Neuron, Vol. 36, 265–284, October 10, 2002, Copyright ©2002 by Cell Press

Neural Economics and the Biological Substrates of Valuation

Review

P. Read Montague[1,3] and Gregory S. Berns[2,3]
[1]Center for Theoretical Neuroscience
Human NeuroImaging Lab
Division of Neuroscience
Baylor College of Medicine
1 Baylor Plaza
Houston, Texas 77030
[2]Department of Psychiatry and Behavioral Sciences
Emory University School of Medicine
1639 Pierce Drive, Suite 4000
Atlanta, Georgia 30322

A recent flurry of neuroimaging and decision-making experiments in humans, when combined with single-unit data from orbitofrontal cortex, suggests major additions to current models of reward processing. We review these data and models and use them to develop a specific computational relationship between the value of a predictor and the future rewards or punishments that it promises. The resulting computational model, the predictor-valuation model (PVM), is shown to anticipate a class of single-unit neural responses in orbitofrontal and striatal neurons. The model also suggests how neural responses in the orbitofrontal-striatal circuit may support the conversion of disparate types of future rewards into a kind of internal currency, that is, a common scale used to compare the valuation of future behavioral acts or stimuli.

Introduction

A general function of neural tissue is ongoing economic evaluation, a central function for any system that must operate with limited resources, that is, all mobile creatures. All mobile creatures run on batteries; they must continually acquire nutrients and expel wastes in order to reproduce and survive. Consequently, the way that mobile creatures value their internal states, sensory experience, and behavioral output influences directly how they will invest their time and energy. Our perspective here is focused. By economic evaluation, we refer to the problems that an individual nervous system faces when making rapid, moment-to-moment decisions possessing real costs and potential future payoffs (good and bad). A central feature of this problem is the need for an internal currency that can be used as a common scale to value diverse behavioral acts and sensory stimuli.

The need for common valuation scales arises from the sheer breadth and variety of information available to a creature's nervous system. Do I chase this new prey or do I continue nibbling on my last kill? Do I continue to drink from this pond or do I switch to foraging nearby for food? Do I run from the possible predator that I see in the bushes or the one that I hear? Do I chase that potential mate or do I wait around for something better?

[3]Correspondence: read@bcm.tmc.edu (P.R.M.), gberns@emory.edu (G.S.B.)

These questions illustrate issues, behaviors, and stimuli that are fundamentally unmixable; there is no natural way to combine or compare them. To do so, a creature must convert them into some kind of common scale (currency) and use such economic evaluations to choose a proper course of action. The need for valuation schemes guides the structure of our review here.

We first review evidence that midbrain dopamine systems encode errors in reward predictions, an error model that provides a perspective for designing and interpreting reward expectancy experiments in humans. As we detail below, these experiments show that both behavioral tasks and functional magnetic resonance imaging (fMRI) generate a more general view of reward processing than simple prediction errors. In particular, dopaminoceptive regions like the orbitofrontal cortex and striatum (OFS circuits) appear to be involved in valuation schemes meeting some of the needs described above. We discuss one natural model that emerges from the data, the prediction-valuation model, and show that its basic predictions match neural data from orbitofrontal neurons. Remarkably, this model has close ties to the option pricing theory from finance literature and may well represent a biological substrate underlying such approaches to human economic behavior.

Computational Interpretation of Phasic Dopaminergic Activity

During the 1980's and early 1990's, chronic single-unit recordings from midbrain dopaminergic neurons in alert monkeys showed that they give phasic responses to food and fluid rewards, novel stimuli, and stimuli eliciting orienting reactions (e.g., Romo and Schultz, 1990). This experimental work has flourished in recent years and has provided strong connections between dopaminergic activity and behavioral output (for review see Schultz, 2002 [this issue of *Neuron*]). In particular, this work has begun to lean more heavily on the computational learning theory (Montague and Sejnowski, 1994; Montague et al., 1994, 1996; Schultz et al., 1997; Egelman et al., 1998; Dayan et al., 2000) and psychological theories of conditioning (Waelti et al., 2001; Schultz and Dickinson, 2000). Overall, this work suggests that activity changes in a subset of dopamine neurons in the ventral tegmental area and substantia nigra represent prediction errors in the time and amount of future rewarding events; that is, changes in spike rate encode an ongoing difference between experienced reward and long-term predicted reward. Increases in spike rate mean "better than predicted," decreases in spike rate mean "worse than predicted," and no change in spike rate means "just as expected."

Figure 1 illustrates the way in which activity changes in dopaminergic neurons are thought to convey information about prediction errors in future reward. Figure 1A shows a summary of how spike production in dopaminergic neurons changes with learning, and Figure 1B shows a qualitative summary of the results in (A). The basic computational model that captures these data

