Vigor in the Face of Fluctuating Rates of Reward: An Experimental Examination

Marc Guitart-Masip*, Ulrik R. Beierholm*, Raymond Dolan, Emrah Duzel, and Peter Dayan

Abstract

■ Two fundamental questions underlie the expression of behavior, namely what to do and how vigorously to do it. The former is the topic of an overwhelming wealth of theoretical and empirical work particularly in the fields of reinforcement learning and decision-making, with various forms of affective prediction error playing key roles. Although vigor concerns motivation, and so is the subject of many empirical studies in diverse fields, it has suffered a dearth of computational models. Recently, Niv et al. [Niv, Y., Daw, N. D., Joel, D., & Dayan, P. Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology (Berlin), 191,* 507–520, 2007] suggested that vigor should be controlled by the opportunity cost of time, which is itself determined by the average rate of reward. This coupling of reward rate and vigor can be shown to be optimal under the theory of average return reinforcement learning for a particular class of tasks but may also be a more general, perhaps hard-wired, characteristic of the architecture of control. We, therefore, tested the hypothesis that healthy human participants would adjust their RTs on the basis of the average rate of reward. We measured RTs in an odd-ball discrimination task for rewards whose magnitudes varied slowly but systematically. Linear regression on the subjects' individual RTs using the time varying average rate of reward as the regressor of interest, and including nuisance regressors such as the immediate reward in a round and in the preceding round, showed that a significant fraction of the variance in subjects' RTs could indeed be explained by the rate of experienced reward. This validates one of the key proposals associated with the model, illuminating an apparently mandatory form of coupling that may involve tonic levels of dopamine. ■

INTRODUCTION

Maximizing rewards and minimizing punishments requires choosing the best action among the set of available options. Reinforcement learning (Sutton & Barto, 1998) offers ways of formalizing this process that resonate closely with the psychology and the neuroscience of decision-making (Daw & Doya, 2006; Montague & Berns, 2002; Montague, Dayan, & Sejnowski, 1996). In particular, phasic responses of macaque and rodent midbrain dopaminergic neurons to rewards and reward-associated stimuli are akin to the reward prediction error signal from reinforcement learning (Roesch, Calu, & Schoenbaum, 2007; Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006; Bayer & Glimcher, 2005; Schultz, Dayan, & Montague, 1997). Moreover, abundant fMRI studies show that BOLD signal in the striatum, a major target of the dopaminergic system, correlates with the prediction error signals derived from fitting subject's behavior with reinforcement learning models (Rutledge, Dean, Caplin, & Glimcher, 2010; O'Doherty et al., 2004; McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). However, until recently, reinforcement learning models overlooked the important behavioral observation that animals systematically vary the vigor with which they execute their near optimal choices (Phillips, Walton, & Jhou, 2007; Salamone & Correa, 2002). This omission is especially troubling considering the substantial data implicating the dopaminergic system in different aspects of response vigor (Lex & Hauber, 2008; Parkinson et al., 2002; Salamone & Correa, 2002; Taylor & Robbins, 1986; Langston, Forno, Rebert, & Irwin, 1984; Ungerstedt, 1971).

Niv, Daw, Joel, and Dayan (2007) recently developed an reinforcement learning model in which the vigor (defined as the inverse latency) of action can be optimized. The model realizes a trade-off between two costs: one stemming from the harder work assumed necessary to emit faster actions and the other from the opportunity cost inherent in acting more slowly. The latter arises from the delay that results to the next reward and, indeed, all subsequent rewards. Niv et al. (2007) suggested that agents should choose latencies (and actions) to maximize the rate of accumulated reward per unit time and showed that the resulting optimal latencies would be inversely proportional to the average reward rate. On the basis of existing experimental evidence, Niv et al. (2007) proposed that tonic levels of dopamine report the average rate of reward and, thus, tied together prediction error (McClure et al., 2003; Schultz et al., 1997; Montague et al., 1996), incentive salience (Berridge & Robinson, 1998), and invigoration (Salamone & Correa, 2002) theories of dopamine. Furthermore, Cools, Nakamura, and Daw (2011) recently formulated an integrative model of opponency between dopamine and serotonin, which has at its

University College London
*These authors contributed equally to this study.

heart the average rate of reward and the opportunity cost of time. In paradigms such as Pavlovian instrumental transfer, vigor appears to be at least partially under mandatory Pavlovian rather than wholly instrumental control, and so we, therefore, hypothesized that healthy human volunteers would adjust their response vigor on the basis of estimates of the average reward rate, irrespective of any instrumentality.

