.*, 2006). We then attempt to relate these prediction errors to the activity of selected parts of the neural substrate. We concentrate on the case of punishment, as the resulting questions are sharper. Note again that these prediction errors are tied to the optimal latencies; if, for instance, subjects press more than they should, then the expectations about the activity will not be correct, at least in detail.

Figure 10 shows prediction errors for the first few lever presses $\tau_1^*, \tau_2^*, \ldots$ for the three distributions d^l, d^m, d^s and three values of $v = 0.1$, 0.5, 0.9. These are exactly the lever presses whose latencies are reported in Fig. 8. We make the simplifying assumption that the prediction errors only arise when the lever press terminates, rather than being continuous throughout its execution. This is exactly the way in which prediction errors for a particular set of temporally extended actions (called options (Sutton *et al.*, 1999)) have been treated (Duik *et al.*, 2010). We ignore any additional prediction errors that may arise in between terminations.

Each press can terminate in one of three ways: (i) finishing successfully in arranging a transition to the post-outcome state (blue '+'; left-hand axis); or (ii) finishing, but without this transition being successful (blue 'x'; left-hand axis); or (iii) the punishment arriving while the lever is in the process of being pressed (green lines; right-hand axis). In the last of these cases, the punishment could come at any time (between $s_{i-1}^*$and $s_i^* = s_{i-1}^* + \tau_i^*$), and so the prediction errors are indexed by the time s. The cyan 'o's (left-hand axis) show the weighted integrals of these terms, quantifying the net contribution for the possibility that the punishment might arrive before the lever press finishes. If we weight the blue points by their probabilities of occurring, then the sum of those plus the cyan points is 0 at every termination of a lever press, as this is what defines the Q^* and V^* values. The blue asterisk (at $s = 0$; left-hand axis) shows the initial value $V^*(\{1,0\})$ [as it is commonly found, at least in the phasic activity of dopamine cells, that prediction errors at the start of trials are reported relative to a 0 baseline; e.g. (Roesch *et al.*, 2007; Morris *et al.*, 2006; Fiorillo *et al.*, 2003; Tobler *et al.*, 2005)].

The total prediction errors are distinctly different for the different cases. In particular, for d^1, for ineffective presses ($v = 0.1$), when the press terminates but does not succeed, the prediction error (blue 'x') is actually positive. This is because the punishment has at least been successfully postponed. Likewise, if a punishment is provided (green lines), then the later it comes over the course of the lever press, the less negative the prediction error – again from the opportunity benefit of sloth – that is, time having passed. When pressing is more successful, the opportunity benefit is smaller (as the average utility rate is less negative) and so fails to outweigh the cost associated with having had to be pressing the lever for longer. Thus, the green lines reverse their slopes.

For d^s, the punishment is overwhelmingly likely to arrive at $s = 1$. Again, being forced to press the lever early to have the chance of avoiding the punishment successfully has the attendant cost of bringing forward the next opportunity to suffer punishment. It therefore competes with the necessity to press the lever more quickly on subsequent occasions in the trial so that success is likely to be achieved before $s = 1$. For $v = 0.1$, the net expected success is low, and therefore the prediction errors are more positive when the lever press succeeds than they are negative when it fails. For $v = 0.9$, the converse is true. They are evenly balanced for $v = 0.5$. This effect is not apparent for d^l and d^m, as for $v = 0.9$, there is still a substantially greater chance of actually getting the punishment during the time committed to a lever press, and so a successful lever press remains surprising.

The final issue is to estimate how these net prediction errors are represented in neural activity. One of (Niv *et al.*, 2007)'s main claims for the case of reward was that the average utility rate ρ^* is represented by the tonic activity of dopamine neurons, in a way that could influence vigour. This would allow an implementation of the optimal latency being $\tau^* = \sqrt{\frac{-a}{\rho^*}}$, appropriate for $v = 1$ and distribution d^0. This putative relationship with dopamine was based on a variety of results, such as those discussed in (Salamone & Correa, 2002), noting also that tonic and phasic dopamine activity could be subject to different regulation (Goto & Grace, 2005; Goto *et al.*, 2007). This would allow phasic activity to retain its conventional role as a temporal difference prediction error (Sutton & Barto, 1998) for future reward (Montague *et al.*, 1996; Schultz *et al.*, 1997).

However, various factors complicate such assignments for the case of punishment. How is utility rate represented in the case of aversive outcomes; and how do these long-run quantities relate to immediate facets associated with ever-changing predictions and delivery of punishment? The complexity is that the neuromodulatory substrate

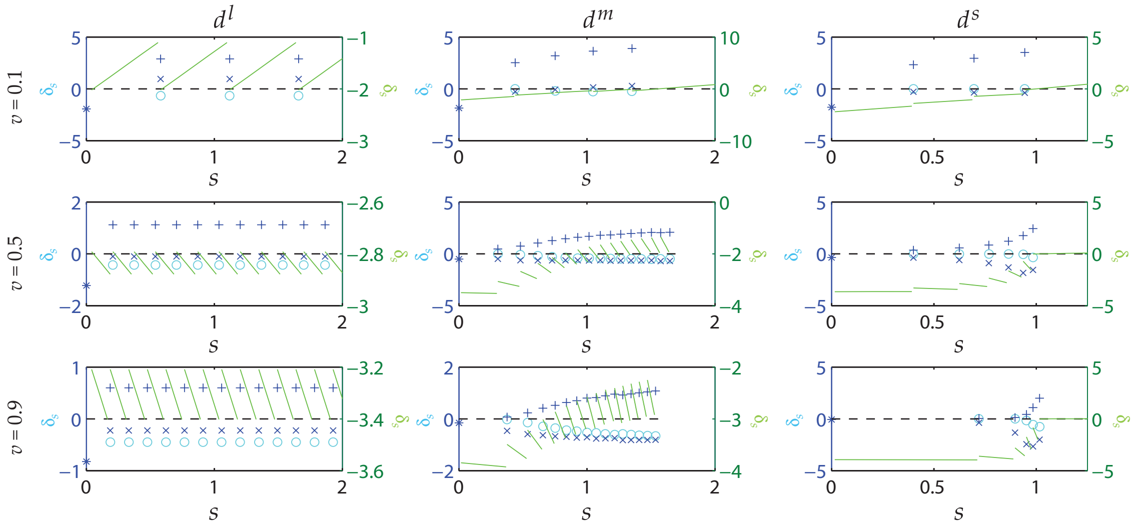

FIG. 10. Complete temporal difference prediction errors at the time of punishment (green lines; right-hand scale) or the successful (blue '+'s) or unsuccessful (blue 'x's) termination of the lever press. The cyan 'o's show the total expected punishment-based prediction error, accumulated across the time devoted to each lever press. The blue '*' at $s = 0$ (left-hand scale) shows the prediction error associated with the initial value V^* ({1, 0}). The results are for the three distributions (d^l, d^m, d^s; columns) and for three probabilities of success ($v = 0.1$, 0.5, 0.9; rows). The other parameters are $z = -4$; $a = -0.05$; $b = 0$, as for Fig. 8.

for the representation of punishment is subject to significant debate – it is unclear whether it is one of the actions of a different neuromodulator such as serotonin (Deakin, 1983; Deakin & Graeff, 1991; Daw *et al.*, 2002; Cools *et al.*, 2008, 2011; Dayan & Huys, 2009; Schweimer & Ungless, 2010; Boureau & Dayan, 2011), dips below baseline of dopaminergic activity (Frank *et al.*, 2004), the elevated activity of some among the population of dopamine neurons (Mirenowicz & Schultz, 1996; Joshua *et al.*, 2008; Matsumoto & Hikosaka, 2009), or an anatomically distinct set thereof (Brischoux *et al.*, 2009; Lammel *et al.*, 2011).