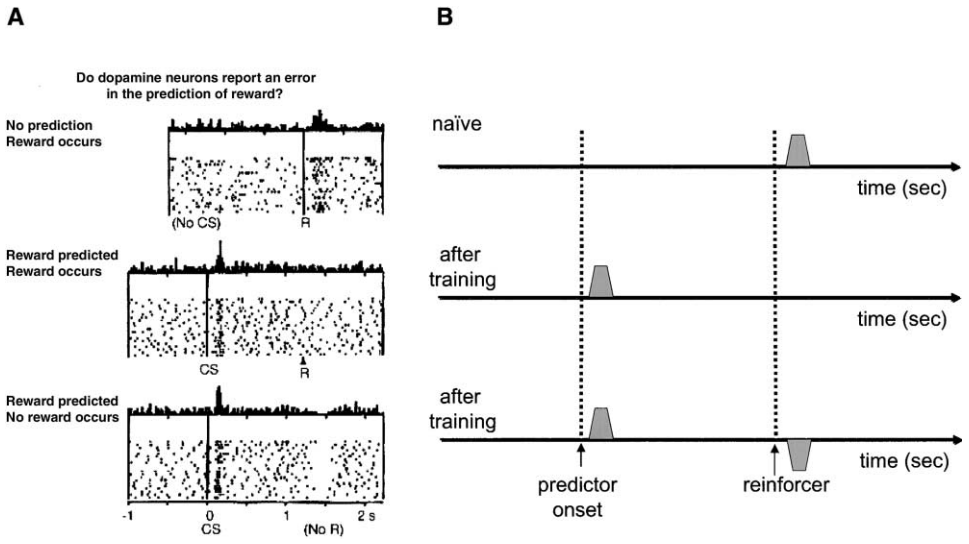

Figure 1. Dopamine Neurons Encode a Temporal Prediction-Error Signal in Their Phasic Spike Activity

(A) Recordings from single units in primate ventral tegmental area during simple learning tasks have shown that dopamine neurons emit a signal consistent with a prediction error in future reward. Raster plots of spike times overlayed with histogram of spike activity. The qualitative changes are most important here, but for scale, the transient in the middle trace is approximately 15 spikes/s. In a naïve animal, the unexpected delivery of a positive reinforcer (R) causes an increase in spike rate (positive prediction error), but the earlier presence of a neutral cue or no cue (labeled no CS) causes no significant change in activity (top traces). If the delivery of reward is consistently preceded by a sensory cue, two changes occur in the phasic firing of the dopamine neurons: (1) the response to the reward (R) sinks back into baseline, and (2) a positive response grows locked to the earliest consistent predictor (middle trace). The omission of an expected reward results in a decrease in spike output (negative prediction error) at the time of the expected reward. All traces are aligned to the solenoid activation that delivers the reward, R (juice squirt) (bottom trace).

(B) Qualitative summary of the measured effects illustrated in (A).

derives from a model for prediction learning called the method of temporal differences or TD learning for short (Sutton, 1988; Sutton and Barto, 1998). The ultimate goal in TD learning is for a system to learn to predict the time and amount of future rewards or punishments. A key signal in this computational learning procedure is the TD error, which represents a prediction error in the long-term estimate of the average reward that is expected from time t into the distant future. It is precisely this TD error signal that has been proposed as the prediction error signal emitted by midbrain dopaminergic neurons (Figure 2; Montague and Sejnowski, 1994; Montague et al., 1996; Schultz et al., 1997). Details of the model can be found in Appendix A, and an example of how it captures the basic electrophysiological results is shown in Figure 2.

Figure 3 compares the model's performance against single-unit recordings for a task in which multiple sensory cues predict a terminal reward, but the temporal consistency of the second cue is varied. For the temporally consistent case, cue 1 is always followed by cue 2 with a fixed 1 s delay. Cue 2 acts as the behavioral trigger, which releases the animal to execute a simple action to obtain reward (here, a juice squirt). Early on during training, the dopamine cells gave a phasic burst to reward delivery and no response to either of the two predictive cues. The results shown here are after training has occurred, and clearly the phasic response is now linked to the earliest consistent reward-predicting cue (instruction light). The situation changes completely if the second cue is made to vary in its time, but not its probability of arrival (100% schedule). After training, both the instruction light and trigger signal elicit phasic activation of these neurons. Moreover, the phasic activation following the trigger cue is more spread out (uncertain) through time, a feature that parallels that uncertainty in the time of arrival of the trigger cue. The prediction-error model responses are shown beneath each of these data plots and show that the model can account for these detailed spike data.

The data discussed above (Figures 1–3) and the model that captures them suggest strongly that the signal emitted by dopamine neurons goes well beyond the simple idea that dopamine delivery simply equals reward. Recent data on dopamine release in freely moving rats now makes such an equivalence untenable (Garris et al., 1999; Kilpatrick et al., 2000). Figure 4 shows an ongoing dopamine measurement made in the nucleus accumbens of alert rats while they press a bar to activate their own ventral tegmental area through an implanted electrode. Early on during the self-stimulation, each bar press elicits large, rapid transients in dopamine release. Shortly thereafter, these same transients in dopamine delivery cannot be detected despite continued bar pressing by the animal (vertical black bars) and, hence, continued electrical stimulation through the implanted electrode (Figure 4). The animal is not simply stuck in a repetitive behavior because the electrical stimulation is required to maintain the bar pressing behavior whether or not dopamine is released at the target structure (here, the nucleus accumbens). A similar self-stimulation experiment making dopamine measurements in the dorsal striatum revealed analogous results (Kilpatrick et al., 2000; data not shown). These results, when considered