Here, we tested this hypothesis using a novel variant of a monetary incentive delay task (Adcock, Thangavel, Whitfield-Gabrieli, Knutson, & Gabrieli, 2006; Knutson, Adams, Fong, & Hommer, 2001). We induced changes in the average reward rate by varying the rewards offered on each trial and studied how the ensuing RTs varied. We deliberately made the task simple to avoid any issues associated with a speed–accuracy trade-off. Following conventional practice (Sutton & Barto, 1998), Niv et al. (2007) suggested that subjects might estimate the average reward rate using the delta or Rescorla–Wagner rule. Consequently, we also tested whether the dependence of RT on recent past rewards was consistent with the operation of such a rule.

METHODS

Subjects and Behavioral Paradigm

Thirty-nine subjects were recruited from the University College London Psychology Department's recruitment pool, received full written instructions, and provided written consent in accordance with the University College London Research Ethics Committee. The experiment used a regular PC monitor and keyboard. The layout of a trial is depicted in Figure 1A. At the beginning of each round, subjects were presented visually with a number representing the potential payout of that round R_t, in the range of 1–100 pence. The potential payouts, R_t, were varied across trials according to a prespecified function that was fixed across subjects and designed to vary over time in a way that was minimally correlated with other potential variables. The potential payout function used is displayed in Figure 1B. After a variable period (750–1250 msec), subjects were shown three visual figures and had to indicate the "odd one out" by pressing a button. For a trial to be counted as successful, subjects had to respond within 500 msec by pressing the button corresponding to the deviant stimulus. In 20% of the trials, this time constraint was lowered to 400 msec to ensure that the task would lead to unexpected misses and, therefore, to keep the participants engaged throughout the whole task. After being shown a blank screen for 500 msec, subjects were informed as to the success of the trial. This feedback was followed by another blank screen and the beginning of the next round.

Subjects performed as many trials as they could within the time limit of 27 min; this varied from 383 to 467 trials. At the end of the study, 10% of the trials were chosen randomly, and subjects were paid the sum of the value of the successful subset of those trials, plus a fixed show up fee of £5.

Data Analysis

We fitted a log normal distribution to each individual's RTs and, thus, a set of associated z scores. We were, therefore, able to study individual and average RTs using a common analysis method we describe below. Missed trials (trials without any behavioral response) were not included in the analysis. For the averaged data, we did not analyze any trials after trial number 400, as the number of trials completed varied across subjects. For both types of analysis, we ignored the first 20 trials to allow subjects to get used to the task.

Given the log-normalized data, we performed a linear regression on the subject RTs using the following regressors:

R_t: available reward for the subjects to win in a given round.