First, a wealth of experiments involving tryptophan depletion [reviewed, for instance, in (Cools *et al.*, 2008, 2011)] certainly implicate serotonin in the processing of aversive information. Unfortunately, although recordings have been made from raphe neurons of awake behaving animals (Jacobs & Fornal, 1997, 1999; Jacobs *et al.*, 2002; Nakamura *et al.*, 2008; Ranade & Mainen, 2009; Bromberg-Martin *et al.*, 2010; Miyazaki *et al.*, 2011a) (albeit with recent investigations considering delays and not punishments), it is well-nigh impossible to discriminate serotonergic from non-serotonergic cells in these nuclei (Allers & Sharp, 2003; Kocsis *et al.*, 2006), and there is evidence for substantial heterogeneity in the behavioural correlates of true serotonergic neurons (Lowry, 2002). These confusions extend to the findings about functional magnetic resonance imaging signals in human experiments investigating gains and losses, which often show correlations between blood oxygen level-dependent signals in structures such as the ventral striatum that are targets of neuromodulatory systems, and prediction error signals (McClure *et al.*, 2003; O'Doherty *et al.*, 2003; Haruno *et al.*, 2004; Seymour *et al.*, 2004; Tom *et al.*, 2007; Delgado *et al.*, 2008). However, the blood oxygen level-dependent signal sometimes increases and sometimes decreases with the prediction error for future punishment, for reasons that have not been fully understood; and there may also be anatomical differentiation of reward and punishment in the striatum that is not always examined (Reynolds & Berridge, 2002; Seymour *et al.*, 2007).

Second, it is important to take account of evidence associated with two-factor theories (Mowrer, 1947) suggesting that a good way to interpret results in active avoidance is to consider that outcomes are measured against the prospect of punishment. That is, achieving safety becomes a surrogate reward, represented by normal systems involved in appetitive learning. Modern reinforcement learning theories based on these principles (Johnson *et al.*, 2002; Moutoussis *et al.*, 2008; Maia, 2010) can account for many of the perplexing results, including those involving manipulations of dopamine [e.g. (Beninger *et al.*, 1980)], which suggest the importance of this neuromodulator for learning of the avoidance response.

Given such uncertainties, it is not possible to be completely confident about the neural realization of prediction errors such as those in Fig. 10. We await electrophysiological or even cyclic voltammetry in experiments of this sort. However, for concreteness, Fig. 11 suggests one possible assignment to the neuromodulators dopamine (upper axis) and serotonin (lower axis), assuming that they are acting as opponents. Some of the many issues associated with opponency are reviewed extensively in (Daw *et al.*, 2002; Cools *et al.*, 2008, 2011; Dayan & Huys, 2009; Boureau & Dayan, 2011), and we will not rehearse those arguments here. Key to understanding this opponency, however, is that the net prediction error only constrains the difference between two signals. We suggest that various factors may be added to both signals, leaving the prediction error the same, but allowing for other important effects.

Each graph in Fig. 11 is for one distribution and one probability of success of the lever pressing, and shows five signals, which we describe in turn. We work with the case that $b = 0$. First, the dashed lines indicate the tonic activity or concentration of dopamine ($\overline{DA}$; red) and serotonin ($\overline{5HT}$; blue). We describe these as being tonic because they are constant across the whole trial. For average case utility learning, the net tonic signal should be the average rate of utility, ρ^*, which is always negative for cases such as this, in which there is at best punishment, which is costly to avoid. Given opponency, we will have $\rho^* = \overline{DA} - \overline{5HT}$. However, in keeping with the two-factor notion above, the dashed lines reflect the possibility that a common factor – related to the net rate of avoidable punishment – is added to both signals. This means that there is an increase above baseline in the tonic dopamine signal, associated with the possible benefit of safety. This is even though, as a pure case of active avoidance, there is no reward.

One obvious possibility for the magnitude of this common signal is the mean rate of potentially avoidable punishment $-z/\bar{t}^*$, where $\bar{t}^*$ is the mean time for a whole trial (from one transition from the post-outcome to the pre-outcome states to the next such transition) when following the optimal policy. This would imply that

$$\overline{DA} = \rho^* - \frac{z}{\bar{t}^*} \tag{13}$$

$$\overline{5HT} = -\frac{z}{\bar{t}^*} \tag{14}$$

As ρ^* is closer to 0 when punishments are more competently avoided, the tonic dopamine signal reflects the net chance of success. In this version, the serotonergic signal is much less sensitive to success, only depending on it weakly via its effect on the average trial length. Note that alternative baselines that themselves take account of how much of that punishment can actually be expected to be avoided might also be possible.

Second, the solid blue line shows the utility per unit time of the ongoing lever press (which has been added to the baseline $\overline{5HT}$ shown by the dashed blue line). This makes a total of

$$\widehat{5HT}_i = \overline{5HT} - \frac{a}{(\tau_i^*)^2}, \tag{15}$$

that we call quasi-tonic, as it has a persistent character, but changes at the times of choices. As is evident, it can take many different values over the course of a single trial (which is why it is not tonic), but it remains at a constant level over the course of a single latency (and so is not phasic). This quasi-tonic signal can be integrated over either a complete or an interrupted lever press to make the appropriate contribution to the total prediction error. As it is a cost, and inspired by findings about the association between the firing of neurons in the raphe and motor activity (Jacobs & Fornal, 1997, 1999; Jacobs *et al.*, 2002) and other heterogeneous behavioural events (Ranade & Mainen, 2009) [but see (Bromberg-Martin *et al.*, 2010)], we suggest that this is represented as an additional serotonergic signal. In Fig. 11, this signal is most dramatic for d^s, high v and as $s \rightarrow 1$, when the imperative is for very quick presses (duly justified by their likely success). Indeed, some of the values are truncated to fit on the plots.