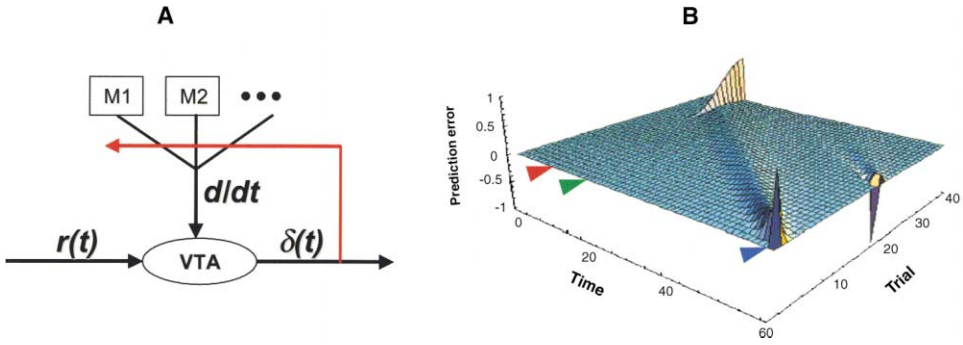

Figure 2. Temporal Difference Model Captures Basic Electrophysiological Observations in Midbrain Dopamine Neurons

Temporal difference model captures basic electrophysiological observations in midbrain dopamine neurons. The basic response features of Figure 1 are reproduced in a temporal difference model.

(A) Hypothesized architecture. Sensory stimuli, processed by modules M1 and M2, provides inputs to ventral tegmental area (VTA) neurons in the form of a temporal derivative (*d*/*dt*). The VTA neuron also receives direct input *r(t)*, representing the time-varying reward signal. By adding these signals, the VTA neuron emits a signal, $\delta(t)$, equivalent to the prediction error used in temporal difference learning (that is, $\delta(t) = r(t) + \gamma V(t+1) - V(t)$, where $\gamma V(t+1) - V(t)$ is one discrete approximation to d/dt). This signed error signal can be used in a correlational synaptic learning rule to permit the system to learn predictions of future reward, which are then stored as synaptic strengths (***W***). Such a rule has been called a predictive Hebbian rule, and weights are updated as: $\boldsymbol{W}_{new} = \boldsymbol{W}_{old} + \lambda \boldsymbol{x}(t)\delta(t)$, where ***x(t)*** is the presynaptic activity and λ is a learning rate (see Montague and Sejnowski, 1994; panels [A] and [B] adapted from Montague et al., 1996 [copyright 1996 by the Society for Neuroscience] and Schultz et al., 1997 [reprinted with permission from Schultz et al., 1997. Copyright 1997, American Association for the Advancement of Science]).

(B) The basic response features of Figure 1 are reproduced in a temporal difference model. In this example, two sensory cues are presented (red light at time 10, green light at time 20) on each trial followed by a reward at time 60. Early in training, the delivery of unexpected reward is accompanied by a phasic increase in the firing rate. As training proceeds, the reward response sinks back to baseline (in this noiseless example, baseline is 0), and a response grows to the earliest consistent sensory cue. This model does not distinguish between sensory-sensory prediction and sensory-reward prediction. The system learns both the time of the second light and the time and magnitude of the reward. Nondelivery of the reward results in a downward deflection in activity to reproduce the data shown in Figure 1. This simple model captures a wide array of experimental data; however, dopamine neuron responses are known to be richer than the collection shown in Figure 1.

in combination with single-unit activity shown in Figures 1 and 3, provide convincing evidence that phasic changes in dopamine delivery are not the singular physical substrate of reinforcement. However, they leave open the possibility that some function of the average dopamine level may contribute directly to the reinforcing qualities of a stimulus. These experiments are consistent with the prediction error hypothesis for dopaminergic spike activity, even though extra filtering is apparently taking place at the level of dopamine release.

Overall, these data suggest a basic computational role for dopamine neurons as detectors of ongoing changes in predictability, an important first-level analysis that must be done on information sequentially streaming into any neural system. Changes in predictability act as markers for epochs during which important information is being detected or processed. Such event markers provide a natural signal to start or stop learning and to direct attention, two informational roles in which dopamine is known to be a major player. This perspective on dopaminergic processing will certainly continue to evolve as new data are generated, but more importantly, they provide a conceptual and computational structure for understanding recent work on reward expectancy in human brains.

Human Brain Imaging of Reward Expectancy

A recent flurry of neuroimaging results in humans has provided a step forward in understanding reward processing in humans. However, there are constraints that limit the study of reward processing in humans with noninvasive brain imaging. For example, as a neural probe, fMRI cannot measure directly the efflux of dopamine or other neuromodulators. Despite this measurement limitation, one can still use fMRI to make meaningful differentiations between reward and the reward expectancy in the human brain. Another major limitation for studying reward processing in humans is the profligate use to which the term reward is subjected. Reward is used in many ways in many contexts, and it is often used interchangeably with the term reinforcement. Hence, one must be careful to delimit its meaning in any particular experiment. Most investigators have taken the definition of reward to be a stimulus that can act as a positive reinforcer, although in any given experiment, a reward may or may not be used to reinforce anything. In humans, reward generally takes the form of appetitive stimuli (food, water, drugs) or money. There has been a proliferation of human experiments probing reward expectancy in humans using fMRI, and these may well serve to define new and more highly differentiated notions of reward, that is, more detailed, algorithmic descriptions.