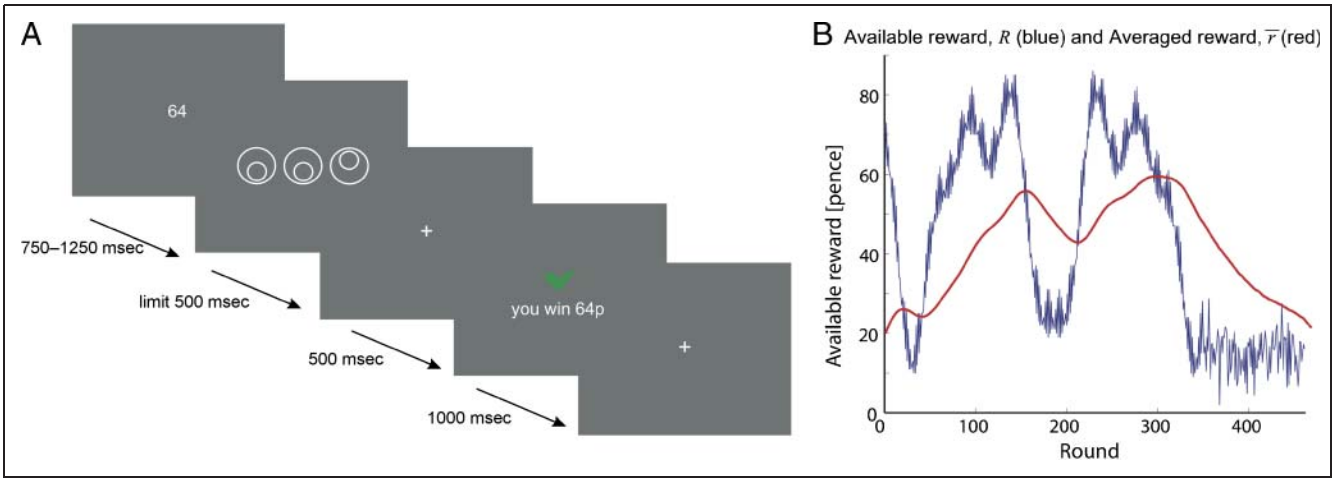

Figure 1. (A) Structure of one round of the behavioral task. Subjects are shown their potential reward, followed by an odd-one-out task to be completed within 500 msec (400 msec for 20% of trials). After a further 500 msec, they were shown their received reward. (B) The induced fluctuation in available reward (blue) and averaged reward (red) varying over time.

r_{t-1}: experienced reward on the previous trial.

$\overline{r}_t$: averaged reward signal, as given by $\overline{r}_t = \overline{r}_{t-1} + \alpha(R_t - \overline{r}_{t-1})$. The update rate for the average reward α was set at 0.012. This was determined by fitting this free parameter to maximize the amount of variance explained by the full regression model using data from a previous pilot study with an identical task. As is standard in learning approaches to average reward reinforcement learning, we used a simple Rescorla–Wagner rule to update the average reward rate.

Repetition of stimulus: binary vector indicating whether the stimulus in the last round was the same as in this round.

Linear: linear function.

Too late: binary return indicating whether the response was too late in the previous round.

Intertrial interval: pretrial interval while waiting for the stimulus to be presented.

A constant term.

The available reward R_t, the immediately experienced reward r_{t-1}, and the averaged reward $\overline{r}_t$ were our regressors of interest (see Figure 1B); the other regressors were included as nuisance variables.

These six regressors, together with the constant term, comprised the input data, X, in the linear regression:

$$\log \vec{R}T = \overline{\overline{X}} \times \vec{\beta} + \vec{\varepsilon},$$

which was performed using standard matrix inversion (MATLAB, Mathworks, Natick, MA). As per standard techniques, we examined which regressors explained a significant amount of the variance in the average subject data as well as a significant amount of the variance in the individual subject data for a significant number of subjects.

RESULTS

Subjects performed an "odd-one-out" task with a response being required within 500 msec (400 msec in 20% of the trials, randomly chosen) for the chance of receiving a monetary reward whose magnitude was announced at the beginning of each trial (Cools et al., 2005; Knutson et al., 2001; see Figure 1A). Following each response, subjects were informed whether they chose correctly and sufficiently quickly. As their eventual payout was directly related to performance, subjects had an incentive to be both fast and accurate. Subjects chose the correct response in 92.8% of trials (standard deviation of 5.5% across subjects; only three subjects had less than 85% correct), suggesting that there was no substantial speed–accuracy trade-off. On average, subjects made their choices within 416 msec (standard deviation of 24 msec across subject means, mean standard deviation of 48 msec) and acted too slowly in 19.1% of trials on average (standard deviation of 8.8). In total, subjects responded correctly and within the allocated time on 73.7% of the trials (with standard deviation of 10.3 across subjects).