Third, the red and blue bars show putative phasic dopamine and serotonin responses, respectively. These reflect abrupt changes to ongoing predictions that arise when aspects of the state change. Such stimulus-locked, and very brief, burst-like responses are well characterized for the case of dopamine (Schultz & Dickinson, 2000); apart from effects at the time of an actual punishment to an anaesthetized subject (Schweimer & Ungless, 2010), they are less clear for the case of serotonin. Again, two-factor theory suggests that there will be a

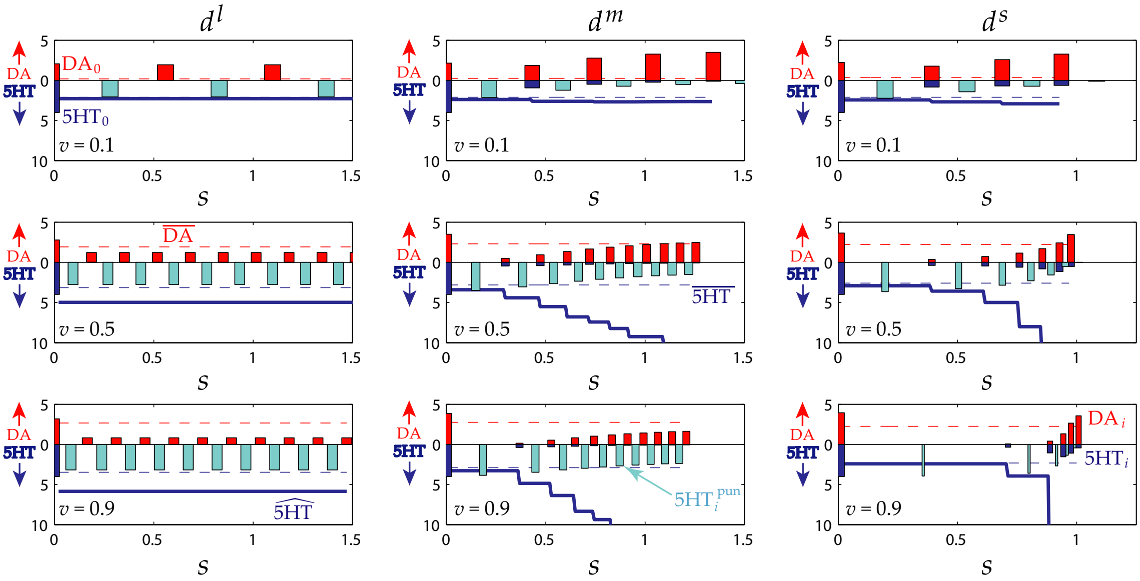

FIG. 11. Putative dopaminergic (DA; red; upper half) and serotonergic (5HT; blue and cyan; lower half) signals associated with the prediction errors of Fig. 10 for distributions d^l, d^m, d^s and $v = 0.1, 0.5, 0.9$. The dashed lines show contributions associated with the overall rate of utility. The blue solid line includes the additional utility per unit time of the current lever press (and so is described as quasi-tonic). The red and blue bars show the phasic contributions associated with success and failure, respectively, of lever pressing, at the times s_i^* in the trial at which they succeed or fail. The cyan bars show phasic consequences for the serotonergic signal when the punishment arrives. These are displayed at time $s_{i-1}^* + \frac{\tau_i^*}{2}$, halfway through the relevant lever press; however, they would actually be time-locked to the punishment itself.

baseline for the signals coming from the possibility of getting the punishment. There are two obvious baselines. Figure 11 reflects the first, which is that the baseline is the cost of potentially avoidable punishment, z, but applied to the differential state values V^* and state-action values Q^*. Temporal difference signals are differential, associated with the change over time in predicted values. Thus, a baseline applied to these predictions would reveal itself only at the onset of the task. That is, the initial signals would be

$$DA_0 = V^*(\{1,0\}) - z \qquad (16)$$

$$5HT_0 = \qquad -z \qquad (17)$$

Thus, as for the tonic dopamine signal, the fact of the avoidable punishment gives rise to a positive phasic dopamine signal. We can understand this by comparison with the familiar phasic prediction error responses associated with reward (Montague *et al.*, 1996; Schultz *et al.*, 1997). There, the prospect of future reward turns gives rise to a positive signal at the onset of an early reliable predictive cue. Here, the future reward is replaced by future safety (i.e. a transition from a state in which punishment is possible to one in which it has at least been postponed); this onset dopamine response is tied to that event. Nevertheless, under opponency, the net value at the outset is still $V^*(\{1,0\})$, once the simultaneous serotonergic activity is taken into account. If presses are routinely successful, then this net value is only very mildly negative (reflecting the cost of the active avoidance lever press).

According to this baseline scheme, subsequent success and failure of lever pressing only implicate either dopamine (red bars) or serotonin (blue bars), but not both, at the termination times $s_1^*, s_2^*, \ldots$ In the case of success, the two-factor manoeuvre of referring outcomes to a baseline expectation of getting the shock (i.e. z) makes the transition to safety appear like an immediate reward worth $-z$. The resulting contributions to the overall prediction error in Fig. 10 are thus:

$$\text{success}: \; DA_i = \qquad - V^*\left(\{1, s_{i-1}^*\}\right) \qquad (18)$$

$$\text{failure}: \; 5HT_i = -V^*\left(\{1, s_i^*\}\right) + V^*\left(\{1, s_{i-1}^*\}\right) \qquad (19)$$

These exhibit rather complex effects for the different distributions and probabilities of success. Of particular note is that the serotonergic signals are almost always very small. For d^l, they are 0, as the memoryless nature of the problem implies that the values $V^*(\{1, s\})$ are constant (and we are assuming that $b = 0$). However, even for d^m and d^s, these signals are rarely substantial. For low v, this is partly because there is little expectation that the press will succeed, and therefore little change in $V^*(\{1, s\})$ when it fails. Even for higher v, 5HT$_i$ becomes large only when the punishment goes from being reasonably to unreasonably avoidable; this happens punctately only for d^s, for which the time of the punishment is well constrained.

Finally, the cyan bars show the signal associated with the actual delivery of punishment. For convenience, these are shown halfway through the presses concerned (i.e. at times $s_{i-1}^* + \frac{\tau_i^*}{2}$, although they would really be time-locked to the punishment itself. Note that this term is constant across the time of the lever press, as the linear variation in the full prediction error shown in the green lines in Fig. 10 is absorbed into the integral of the quasi-tonic solid blue line which comprises the sum of the baseline cost $\overline{5HT}$ associated with the serotonergic component of the overall rate of utility, and the vigour utility per unit time $-\frac{a}{(\tau_i^*)^2}$. Given the two-factor baseline of $-z$, this term is the value associated with the time of the preceding lever press

$$5HT_i^{\text{pun}} = -z + V^*\left(\{1, s_{i-1}^*\}\right) \qquad (20)$$

These terms therefore show the evolution of the subject's expected values. Crudely, these values approach 0 as the expectation of punishment rises (most evident in d^m), although the net utility of future expected avoidance are also included, giving rise to the more complex dependencies evident for d^s.

Discussion

We analysed optimizing appetitive and aversive vigour in a set of tasks that generalize those studied by (Niv *et al.*, 2007). Their tasks implied that there would be an opportunity cost for the passage of time; in our tasks, there was also a direct instrumental effect of speed. Further, we considered the consequences of only stochastically successful lever pressing. Finally, we suggested how the resulting rather complex collection of appetitive and aversive prediction errors might be represented by tonic, quasi-tonic and phasic aspects of dopaminergic and serotonergic neuromodulation.