Brain Response to Primary Reward and Its Predictability

A straightforward approach to the study of reward processing in humans is to probe the brain's response to primary reward, i.e., appetitive stimuli. The acute effects of cocaine, for example, are associated with increased striatal activity, which also correlates with subjective responses (Breiter et al., 1997). In our work, we have focused on the role of predictability in the brain's response to gustatory stimuli. The approach is simple: take a particular stimulus, render it either predictable or

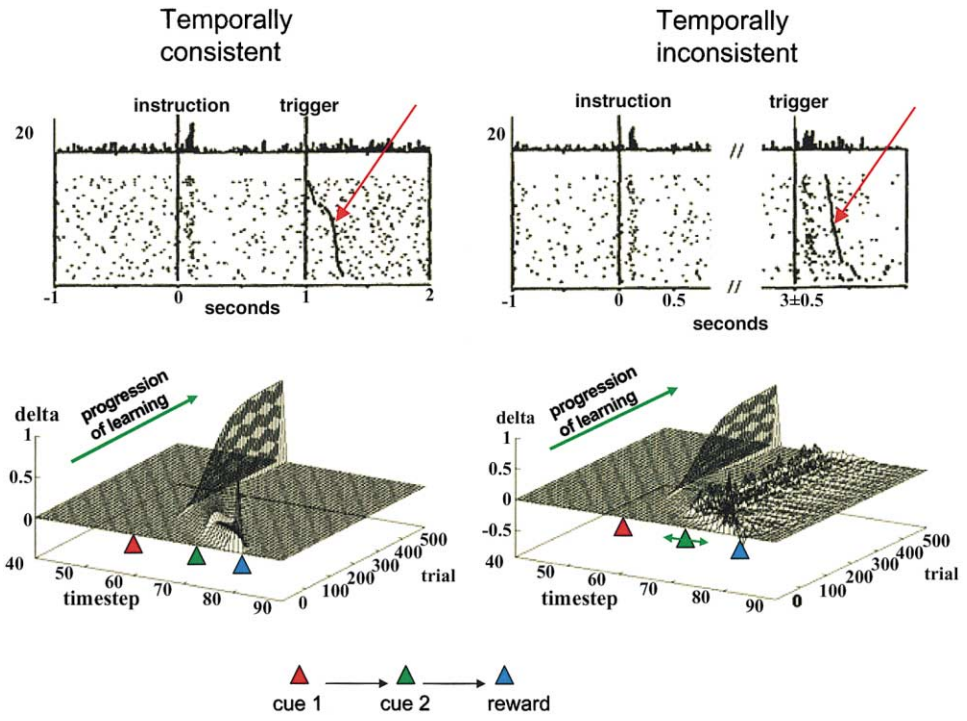

Figure 3. Response of Prediction-Error Model to Temporally Consistent and Temporally Inconsistent Sensory Cues

Development of prediction error signal through training, using two successive sensory cues followed by reward (juice). The variable of interest in these experiments is the consistency of the temporal relationship between the two predictive sensory stimuli. A sequence, cue 1 (instruction) → cue 2 (trigger) → reward, was given repeatedly to an alert primate while recording from single midbrain dopamine neurons. Each cue always occurred prior to reward delivery and in the order shown; therefore, both cues were predictors of reward. These raster plots and firing rate histograms are taken from experiments with a monkey already trained on the tasks.

(Top) In the temporally consistent case (left, top), cue 1 (instruction) and cue 2 (trigger) were separated by a fixed 1 s delay. In the temporally inconsistent cases (right, top), cue 1 (instruction) and cue 2 (trigger) were separated by a variable and unpredictable 1.5–3.0 s delay (random times in this range were used). In the temporally consistent case, the phasic change in dopamine neuron firing rate is associated only with the time of cue 1 onset. However, for the inconsistent case, the phasic change in dopamine neuron firing rate is associated with the onset of both cue 1 and cue 2. In addition, the extra spikes produced subsequent to the trigger stimulus are more spread out through time, as though the circuit is more uncertain about the expected time of onset of the second cue. The time of movement onset in each trial is indicated by a larger black marker, and the trials have been shuffled and ordered according to this time of movement onset (red arrows). Scale bar at top is 20 spikes/s.

(Bottom) Prediction-error model response. These effects of temporal consistency are captured in the prediction-error model response shown below each respective experiment. Time progresses from left to right, and individual learning trials progress in the direction of the green arrows marked progression of learning. As before, the phasic change in dopamergic activity is shown as a signed quantity labeled delta and represents the change in activity from baseline. In the temporally inconsistent case, the model is unable to discount the arrival of cue 2 because its time of arrival after cue 1 is not sufficiently consistent (predictable), a finding that reproduces the experimental result (adapted from Montague et al., 1996 [copyright 1996 by the Society for Neuroscience]).

unpredictable, and measure the associated brain response in systems thought to process the stimulus. This has been applied in the context of visual and motor tasks (Berns et al., 1997; Bischoff-Grethe et al., 2000), but here, we focus on the brain response to changes in predictability for rewarding gustatory stimuli. The human brain response to gustatory stimuli is robust; both gustatory stimuli and their predictors can generate responses detectable using fMRI. Several brain imaging studies have demonstrated activation of orbitofrontal cortex by a range of gustatory stimuli (Rolls, 2000; O'Doherty et al., 2001, 2002; Zald et al., 2002).