To vary the perceived average reward rate, the potential reward (R_t on trial t) was changed across trials (see Figure 1B) in a pseudorandom way. Although all subjects were presented with the same sequence, their individual errors implied that each subject would have his or her own individual actual immediate reward, actual previous reward r_{t-1}, and actual average reward rate $\overline{r}_t$ (see Methods). To study the effect of these three quantities on subjects' RTs, we performed a linear regression on the logarithm of the RTs including R_t, r_{t-1}, and $\overline{r}_t$ as explanatory variables. We also included four nuisance regressors in the model to eliminate the effect of factors that might impact behavior but were orthogonal to our hypotheses. These comprised a binary variable that indicated that if the stimuli in the current round were identical to those in the previous round (to address response priming), a linear term (to address fatigue and/or training), a binary variable that indicated if the subject's response had been too slow in the previous round (which might hasten their current response), and a variable indicating the time between the start of the trial (available reward presented on the screen) and the presentation of the oddball (to allow for preparation). We performed this regression analysis on the individually z-scored log RTs, both for individual subjects, and averaged across subjects. For the averaged subjects, we used the average over the individual perceived reward rates for the regression.

When performing this regression on the average RTs (Figure 2A), we were able to explain 21.4% of the variance, finding that the average rate of reward ($t(37.9) = -3.44$), the repetition of stimulus ($t(37.9) = -5.75$), and the time to the oddball ($t(37.9) = -4.13$) also contributed significantly ($p < .05$) to the variance. Rather surprisingly neither the available reward, R_t, nor the immediately experienced reward, $r\_(t-1)$, significantly influenced the RTs ($t(37.9) = 0.16$ and $t(37.9) = -0.09$, respectively; see Figure 2B). The negative sign of the beta value for the average reward rate indicated that the regressor had a negative effect on the RTs, that is, causing subjects to speed up, as predicted by our original hypothesis.

We found similar results when performing the regression analysis on individual subjects' data, while explaining 7.7% of the variance on average (range of 1.6–22.2%). We performed a random effects two-tailed t test over beta values for each regressor across subjects. The beta values for the average reward rate ($t(38) = -4.68, p < .0001$), the repetition of stimulus ($t(38) = -7.24, p < .0001$), and the intertrial interval ($t(38) = -5.48, p < .0001$) regressors were all significantly different from zero (indicating their contribution towards the variance in the RTs), whereas the available reward, R_t, was not significantly different from zero ($t(38) = 1.13, p = .264$). Again, the negative sign for the beta value for the average reward rate indicated that increases in the average reward led to subjects increasing their speed, in accordance with the hypothesis derived from the model. Unlike the analysis of the averages across subjects, the immediate reward obtained in the previous