The starting point for these models was analysis suggesting that, in appetitive circumstances, vigour should depend on the average rate of utility (Niv *et al.*, 2007). We have seen that the picture is significantly richer in the case of variable interval schedules for reward or stochastic punishment times. As one might expect, when rewards or punishments are postponed ($T \gg 0$) and acting quickly is costly, latencies are long. However, if actions can fail, then it may be worth trying to act more hastily, in order to preserve the possibility of trying again relatively cheaply. In the case of reward, in the unrealistic case that there is no unit utility of an action, but only a vigour utility that decreases with latency, an action will always ultimately be emitted. However, for the case of punishment, it may be better to suffer the consequences than to act sufficiently quickly to avert them.

For the case of active avoidance, the apparent paradox for the previous theory was that the average rate of utility is negative, and would therefore be incapable of determining vigour according to that theory's central result. In fact, this relationship is actually true – if acting quickly leads to the post-outcome state being reached more rapidly, it can also imply reaching the next pre-outcome state more quickly, and thus an earlier chance of future punishment. Thus, there is indeed an opportunity benefit for sloth that should make subjects slow down. This is, however, balanced against the immediate instrumental benefit for the action, which can dominate. This amounts to an interesting form of instrumental approach–avoidance conflict, where speedy approach itself is engendered by a need to be sure to avoid the immediate punishment, whereas sloth (which can be seen as a form of avoidance) comes from the need to avoid subsequent punishments. If rewards were also available (e.g. at the post-outcome state), then this would have the effect of increasing the average rate of utility, and thus decreasing the latency of any avoidance action through the contribution of the term $-\rho^*\tau$ to the state-action values. However, if different outcomes were indeed mixed, it would be important to take account of the surprising lability and adaptation in their subjective utilities (Vlaev *et al.*, 2009).

Unfortunately, I am not aware of any existing paradigms that systematically explore the effect on reaction times in an active avoidance task of either changing the probability of success of each action or manipulating the inter-trial interval in such a way that any benefit of sloth would be apparent. They would, however, be straightforward to test, at least in human subjects. It would be interesting to look for the rather detailed signatures of optimal behaviour evident in the effects in Figs 6 and 8 on successive lever press latencies of the success probability v.

The various contingencies in the active avoidance task lead to a varied collection of prediction errors that evolve in regular and revealing ways over the course of each trial. This is even true after learning is complete, particularly when lever pressing is only partially successful. In assigning parts of the prediction errors to various facets of the activity and concentration of dopamine and serotonin, we have been inspired by a wide range of evidence, notably on appetitive aspects of phasic dopamine (Montague *et al.*, 1996; Schultz *et al.*, 1997), vigour associations of tonic dopamine in the context of reward (Salamone & Correa, 2002; Niv *et al.*, 2007), and two-factor theories of active avoidance (Mowrer, 1947; Moutoussis *et al.*, 2008; Maia, 2010), but have nevertheless had to make some significant speculations, particularly about opponency between serotonin and dopamine (Deakin, 1983; Deakin & Graeff, 1991; Daw *et al.*, 2002; Cools *et al.*, 2008, 2011; Dayan & Huys, 2009; Boureau & Dayan, 2011). Even then, we have passed over a huge wealth of issues discussed in these reviews, including other influential views of serotonin, such as its potential role in temporal discounting (Tanaka *et al.*, 2004, 2007; Doya, 2008; Schweighofer *et al.*, 2008).

One constraint on this speculation comes from richer analyses of the relationship between Pavlovian and instrumental influences on action (Breland & Breland, 1961; Panksepp, 1998; Dickinson & Balleine, 2002; Dayan *et al.*, 2006b; Crockett *et al.*, 2009; Boureau & Dayan, 2011; Cools *et al.*, 2011). Crudely, stimuli associated with rewards lead to vigorous approach and engagement, at least partly under the influence of dopamine in the nucleus accumbens (Ikemoto & Panksepp, 1999; Lex & Hauber, 2008). Conversely, stimuli associated with punishments lead to a rich set of species-specific defensive actions (Bolles, 1970) that notably include forms of behavioural inhibition influenced by serotonin (Deakin, 1983; Soubrié, 1986; Deakin & Graeff, 1991; Graeff, 2002; Cools *et al.*, 2008; Tops *et al.*, 2009). Active defensive responses have themselves been associated with the accumbens (Reynolds & Berridge, 2002), and dopamine is also known to play a role (Faure *et al.*, 2008). One might also see the competition between ergotropic (work-directed, energy-expending) and trophotropic (internally directed, energy-conserving) behaviours in similar opponent terms, with the latter under serotonergic influence (Ellison, 1979; Handley & McBlane, 1991; Tops *et al.*, 2009; Boureau & Dayan, 2011).

These appetitive and aversive responses are Pavlovian in the sense that they are emitted without regard to their actual consequences. In the appetitive case, even if rewards are not delivered in trials on which engagement and approach occur, then subjects cannot in general help themselves, but still engage and approach, at least to some extent (Sheffield, 1965; Williams & Williams, 1969). Concomitantly, in the aversive case of active avoidance, this would imply that behavioural inhibition engendered by the prediction of punishment would work against the subjects' need to act quickly to avoid the punishment (hence the investigations in (Crockett *et al.*, 2009; Guitart-Masip *et al.*, 2011b), with the latter suggesting that conventional prediction errors are only seen in the face of a potential requirement to act). One can interpret two-factor theories of avoidance (Mowrer, 1947; Johnson *et al.*, 2002; Moutoussis *et al.*, 2008; Maia, 2010) as suggesting the resolution to this problem of measuring the worth of outcomes against the prospect of punishment. That is, achieving safety becomes a surrogate reward.

To put this another way, phasic dopamine could report on all improvements in state, whether these are based on the potential for gaining reward or the potential for avoiding punishment. Likewise, tonic dopamine could report on the appetitive utility rate associated either with reward or with the potential for avoiding punishment. In the aversive cases, these signals are relative to a baseline provided by phasic and tonic serotonin. The result, as pictured in Fig. 11, can be seen as partly reconciling the original theory of dopamine–serotonin opponency, which suggested that tonic dopamine might report average rate of punishment (Daw *et al.*, 2002), with the more recent suggestion that it might report the average rate of reward (Niv *et al.*, 2007).

Figure 11 also goes beyond (Niv *et al.*, 2007) in assigning the time-dependent component of the utility of the lever pressing to a quasi-tonic serotonergic signal. The assumption is that this is integrated over time – for instance, by receptors or even in the tissue – with its integral forming a central component of the prediction error. There is recent evidence for the release of serotonin while animals had to wait during a delay period (Miyazaki *et al.*, 2011b), which could be the consequence of such quasi-tonic activity. However, we should note that the activity of neurons in the raphe shows a range of responses during delays in reward-related tasks, including both activation and suppression (Nakamura *et al.*, 2008; Ranade & Mainen, 2009; Bromberg-Martin *et al.*, 2010; Miyazaki *et al.*, 2011a). In particular, (Bromberg-Martin *et al.*, 2010) note that the raising or lowering of what we could consider to be quasi-tonic activity of raphe neurons during a delay period was correlated with their excitation to a subsequent reward or non-reward; however, as it is impossible to know which, if any, of these neurons were serotonergic (Allers & Sharp, 2003; Kocsis *et al.*, 2006), and frank punishments were not employed, it is hard to draw firm conclusions with respect to our task.