As discussed above (Figures 1–3), the predictability of a primary rewarding stimulus is a critical parameter for activation of reward pathways in both rats and monkeys. Consequently, the predictability of a sequence of stimuli may itself recruit reward-related neural structures in a manner detectable with fMRI in humans. We chose gustatory stimuli (juice and water) because: (1) they represent a very basic form of appetitive reward, (2) they are used routinely in primate experiments on reward expectancy (Schultz, 1998), and (3) they can both be used to reinforce behavior.

In our first experiment, we delivered sequences of small (0.8 ml) juice and water squirts to humans, manipulated the predictability of the sequences, and measured the brain response using fMRI (Berns et al., 2001; Figure 5). The receipt of juice or water alone evoked widespread activation throughout the brain, including orbitofrontal cortex and motor areas (data not shown). The brain response to juice alone did not differ from the brain response to water alone; that is, irrespective of predictability, subtracting the juice alone fMRI response from the water alone fMRI response yielded no significant activation in any voxels. However, as indicated in Figure 5, when difference images were computed to uncover

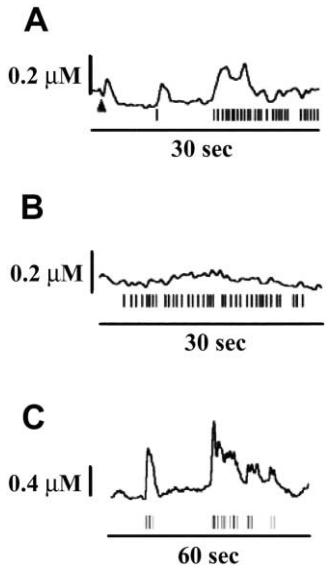

Figure 4. Dopamine Release Does Not Equal Pleasure and Is Not Linearly Related to Changes in Spiking

Ascending connections from the midbrain distribute a prediction error signal $\delta(t)$ in the form of increases (better than expected), decreases (worse than expected), or no change (just as expected) in spike activity. It has long been thought that dopamine release was equivalent to the subjective pleasure or euphoria that attends stimuli that increase dopamine release in target neural structures. Recent experiments by Wightman and colleagues (Garris et al., 1999) show that there is not a simple equivalence between dopamine release and pleasure. They used cyclic voltammetry to measure dopamine release in the nucleus accumbens and striatum of alert, freely moving rats, while the rats pressed a bar to stimulate their own ventral tegmental area through an implanted electrode.

(A) Vertical black bars indicate times at which the animal pressed the bar to receive stimulation. Each stimulus (single vertical bar) was a repetitive train of 24 current pulses delivered at 60 Hz. The filled arrowhead represents an experimenter-delivered stimulation. Early on during self-stimulation, large transients in dopamine were measured; however, as the animal continued to bar press (self-stimulate), the measured dopamine release dropped to 0 despite continued stimulation.

(B) Same animal 30 min after the self-stimulation experiment in (A). As before, the rat will continue self-stimulate to deliver an electrical stimulus to the implanted electrode; however, there was no change in the measured dopamine release.

(C) Dopamine transients can still be elicited by unexpected experimenter-delivered stimuli delivered to the same implanted electrode used for the self-stimulation experiments in (A) and (B). These experiments show clearly that dopamine release does not equate with pleasure and that the spikes that travel from dopamine neurons, regardless of what they represent computationally, may or may not elicit dopamine release onto target structures (adapted from Garris et al., 1999; copyright 1999 by Nature, www.nature.com).

the brain response to changes in the predictability of the sequences, highly significant activation occurred throughout the ventral striatum including strong activation in the nucleus accumbens. In this first experiment, we sought to change only predictability; however, there are two sources of prediction–what stimulus arrives next (stimulus prediction) and when the next stimulus occurs (temporal prediction). In the predictable run, both the time and identity of the next stimulus is predictable. In the unpredictable run, neither the time nor the identity of the next stimulus is predictable. In our subsequent experiments, we sought to separate the influence of these two sources of predictability changes.

To separate the influence of stimulus predictability and temporal predictability, we broke the sequential stimulus (Figure 6A) into its component parts (Figures 6B and 6C). In this event-related design, we sought to characterize the influence of changes in temporal predictability. A juice squirt was paired in time with a light that preceded it by a fixed time interval. During training, the time between the light-juice pairs was randomized, but the schedule of reinforcement was 100%. After training, catch trials with a new (unexpected) time for juice delivery were inserted at random in an otherwise normal training sequence. Contrast images were formed between the brain response to juice at the expected and unexpected times. As illustrated (Figures 6B and 6C), this paradigm has been carried out using both a passive and active (instrumental) design.

In the active task, subjects pressed a button after onset of a green light, and juice was delivered 4 s later. The average time from light onset to button press was 1.5 s. In the catch trials, juice delivery was delayed an extra 4 s, allowing for the measurement of the brain response to a time-locked reward prediction error (Figure 6C). Early reward delivery was not investigated due to the difficulty in separating hemodynamic responses less than 4 s apart. As shown in Figure 7, the time course of the hemodynamic response in the ventral striatum indicated a robust activation at precisely the time when juice was expected, but not received (Pagnoni et al., 2002). Here, the change in activity in the ventral striatum appears to be locked to the error in reward prediction.