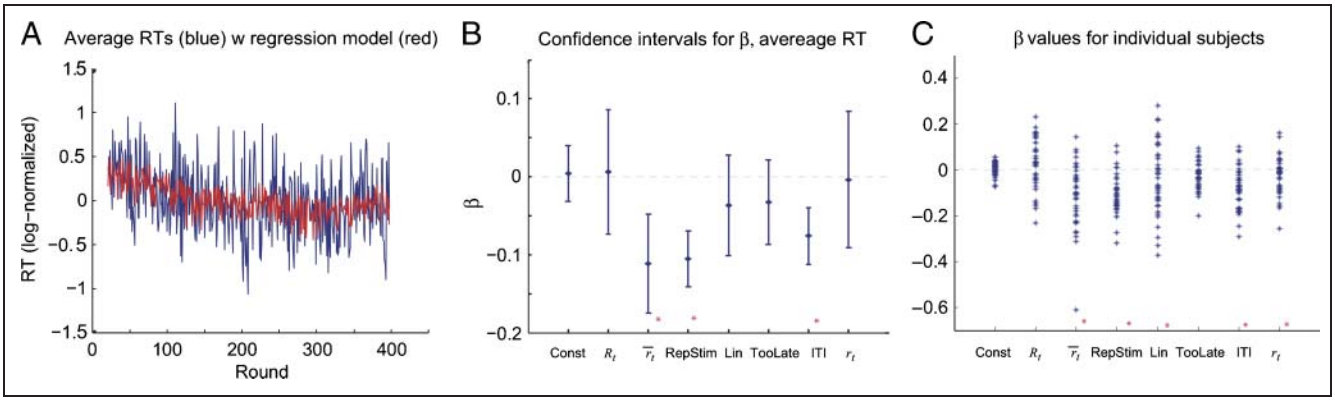

Figure 2. (A) The RTs after averaging across subjects (blue) with the fitted linear regression (red). (B) A linear regression on the average data across subjects yields beta values for each regressor. Confidence intervals ($p < .05$) are shown, with significant regressors indicated by a red asterisk (*). (C) A linear regression on the data for individual subjects yields beta values for each regressor. A two-tailed t test indicates which regressors are significant across subjects (red *).

trial $r\_(t - 1; t(38) = -2.38, p = .0226)$, and the too late regressor ($t(38) = -3.08, p = .0038$) also accounted significantly for more modest percentages of the variance in RTs across participants.

Finally, we considered whether subjects might have strategically slowed their responses on trials with large available rewards to optimize a speed–accuracy trade-off. Inconsistent with this possibility, we found that there was no significant correlation between the number of participants that performed correctly on a given trial with the available reward on that trial ($r = -0.056, p = .276$).

DISCUSSION

Exactly in line with our hypothesis, we showed that healthy human participants adjusted the vigor of motor behavior on the basis of an estimate of the average reward. This was seen both when the data were averaged across all participants and also when each participant's RT was analyzed individually.

It is well known that the RTs of humans and other animals (i.e., our definition of vigor) are influenced by the incentive motivational value of the goal toward which their actions are directed (e.g., Cools et al., 2005; Wittmann et al., 2005; Satoh, Nakai, Sato, & Kimura, 2003; Takikawa, Kawagoe, & Hikosaka, 2002; Watanabe et al., 2001). The evidence supporting an involvement of the midbrain dopaminergic system in the regulation of response vigor is also broad (Niv et al., 2007; Phillips et al., 2007; Satoh et al., 2003; Salamone & Correa, 2002; Berridge & Robinson, 1998). However, the computational basis of this process is rather less well studied (see Niv et al., 2007; Phillips et al., 2007; McClure et al., 2003). In the computational account of vigor suggested by Niv et al. (2007), vigor is proportional to the average rate of reinforcement.

The model of Niv et al. (2007) is based on a slightly different task, which lacks the immediate link between RT and reward that we employed here. That is, in the model, the only virtue of vigor is the opportunity cost of being slothful. However, in our task, subjects actually have to react quickly to win. It is straightforward to include the penalty of missing a reward because of too slow responses into the model, although we would then need to model more accurately the minimum possible RT. However, if anything, the addition of an imperative to respond quickly in our task should have strengthened the importance of the immediate reward R_t on RTs; thus, our experiment tests the stronger version of the underlying hypothesis, that is, that the coupling between average reward and RT is effectively mandatory, even when it is not instrumental. A less noisy measure of vigor than RT would in any case be needed to enable a more fully quantitative test of the model. We did not find any influence on response vigor of the reward available in a given trial. This came as a surprise, given previous observations that the motivational value of an outcome influences the latency of the associated response (see e.g., Cools et al., 2005; Wittmann et al., 2005). One further possibility is that RTs in our task could have been driven by a temporally local prediction error between the offered reward and the average rate of reward ($R_t - \overline{r}_t$; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008). The positive excursions of this reward prediction error would be coded by the phasic activity of dopamine neurons (Schultz et al., 1997), which have been shown to influence vigor in monkeys (Satoh et al., 2003). However, the dependence of RT on the average reward rate $\overline{r}_t$ predicted from this relationship would be the inverse of what we observed (a higher reward rate would lead to a lower reward contrast and, thus, slower RTs).