The role of these baselines raises a structural question about the framing of a task. If there are both rewards and punishments, should the baseline be associated with the worst possible punishment, the most likely punishment, the average negative reinforcement per recurrent cycle, or some other quantity? How should controllability influence the baseline (Huys & Dayan, 2009) – should only punishments that can be avoided be able to influence it? In fact, this would make sense of the transient behaviour of dopamine in learned helplessness, as measured by microdialysis (Bland *et al.*, 2003). Given results on the dramatic effects of re-framing tasks (Tversky & Kahneman, 1981), it seems likely that there is no single answer to this question – it will depend on context and circumstance. This could therefore help to explain some of the apparent inconsistencies referred to above concerning the involvement of dopamine and serotonin in aversive and appetitive cases. However, it does not resolve other questions, such as the distinct groups of dopamine neurons. Certainly, framing-like manipulations of the baseline might have a powerful effect on the way in which Pavlovian mechanisms might offer a form of crutch for instrumental behaviour (Dayan & Huys, 2008) – that is, by engaging dopamine or serotonin appropriately, vigour and sloth could be automatic rather than having to be learnt. Of course, such Pavlovian influences might complicate our predictions, for instance if they took effect after rather than before the selection of instrumentally optimal behaviour.

Another set of paradigms that could be explored with a variant of the framework described here is differential reinforcement of low rates of responding (DRL) (Kramer & Rilling, 1970), which is the dual of active avoidance. In DRL, animals are typically rewarded provided that they can refrain for a given period from executing the action that will lead to the reward. Given timing uncertainty, there will be an optimal action latency; we could seek to model dopaminergic and serotonergic activity associated with the wait and the ultimate execution. Tasks such as those used by (Nakamura *et al.*, 2008; Ranade & Mainen, 2009; Miyazaki *et al.*, 2011a,b) to examine serotonin release and the activity of raphe neurons can involve delay periods, making this an attractive target. Serotonin depletion is known to affect the ability to wait that these paradigms require (Eagle *et al.*, 2009); it also impairs performance in DRL (Wogar *et al.*, 1992; Fletcher, 1995).

One issue that the prediction errors of Fig. 11 are likely to get wrong is that, apart from at the onset of a trial, the phasic dopaminergic (red bars) and serotonergic (blue bars) never occur at the same time. A transition to safety from a successful press leads to a phasic dopamine signal; a failed press or actual punishment leads to a phasic serotonin signal. However, when less reward is delivered than expected, there is ample evidence that the activity of dopamine cells dips below baseline, and substantial theories resulting from this explaining asymmetric findings about learning in patients with dopamine deficiency (Frank *et al.*, 2004; Frank, 2005) or normal subjects with different genetic endowments concerned with dopamine processing (Frank *et al.*, 2007; Frank & Hutchison, 2009). This suggests that there might be an additional baseline.

We only studied optimal behaviour, and the prediction errors that would result from the associated optimal values of states. Most aspects of behaviour change continuously with small changes to the parameters, and so would change continuously with small approximations to optimality. Some (e.g. the exact number of unsuccessful lever presses before a punishment) would be more sensitive. The tasks do certainly pose a panoply of learning problems. (Niv, 2008) described a form of temporal difference-based learning for acquiring responding for the case of distribution d^0, and this should also work for the cases that we have studied here. Exploring different latencies would, of course, be important to find optima; one might discover that a (possibly Pavlovian-induced) reluctance to sample long latencies in the face of impending shocks could make for a systematic bias shortening of the latencies as compared with the optimum. Exploration would also complicate the predictions about neuromodulatory coding, as there is evidence of phasic activation of dopamine neurons reporting forms of bonuses associated with exploration (Kakade & Dayan, 2002).

Our tasks place a potentially substantial burden on the representations that the subjects form about the task, notably the partial observability in the case of reward with stochastically successful lever pressing, and disentangling externally imposed randomness in timing from noise in internal timing mechanisms. Of course, the source of noise in timing need not matter for the predictions that we can make about behaviour. It would certainly be important to explore the effects of different sorts of suboptimality, such as noisy temporal choices, particularly in any attempt to fit human or animal decision-making.

One issue that emerges from consideration of dopaminergic manipulation in active avoidance and, indeed, the escape from fear paradigm (Moutoussis *et al.*, 2008) is that there is an asymmetry between the roles of dopamine and serotonin in aversion, with the former playing a central role in both value and action learning, but the latter only in the acquisition of values. One rationale for this is that when there are very many possible actions, finding out that executing one led to punishment is very weak information about what to do, whereas finding out that it led to reward, or to safety, is very strong information (Dayan & Huys, 2009).

Of course, there are known to be multiple neural systems controlling action (Dickinson & Balleine, 2002; Daw *et al.*, 2005; Daw & Doya, 2006; Samejima & Doya, 2007). One of these systems, implicated in model-free reinforcement learning or habitual responding, has historically been closely associated with the neuromodulators and regions such as the dorsolateral striatum (Killcross & Coutureau, 2003; Balleine, 2005). (Niv, 2008)'s temporal difference learning method associated with d^0 is model-free in this sense. Conversely, a second system, implicated in model-based reinforcement learning or goal-directed behaviour, has been associated with prefrontal processing and the dorsomedial striatum. These systems are believed to cooperate and compete in ways that are only just starting to be investigated (Gläscher *et al.*, 2010; Daw *et al.*, 2011; Simon & Daw, 2011). The exact nature of their interaction could have an important bearing on the predictions that we might make for the behaviour of neuromodulatory systems in the tasks that we describe here, if different parts of the signals arise from different systems.

One important direction in which the model requires improvement to be more directly testable is in its conceptualization of action. Along with (Niv *et al.*, 2007), we assumed that each lever press may be significantly costly, and so will be spared if possible. This may be the case for some actions, but certainly not all. An alternative suggested in the literature would be to have the subjects switch between two different rates of lever pressing (Gallistel *et al.*, 2004), or indeed more complex patterns of responding (Eldar *et al.*, 2011); it would certainly be possible to adapt our analysis accordingly. There are very many detailed studies of operant timing [from seminal work such as (Catania & Reynolds, 1968) to modern analyses such as (Williams *et al.*, 2009a,b,c)], which would be attractive targets. There are also structurally different schedules of reward or punishment that do not fit quite so easily into the current scheme, such as the extended working time required in (Arvanitogiannis & Shizgal, 2008; Breton *et al.*, 2009), but that would also be of great interest to capture.

In conclusion, even the simplest of active avoidance tasks poses a set of fascinating and powerful constraints on the nature of optimal behaviour, and the forms of prediction error that arise. These results, and the predictions and speculations with which they are associated, take us nicely beyond conventional neural reinforcement learning.