The passive task, as illustrated in Figure 6B, presented an analogous design to subjects. A yellow light was followed 6 s later by the same juice squirt used in the active task. This time is comparable to the active task since the average time to button press was $\sim$1.5 s, making a total of 5.5 s from light to reward delivery in training runs. In both tasks, the extra delay on catch trials was 4 s. Remarkably, the passive task uncovered strong, specific activity in the dorsal striatum locked to the reward prediction error (data not shown; Samuel McClure, personal communication). This work is still underway with many important questions still open. These activations may reflect responses participating in making predictions or may reflect true prediction error responses. In either case, one would expect a strong brain response to be temporally locked to the reward prediction error signal. Further experiments are required to separate the contribution of the predictions (inputs) and the prediction error (output).

Brain Response to Monetary Reward, Its Anticipation, and Its Context

Recent studies have demonstrated a correlation between the relative magnitude of monetary reinforcement and the fMRI signal in both ventral striatum (Delgado et al., 2000) and orbitofrontal cortex (Knutson et al., 2000; Breiter et al., 2001). The importance of this work derives in part from the abstract nature of a monetary reward. Money represents value to an individual but in a potentially idiosyncratic way. It is therefore significant that the brain response in the orbitofrontal-striatal (OFS) circuit correlates with relative monetary reward, and this response is consistent across subjects.

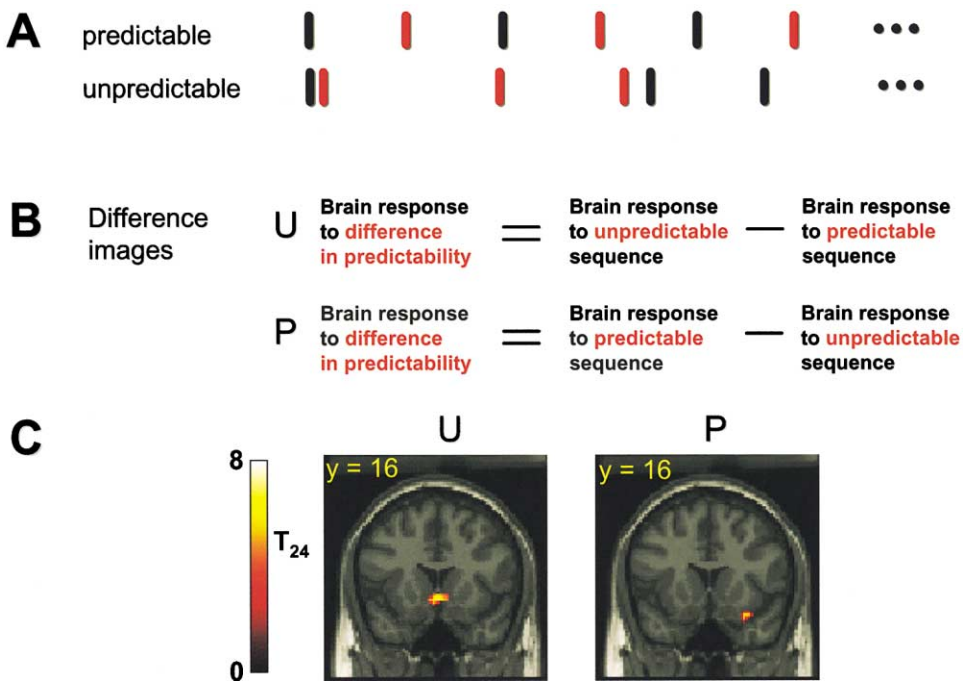

Figure 5. Change in Predictability of Sequential Gustatory Stimuli Activates Dopaminergic Targets

Changes in predictability of sequential stimuli mark epochs during which important information is being detected or processed. As shown above in Figure 1, midbrain dopamine systems give phasic responses to such changes in predictability. Dopamine has a powerful influence on neural activity and a direct effect on microvasculature (Krimer et al., 1998); therefore, we sought to test whether changes in predictability in sequential stimuli could be detected in human subjects using fMRI. In this experiment, two sources of predictability were concurrently changed in an effort to seek a maximal response. (1) The probability, P_d, that the next stimulus was different from the current stimulus changed from $P_d = 1$ in the predictable sequence to $P_d = 1/2$ in the unpredictable sequence. (2) The time boundaries of each fluid stimulus were randomized, but the average inter-stimulus time was held at a constant 10 s. This latter change was accomplished by randomly sampling a Poisson interval distribution with an average interval of 10 s. The fixed time between stimuli in the predictable trial was 10 s.

(A) Boluses of juice (red bars) and water (black bars) were delivered in either predictable or unpredictable sequences, and changes in the BOLD response were measured.

(B) Two difference images were computed for all brain voxels and are labeled here as U and P. These difference images will exclude common regions of activation associated with juice and water delivery and the swallowing movement that ensues. In this manner, these contrasts reveal the brain response only to changes in predictability for the sequential gustatory stimuli.