At a single subject level, we did find a modestly significant relationship between the vigor on a trial and the reward obtained on the previous trial. Of course, this previous reward is the most significant single contributor to the running estimate of the average reward. This link was not significant in the overall average data. The discrepancy may, for instance, reflect a positively skewed distribution of learning rates for the average reward across the subjects, leaving a net explanatory gap for this effect of the previous trial.

Niv and colleagues suggested that the effects of average reward on vigor would be mediated by tonic dopamine levels. Obviously, the present experiment did not test this possibility, and therefore, future research should directly address the relationship between the computation of the average rate of reward, vigor, and dopaminergic neurotransmission in humans and experimental animals. However, in relation to dopamine, it is possible that the correlation between phasic dopamine release and vigor observed in monkeys (Satoh et al., 2003) arose as a result of influences on dopamine activity that are associated with control over tonic rather than phasic firing. Certainly, the mechanisms controlling tonic levels of extrasynaptic dopamine and its relationship with phasic dopamine appear complex (Floresco, West, Ash, Moore, & Grace, 2003; Grace, 1991), with the two signals possibly being independently regulated (Grace, Floresco, Goto, & Lodge, 2007; Lodge & Grace, 2006). Niv et al. (2007) explicitly discuss the association between such influences and their normative account; it is reminiscent of other apparent influences as in appetitive Pavlovian to instrumental transfer.

According to some influential models of decision-making such as the drift–diffusion or the LATER models, observed behavioral responses result from the accumulation of evidence for execution until a certain threshold is reached (Ratcliff & Rouder, 2000; Reddi & Carpenter, 2000). In this framework, changes in RT may arise because of changes either in the rate of evidence accumulation or in the threshold (Ratcliff & Rouder, 2000; Reddi & Carpenter, 2000). Previous research has shown that the parameters of the drift–diffusion model may be modified depending on whether participants are instructed to perform as quickly or as accurately as they can, suggesting that the RT may be affected by a speed–accuracy trade-off (Ratcliff & McKoon, 2008). It could, thus, be argued that the lack of effect of available reward on RT could be a result of strategic slowing down to increase accuracy when the available reward was high. However, such an interpretation is unlikely because our participants showed a high accuracy in their performance across the whole task and we did not observe any between-subjects correlation between the available reward and the percentage of correct responses. Whether slow adjustments in the threshold of a decision process as specified in these models may have contributed to the observed effects of average reward on RT remains unclear. Such a possibility is orthogonal to our hypothesis, and our design does not permit a test of this possibility.

If the average rate of rewards enhances the vigor of responding, it is natural to consider whether the average rate of punishments enhances sloth and how this might be realized in the brain. In fact, one major pillar of the current version of the computational proposal that serotonin acts as an opponent to dopamine (Boureau & Dayan, 2011; Deakin & Graeff, 1991) is that serotonin is directly implicated in behavioral inhibition and behavioral quiescence (Dayan & Huys, 2009) as a contrast to dopamine's involvement in behavioral activation (Cools et al., 2011). However, there is a key asymmetry to consider: The average reward is an opportunity cost for time, because not acting swiftly postpones rewards that can be earned. By contrast, at least some forms of punishment cannot be postponed by sloth, and the formalism would need to be extended to capture this asymmetry fully.

In conclusion, we found that human vigor in an RT task was directly influenced by the local average reward rate, with higher rates leading to faster reactions. This is in direct contradiction of the intuitive notion that subjects speed up on the basis of immediate fluctuations in potential rewards but supports a recent normative theory of vigor.

Reprint requests should be sent to Marc Guitart-Masip, ICN, University College London, 17 Queen Square, London, WC1N 3AR, United Kingdom, or via e-mail: m.guitart@ucl.ac.uk.

REFERENCES

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006). Reward-motivated learning: Mesolimbic activation precedes memory formation. *Neuron, 50,* 507–517.