Acknowledgements

I am very grateful to my collaborators Y-Lan Boureau, Roshan Cools, Nathaniel Daw, Ray Dolan, Emrah Düzel, Marc Guitart, Quentin Huys, Yael Niv and Ben Seymour for their inspiration for and interaction with this work. Y-Lan Boureau, Roshan Cools, Nathaniel Daw, Marc Guitart and Quentin Huys and two anonymous reviewers provided very helpful comments on an earlier version. Funding was from the Gatsby Charitable Foundation.

Abbreviations

DRL, differential reinforcement of low rates of responding; SMDP, semi-Markov decision process.

References

Allers, K.A. & Sharp, T. (2003) Neurochemical and anatomical identification of fast- and slow-firing neurones in the rat dorsal raphe nucleus using juxtacellular labelling methods in vivo *Neuroscience*, **122**, 193–204.

Arvanitogiannis, A. & Shizgal, P. (2008) The reinforcement mountain: allocation of behavior as a function of the rate and intensity of rewarding brain stimulation *Behav. Neurosci.*, **122**, 1126–1138.

Balleine, B.W. (2005) Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits *Physiol. Behav.*, **86**, 717–730.

Beninger, R., Mason, S., Phillips, A. & Fibiger, H. (1980) The use of extinction to investigate the nature of neuroleptic-induced avoidance deficits *Psychopharmacology*, **69**, 11–18.

Bland, S.T., Hargrave, D., Pepin, J.L., Amat, J., Watkins, L.R. & Maier, S.F. (2003) Stressor controllability modulates stress-induced dopamine and serotonin efflux and morphine-induced serotonin efflux in the medial prefrontal cortex *Neuropsychopharmacology*, **28**, 1589–1596.

Bolles, R.C. (1970) Species-specific defense reactions and avoidance learning *Psychol. Rev.*, **77**, 32–48.

Boureau, Y.-L. & Dayan, P. (2011) Opponency revisited: competition and cooperation between dopamine and serotonin *Neuropsychopharmacology*, **36**, 74–97.

Breland, K. & Breland, M. (1961) The misbehavior of organisms *Am. Psychol.*, **16**, 681–684.

Breton, Y.-A., Marcus, J.C. & Shizgal, P. (2009) Rattus psychologicus: construction of preferences by self-stimulating rats *Behav. Brain Res.*, **202**, 77–91.

Brischoux, F., Chakraborty, S., Brierley, D.I. & Ungless, M.A. (2009) Phasic excitation of dopamine neurons in ventral vta by noxious stimuli *Proc. Natl. Acad. Sci. USA*, **106**, 4894–4899.

Bromberg-Martin, E.S., Hikosaka, O. & Nakamura, K. (2010) Coding of task reward value in the dorsal raphe nucleus *J. Neurosci.*, **30**, 6262–6272.

Catania, A.C. & Reynolds, G.S. (1968) A quantitative analysis of the responding maintained by interval schedules of reinforcement *J. Exp. Anal. Behav.*, **11**(3 Suppl), 327–383.

Cools, R., Roberts, A.C. & Robbins, T.W. (2008) Serotonergic regulation of emotional and behavioural control processes *Trends Cogn. Sci.*, **12**, 31–40.

Cools, R., Nakamura, K. & Daw, N.D. (2011) Serotonin and dopamine: unifying affective, activational, and decision functions *Neuropsychopharmacology*, **36**, 98–113.

Crockett, M.J., Clark, L. & Robbins, T.W. (2009) Reconciling the role of serotonin in behavioral inhibition and aversion: acute tryptophan depletion abolishes punishment-induced inhibition in humans *J. Neurosci.*, **29**, 11993–11999.

Daw, N. & Courville, A. (2008)The pigeon as particle filter In Platt, J.C., Koller, D., Singer, Y. & Roweis, S. (Eds), *Advances in Neural Information Processing Systems*, vol. 20. MIT Press, Cambridge, MA, pp. 369–376.

Daw, N.D. & Doya, K. (2006) The computational neurobiology of learning and reward *Curr. Opin. Neurobiol.*, **16**, 199–204.

Daw, N.D., Kakade, S. & Dayan, P. (2002) Opponent interactions between serotonin and dopamine *Neural Netw.*, **15**, 603–616.

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., Courville, A. & Touretzky, D. (2006) Representation and timing in theories of the dopamine system *Neural Comput.*, **18**, 1637–1677.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P. & Dolan, R.J. (2011) Model-based influences on humans' choices and striatal prediction errors *Neuron*, **69**, 1204–1215.

Dayan, P. & Huys, Q.J.M. (2008) Serotonin, inhibition, and negative mood *PLoS Comput. Biol.*, **4**, e4.

Dayan, P. & Huys, Q.J.M. (2009) Serotonin in affective control *Annu. Rev. Neurosci.*, **32**, 95–126.

Dayan, P., Niv, Y., Seymour, B. & Daw, N.D. (2006a) The misbehavior of value and the discipline of the will *Neural Netw.*, **19**, 1153–1160.

Dayan, P., Niv, Y., Seymour, B. & Daw, N.D. (2006b) The misbehavior of value and the discipline of the will *Neural Netw.*, **19**, 1153–1160.

Deakin, J.F.W. (1983) Roles of brain serotonergic neurons in escape, avoidance and other behaviors *J. Psychopharmacol.*, **43**, 563–577.

Deakin, J.F.W. & Graeff, F.G. (1991) 5-HT and mechanisms of defence *J. Psychopharmacol.*, **5**, 305–316.

Delgado, M.R., Li, J., Schiller, D. & Phelps, E.A. (2008) The role of the striatum in aversive learning and aversive prediction errors *Phil. Trans. R Soc. Lond. B Biol. Sci.*, **363**, 3787–3800.

Dickinson, A. & Balleine, B. (1994) Motivational control of goal-directed action *Learn. Behav.*, **22**, 1–18.

Dickinson, A. & Balleine, B. (2002) The role of learning in motivation In Gallistel, C., (Ed.), *Stevens' Handbook of Experimental Psychology, vol 3*, Wiley, New York, NY, pp. 497–533.

Dickinson, A., Smith, J. & Mirenowicz, J. (2000) Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists *Behav. Neurosci.*, **114**, 468–483.

Doya, K. (2008) Modulators of decision making *Nat. Neurosci.*, **11**, 410–416.

Duik, C., Botvinick, M., Barto, A.G. & Niv, Y. (2010)Hierarchical reinforcement learning: an FMRI study of learning in a two-level gambling task *Society for Neuroscience Abstracts*, vol. 36. *Society for Neuroscience*, San Diego, CA, p. 907.14.

Eagle, D.M., Lehmann, O., Theobald, D.E.H., Pena, Y., Zakaria, R., Ghosh, R., Dalley, J.W. & Robbins, T.W. (2009) Serotonin depletion impairs waiting but not stop-signal reaction time in rats: implications for theories of the role of 5-HT in behavioral inhibition *Neuropsychopharmacology*, **34**, 1311–1321.

Eldar, E., Morris, G. & Niv, Y. (2011) The effects of motivation on response rate: a hidden semi-Markov model analysis of behavioral dynamics *J. Neurosci. Methods*, **201**, 251–261.

Ellison, G. (1979)Chemical systems of the brain and evolution. In Oakley, D. & Plotkin, H. (Eds), *Brain, Behaviour and Evolution*. Methuen, London, pp. 78–98.