(C) Reward-related regions had a greater BOLD response to the unpredictable sequence than to the predictable sequence (U). Note the bilateral activation of the nucleus accumbens (left). Although not visible in this coronal view, the ventromedial prefrontal cortex, a dopaminergic target structure, also showed strong activation for the U contrast. Significance thresholded at $p < 0.001$ and a spatial extent of greater than ten contiguous voxels. Pseudo-color scale shows the results of a T score with 24 degrees of freedom ($n = 25$ subjects). In this initial experiment, both stimulus predictability (what comes next) and temporal predictability (when next stimulus arrives) were changed. In Figure 6, we break the sequence into its components and show how we test only the contribution of changes in temporal predictability (see Berns et al., 2001).

As with appetitive rewards, the brain response to monetary reward is not static. For example, the brain response for an impending monetary reward (or punishment) will transfer to a conditioned stimulus (light or abstract figure) through the appropriate pairing of the conditioned stimulus and monetary reward. The conditioned stimulus evokes a response in the striatum (Knutson et al., 2001) and orbitofrontal cortex (O'Doherty et al., 2001) that scales with the size of the reward prediction. Consistent with these findings, other work has shown that the striatal response appears to be maximal during anticipation of impending reward (Breiter et al., 2001). It is well known that the context in which a reward is offered ("framed") affects the behavioral valuation of the reward (Kahneman and Tverksy, 1979).

In two recent studies, the contextual effect was the dominant influence on the OFS brain response (Elliott et al., 2000; Knutson et al., 2001). The powerful influence of context has also been seen by Breiter and colleagues in the striatum, where the response to anticipated reward is largest when contrasted to the anticipation of impending loss (Breiter et al., 2001). These experiments suggest that the OFS represents more than a simple reward prediction error. We develop this possibility below but first present behavioral tasks designed to probe directly the prediction-error model as a biasing signal for action choice.

Biasing Action Choice Using Prediction Errors

As discussed above, the prediction-error model is a fruitful starting point for understanding the possible meaning of spike activity in midbrain dopamine neurons and interpreting fMRI experiments on reward expectancy in human subjects. There are numerous reasons to suspect that the same prediction error signal, encoded as phasic changes in dopamine delivery, is used directly to bias action selection. There are heavy dopaminergic projections to neural targets thought to partici-

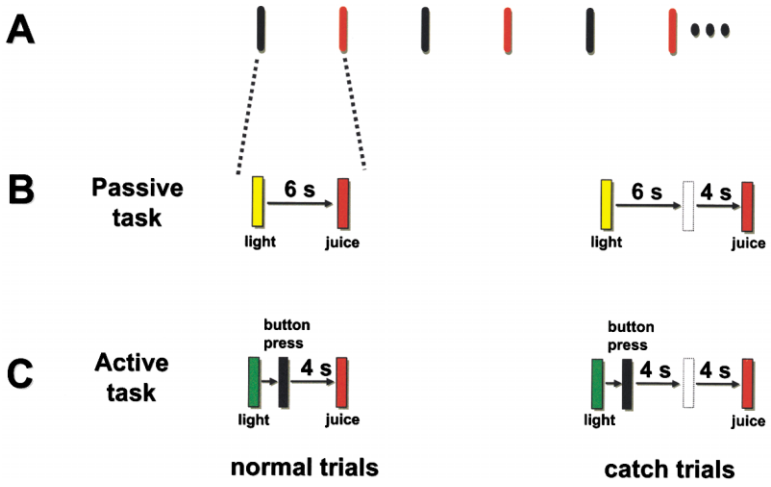

Figure 6. Role of Temporal Predictability in Human Brain Response to Changes in Expectations

In Figure 5, both the time and character of the next stimulus was changed between the predictable and unpredictable sequences in order to achieve maximal predictability and unpredictability for the sequential stimuli that we used.

(A) The basic strategy here is to break the sequential stimulus into its component parts and separate the influence of changes in temporal expectations and stimulus expectations (what comes next).

(B) Passive task. Light cue is presented for 1 s followed 6 s later by a juice squirt. These pairings are presented with randomized interpair times (roughly 50 pairings per subject); after this training, a similar run is carried out containing six catch trials, where the juice is delivered at an unexpected time (10 s). The contrast to examine is between juice delivered at 10 s (unexpected) and juice delivered at 6 s (expected) (data not shown).

(C) Active task. Light cue is again presented and subject presses a button in free response (average time to press = 1.5 s). During training, juice is delivered exactly 4 s after button press. After training, catch trials (where juice delivery is delayed by 4 s) are again interspersed with normal trials. Rather than show a single contrast image, we have plotted the average hemodynamic response during the active task (Figure 7).

pate in the sequencing and selection of motor action, that is, the dorsal striatum. Disease states that perturb dopamine levels (Parkinson's disease) interfere catastrophically with sequenced motor output. Single-unit electrophysiology experiments in alert primates carrying out behavioral tasks show that the learning displayed by the dopamine circuit always occurs before measurable changes in the animals' behavior (Hollerman and Schultz, 1998). Lastly, from a computational perspective, the prediction error signal is ideally suited to act as a critic of actions that lead to immediate changes in reward (see Dayan and Abbott, 2001; Dayan and Balleine, 2002 [this issue of *Neuron*]). Taken as a whole, these diverse lines of evidence suggest that transient changes in dopamine release, captured in part by the prediction-error model, may participate directly in the selection of actions that lead to reward. Below, we review the rationale for this hypothesis and its relationship to behavioral and physiological experiments carried out in honeybees and humans.