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

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research, Brain Research Reviews, 28,* 309–369.

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

Cools, R., Blackwell, A., Clark, L., Menzies, L., Cox, S., & Robbins, T. W. (2005). Tryptophan depletion disrupts the motivational guidance of goal-directed behavior as a function of trait impulsivity. *Neuropsychopharmacology, 30,* 1362–1373.

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

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology, 16,* 199–204.

Dayan, P., & Huys, Q. J. (2009). Serotonin in affective control. *Annual Review of Neuroscience, 32,* 95–126.

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

Floresco, S. B., West, A. R., Ash, B., Moore, H., & Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature Neuroscience, 6,* 968–973.

Grace, A. A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience, 41,* 1–24.

Grace, A. A., Floresco, S. B., Goto, Y., & Lodge, D. J. (2007). Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends in Neurosciences, 30,* 220–227.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbito-frontal

cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience, 28,* 5623–5630.
Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience, 21,* RC159.
Langston, J. W., Forno, L. S., Rebert, C. S., & Irwin, I. (1984). Selective nigral toxicity after systemic administration of 1-methyl-4-phenyl-1,2,5,6-tetrahydropyrine (MPTP) in the squirrel monkey. *Brain Research, 292,* 390–394.
Lex, A., & Hauber, W. (2008). Dopamine D1 and D2 receptors in the nucleus accumbens core and shell mediate Pavlovian-instrumental transfer. *Learning and Memory, 15,* 483–491.
Lodge, D. J., & Grace, A. A. (2006). The hippocampus modulates dopamine neuron responsivity by regulating the intensity of phasic neuron activation. *Neuropsychopharmacology, 31,* 1356–1361.
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.
Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron, 36,* 265–284.
Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16,* 1936–1947.
Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9,* 1057–1063.
Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology (Berlin), 191,* 507–520.
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452–454.
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.
Parkinson, J. A., Dalley, J. W., Cardinal, R. N., Bamford, A., Fehnert, B., Lachenal, G., et al. (2002). Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: Implications for mesoaccumbens dopamine function. *Behavioural Brain Research, 137,* 149–163.
Phillips, P. E., Walton, M. E., & Jhou, T. C. (2007). Calculating utility: Preclinical evidence for cost-benefit analysis by mesolimbic dopamine. *Psychopharmacology (Berlin), 191,* 483–495.
Ratcliff, R., & Mckoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922.
Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 127–140.
Reddi, B. A., & Carpenter, R. H. (2000). The influence of urgency on decision time. *Nature Neuroscience, 3,* 827–830.
Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience, 10,* 1615–1624.
Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience, 30,* 13525–13536.
Salamone, J. D., & Correa, M. (2002). Motivational views of reinforcement: Implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behavioural Brain Research, 137,* 3–25.
Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *Journal of Neuroscience, 23,* 9913–9923.
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2002). Reward-dependent spatial selectivity of anticipatory activity in monkey caudate neurons. *Journal of Neurophysiology, 87,* 508–515.
Taylor, J. R., & Robbins, T. W. (1986). 6-Hydroxydopamine lesions of the nucleus accumbens, but not of the caudate nucleus, attenuate enhanced responding with reward-related stimuli produced by intra-accumbens d-amphetamine. *Psychopharmacology (Berlin), 90,* 390–397.
Ungerstedt, U. (1971). Adipsia and aphagia after 6-hydroxydopamine induced degeneration of the nigro-striatal dopamine system. *Acta Physiologica Scandinavica Supplement, 367,* 95–122.
Watanabe, M., Cromwell, H. C., Tremblay, L., Hollerman, J. R., Hikosaka, K., & Schultz, W. (2001). Behavioral reactions reflecting differential reward expectations in monkeys. *Experimental Brain Research, 140,* 511–518.
Wittmann, B. C., Schott, B. H., Guderian, S., Frey, J. U., Heinze, H. J., & Duzel, E. (2005). Reward-related fMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron, 45,* 459–467.