Estes, W. (1943) Discriminative conditioning. I. A discriminative property of conditioned anticipation *J. Exp. Psychol.*, **32**, 150–155.

Faure, A., Reynolds, S.M., Richard, J.M. & Berridge, K.C. (2008) Mesolimbic dopamine in desire and dread: enabling motivation to be generated by localized glutamate disruptions in nucleus accumbens *J. Neurosci.*, **28**, 7184–7192.

Fiorillo, C.D., Tobler, P.N. & Schultz, W. (2003) Discrete coding of reward probability and uncertainty by dopamine neurons *Science*, **299**, 1898–1902.

Fletcher, P.J. (1995) Effects of combined or separate 5,7-dihydroxytryptamine lesions of the dorsal and median raphe nuclei on responding maintained by a DRL 20s schedule of food reinforcement *Brain Res.*, **675**, 45–54.

Frank, M.J. (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism *J. Cogn. Neurosci.*, **17**, 51–72.

Frank, M.J. & Hutchison, K. (2009) Genetic contributions to avoidance-based decisions: striatal d2 receptor polymorphisms *Neuroscience*, **164**, 131–140.

Frank, M.J., Seeberger, L.C. & O'Reilly, R.C. (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism *Science*, **306**, 1940–1943.

Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T. & Hutchison, K.E. (2007) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning *Proc. Natl. Acad. Sci. USA*, **104**, 16311–16316.

Gallistel, C.R., Fairhurst, S. & Balsam, P. (2004) The learning curve: implications of a quantitative analysis *Proc. Natl. Acad. Sci. USA*, **101**, 13124–13131.

Gibbon, J. (1977) Scalar expectancy theory and Weber's law in animal timing *Psychol. Rev.*, **84**, 279–325.

Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J.P. (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning *Neuron*, **66**, 585–595.

Goto, Y. & Grace, A.A. (2005) Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior *Nat. Neurosci.*, **8**, 805–812.

Goto, Y., Otani, S. & Grace, A.A. (2007) The yin and yang of dopamine release: a new perspective *Neuropharmacology*, **53**, 583–587.

Graeff, F.G. (2002) On serotonin and experimental anxiety *Psychopharmacology*, **163**, 467–476.

Guitart-Masip, M., Beierholm, U.R., Dolan, R., Duzel, E. & Dayan, P. (2011a) Vigor in the face of fluctuating rates of reward: an experimental examination *J. Cogn. Neurosci*, **23**, 3933–3938.

Guitart-Masip, M., Fuentemilla, L., Bach, D.R., Huys, Q.J.M., Dayan, P., Dolan, R.J. & Duzel, E. (2011b) Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain *J. Neurosci.*, **31**, 7867–7875.

Handley, S. & McBlane, J. (1991) 5-HT: the disengaging transmitter *J. Psychopharmacol.*, **5**, 322–326.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., Imamizu, H. & Kawato, M. (2004) A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task *J. Neurosci.*, **24**, 1660–1665.

Huys, Q.J.M. & Dayan, P. (2009) A Bayesian formulation of behavioral control *Cognition*, **113**, 314–328.

Ikemoto, S. & Panksepp, J. (1999) The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking *Brain Res. Brain Res. Rev.*, **31**, 6–41.

Jacobs, B.L. & Fornal, C.A. (1997) Serotonin and motor activity *Curr. Opin. Neurobiol.*, **7**, 820–825.

Jacobs, B.L. & Fornal, C.A. (1999) Activity of serotonergic neurons in behaving animals *Neuropsychopharmacology*, **21**(2 Suppl), 9S–15S.

Jacobs, B.L., Martín-Cora, F.J. & Fornal, C.A. (2002) Activity of medullary serotonergic neurons in freely moving animals *Brain Res. Brain Res. Rev.*, **40**, 45–52.

Johnson, J., Li, W., Li, J. & Klopf, A. (2002) A computational model of learned avoidance behavior in a one-way avoidance experiment *Adapt. Behav.*, **9**, 91–104.

Joshua, M., Adler, A., Mitelman, R., Vaadia, E. & Bergman, H. (2008) Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials *J. Neurosci.*, **28**, 11673–11684.

Kakade, S. & Dayan, P. (2002) Dopamine: generalization and bonuses *Neural Netw.*, **15**, 549–559.

Killcross, S. & Coutureau, E. (2003) Coordination of actions and habits in the medial prefrontal cortex of rats *Cereb. Cortex*, **13**, 400–408.

Kocsis, B., Varga, V., Dahan, L. & Sik, A. (2006) Serotonergic neuron diversity: identification of raphe neurons with discharges time-locked to the hippocampal theta rhythm *Proc. Natl. Acad. Sci. USA*, **103**, 1059–1064.

Kramer, T. & Rilling, M. (1970) Differential reinforcement of low rates: a selective critique *Psychol. Bull.*, **74**, 225–254.

Lammel, S., Ion, D.I., Roeper, J. & Malenka, R.C. (2011) Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli *Neuron*, **70**, 855–862.

Lex, A. & Hauber, W. (2008) Dopamine D1 and D2 receptors in the nucleus accumbens core and shell mediate Pavlovian–instrumental transfer *Learn. Mem.*, **15**, 483–491.

Lovibond, P.F. (1981) Appetitive Pavlovian–instrumental interactions: effects of inter-stimulus interval and baseline reinforcement conditions *Q. J. Exp. Psychol. B*, **33**(Pt 4), 257–269.

Lovibond, P.F. (1983) Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus *J. Exp. Psychol. Anim. Behav. Process.*, **9**, 225–247.

Lowry, C.A. (2002) Functional subsets of serotonergic neurones: implications for control of the hypothalamic-pituitary-adrenal axis *J. Neuroendocrinol.*, **14**, 911–923.

Mahadevan, S. (1996) Average reward reinforcement learning: foundations, algorithms, and empirical results *Machine Learn.*, **22**, 159–195.

Maia, T.V. (2010) Two-factor theory, the actor-critic model, and conditioned avoidance *Learn. Behav.*, **38**, 50–67.

Matsumoto, M. & Hikosaka, O. (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals *Nature*, **459**, 837–841.

McClure, S.M., Berns, G.S. & Montague, P.R. (2003) Temporal prediction errors in a passive learning task activate human striatum *Neuron*, **38**, 339–346.

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

Miyazaki, K., Miyazaki, K.W. & Doya, K. (2011a) Activation of dorsal raphe serotonin neurons underlies waiting for delayed rewards *J. Neurosci.*, **31**, 469–479.

Miyazaki, K.W., Miyazaki, K. & Doya, K. (2011b) Activation of the central serotonergic system in response to delayed but not omitted rewards *Eur. J. Neurosci.*, **33**, 153–160.

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.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. (2006) Midbrain dopamine neurons encode decisions for future action *Nat. Neurosci.*, **9**, 1057–1063.

Moutoussis, M., Bentall, R.P., Williams, J. & Dayan, P. (2008) A temporal difference account of avoidance learning *Network*, **19**, 137–160.

Mowrer, O. (1947) On the dual nature of learning: a reinterpretation of conditioning and problem-solving *Harvard Educ. Rev.*, **17**, 102–150.

Nakamura, K., Matsumoto, M. & Hikosaka, O. (2008) Reward-dependent modulation of neuronal activity in the primate dorsal raphe nucleus *J. Neurosci.*, **28**, 5331–5343.

Nicola, S.M. (2010) The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior *J. Neurosci.*, **30**, 16585–16600.

Niv, Y. (2008) *The Effects of Motivation on Habitual Instrumental Behavior*. PhD thesis, ICNC, Hebrew University of Jerusalem, Jerusalem, Israel.

Niv, Y., Daw, N. & Dayan, P. (2005)How fast to work: response vigor, motivation and tonic dopamine. In Weiss, Y., Scholkopf, B. & Platt, J. (Eds), *Advances in Neural Information Processing*. MIT Press, Cambridge, MA, pp. 1019–1026.

Niv, Y., Daw, N.D., Joel, D. & Dayan, P. (2007) Tonic dopamine: opportunity costs and the control of response vigor *Psychopharmacology*, **191**, 507–520.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. (2003) Temporal difference models and reward-related learning in the human brain *Neuron*, **38**, 329–337.

Panksepp, J. (1998) *Affective Neuroscience*. OUP, New York, NY.

Puterman, M.L. (2005) *Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics)*. Wiley-Interscience, Hoboken, NJ.

Ranade, S.P. & Mainen, Z.F. (2009) Transient firing of dorsal raphe neurons encodes diverse and specific sensory, motor, and reward events *J. Neurophysiol.*, **102**, 3026–3037.

Rescorla, R.A. & Solomon, R.L. (1967) Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning *Psychol. Rev.*, **74**, 151–182.

Reynolds, S.M. & Berridge, K.C. (2002) Positive and negative motivation in nucleus accumbens shell: bivalent rostrocaudal gradients for GABA-elicited eating, taste 'liking'/'disliking' reactions, place preference/avoidance, and fear *J. Neurosci.*, **22**, 7308–7320.

Roesch, M.R., Calu, D.J. & Schoenbaum, G. (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards *Nat. Neurosci.*, **10**, 1615–1624.

Rummery, G. & Niranjan, M. (1994) On-line *Q*-learning using connectionist systemsTechnical Report CUED/F-INENG/TR 166, Cambridge University Engineering Department.

Salamone, J.D. & Correa, M. (2002) Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine *Behav. Brain Res.*, **137**, 3–25.

Samejima, K. & Doya, K. (2007) Multiple representations of belief states and action values in corticobasal ganglia loops *Ann. N. Y. Acad. Sci.*, **1104**, 213–228.

Schultz, W. & Dickinson, A. (2000) Neuronal coding of prediction errors *Annu. Rev. Neurosci.*, **23**, 473–500.

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

Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S.C., Yamawaki, S. & Doya, K. (2008) Low-serotonin levels increase delayed reward discounting in humans *J. Neurosci.*, **28**, 4528–4532.

Schweimer, J.V. & Ungless, M.A. (2010) Phasic responses in dorsal raphe serotonin neurons to noxious stimuli *Neuroscience*, **171**, 1209–1215.

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J. & Frackowiak, R.S. (2004) Temporal difference models describe higher-order learning in humans *Nature*, **429**, 664–667.

Seymour, B., Daw, N., Dayan, P., Singer, T. & Dolan, R. (2007) Differential encoding of losses and gains in the human striatum *J. Neurosci.*, **27**, 4826–4831.

Sheffield, F. (1965)Relation between classical conditioning and instrumental learning In Prokasy, W. (Eds), *Classical Conditioning*. Appelton-Century-Crofts, New York, NY, pp. 302–322.

Simon, D.A. & Daw, N.D. (2011) Neural correlates of forward planning in a spatial decision task in humans *J. Neurosci.*, **31**, 5526–5539.

Smith, J. & Dickinson, A. (1998) The dopamine antagonist, pimozide, abolishes Pavlovian–instrumental transfer *J. Psychopharmacol.*, **12**, A6.

Soubrié, P. (1986) Reconciling the role of central serotonin neurons in human and animal behaviour *Behav. Brain Sci.*, **9**, 319–364.

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

Sutton, R.S. & Barto, A.G. (1998) *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Sutton, R., Precup, D. & Singh, S. (1999) Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning *Artific. Intell.*, **112**, 181–211.

Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y. & Yamawaki, S. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops *Nat. Neurosci.*, **7**, 887–893.

Tanaka, S.C., Schweighofer, N., Asahi, S., Shishida, K., Okamoto, Y., Yamawaki, S. & Doya, K. (2007) Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum *PLoS ONE*, **2**, e1333.

Tobler, P.N., Fiorillo, C.D. & Schultz, W. (2005) Adaptive coding of reward value by dopamine neurons *Science*, **307**, 1642–1645.

Tom, S.M., Fox, C.R., Trepel, C. & Poldrack, R.A. (2007) The neural basis of loss aversion in decision-making under risk *Science*, **315**, 515–518.

Tops, M., Russo, S., Boksem, M.A.S. & Tucker, D.M. (2009) Serotonin: modulator of a drive to withdraw *Brain Cogn.*, **71**, 427–436.

Tversky, A. & Kahneman, D. (1981) The framing of decisions and the psychology of choice *Science*, **211**, 453–458.

Vlaev, I., Seymour, B., Dolan, R.J. & Chater, N. (2009) The price of pain and the value of suffering *Psychol. Sci.*, **20**, 309–317.

Watkins, C. (1989) *Learning from Delayed Rewards*, PhD thesis. University of Cambridge, Cambridge.

Williams, D.R. & Williams, H. (1969) Auto-maintenance in the pigeon: sustained pecking despite contingent non-reinforcement *J. Exp. Anal. Behav.*, **12**, 511–520.

Williams, J., Sagvolden, G., Taylor, E. & Sagvolden, T. (2009a) Dynamic behavioural changes in the spontaneously hyperactive rat: 1. Control by place, timing, and reinforcement rate *Behav. Brain Res.*, **198**, 273–282.

Williams, J., Sagvolden, G., Taylor, E. & Sagvolden, T. (2009b) Dynamic behavioural changes in the spontaneously hyperactive rat: 2. Control by novelty *Behav. Brain Res.*, **198**, 283–290.

Williams, J., Sagvolden, G., Taylor, E. & Sagvolden, T. (2009c) Dynamic behavioural changes in the spontaneously hyperactive rat: 3. Control by reinforcer rate changes and predictability *Behav. Brain Res.*, **198**, 291–297.

Wogar, M.A., Bradshaw, C.M. & Szabadi, E. (1992) Impaired acquisition of temporal differentiation performance following lesions of the ascending 5-hydroxytryptaminergic pathways *Psychopharmacology*, **107**, 373–378.

Wyvell, C.L. & Berridge, K.C. (2001) Incentive sensitization by previous amphetamine exposurencreased cue-triggered 'wanting' for sucrose reward *J. Neurosci.*, **21**, 7831–7840.