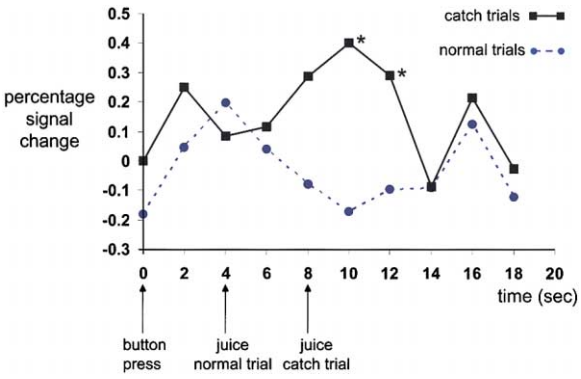

Figure 7. Average Hemodynamic Response during Active Temporal Predictability Task

Average BOLD response for normal trials and catch trials. Statistically significant difference between normal trial average and catch trial average was found at 10 s ($p < 0.0036$) and 12 s ($p < 0.0489$) after button press. Assuming a hemodynamic delay of 6 to 8 s, the curves diverge only at the time of the prediction error (adapted from Pagnoni et al., 2002 [copyright 2002 by Nature Neuroscience, www.nature.com/neurosci). A similar divergence occurs in the passive task (data not shown; Sam McClure, personal communication).

Uncertainty in Immediate Reward

The prediction error interpretation of dopaminergic activity works best in a perfect world where there is no uncertainty in the time or magnitude of future rewards. In classical statistical estimation, uncertainty is defined as the estimated amount by which an observed or calculated value, $\hat{A}$, may differ from the true value, A. This is typically computed as the square root of the mean squared deviation from the true value, $<(\hat{A}-A)^2>^{1/2}$, and is accordingly called the estimation error. This notion of uncertainty captures formally the idea that there is a possible spread in the values that would be obtained during an observation or within a finite sample. Hence, uncertainty quantifies a kind of degree of ignorance about a specific variable, and such ignorance always possesses a cost. A real-world sensory cue that predicts (estimates) the future time and magnitude of a reward will have uncertainty associated with that estimation, and that uncertainty has measurable costs to a mobile creature. These costs influence the behavioral investments that an animal is willing to make with respect to a reward predictor, and they influence the character of the underlying neural computations that support such behavioral choices. Above, we reviewed experiments that addressed temporal uncertainty in reward delivery in passive tasks, but here, we focus on uncertainty in reward magnitude linked directly to actions.

Two-Choice Sequential Decision Tasks for Bees and Humans

Uncertainty in reward magnitude has been tested behaviorally in honeybees (Real, 1991) and provides a starting point for connecting the prediction-error signal to action selection. Honeybees possess octopaminergic neurons in their subesophogeal ganglion that are responsible for reward-based learning in a fashion equiva-

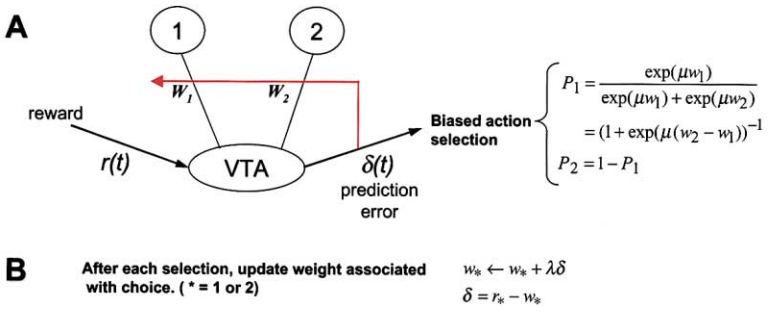

Figure 8. Using the Prediction as an Incentive to Act

(A) Specific action selection model that uses a reward prediction error to bias action choice. Here, two actions, 1 and 2, are available to the system, and in analogy with reward predictions learned for sensory cues (Figure 1 and 2), each action has an associated weight that is modified using the prediction error signal according to the time-independent version of a predictive Hebbian rule, the Rescorla-Wagner rule. As indicated, the weights are used as a drive or incentive to take one of the two available actions, which are chosen using a probabilistic policy.

(B) Once an action is selected and an immediate reward received, the associated weight is updated according to the Rescorla-Wagner rule. This setup can be used to model decisions made by both bees and humans on sequential decision-making tasks where the rewards are respectively, nectar and money.

lent to that seen in the primate dopamine work (Hammer, 1993); that is, they too appear to encode reward prediction errors in their spike output (Montague et al., 1994, 1995). This prediction-error view of the octopaminergic neurons has yielded insights into the computational mechanisms that bees use to make decisions about reward-yielding actions and has provided direction for related work in humans.

Real and colleagues allowed honeybees to forage over an artificial field of flowers where flower color (blue or yellow) was the only predictor of the nectar volumes (see Real, 1991). This arrangement represents a sequential decision task where a reward follows immediately after each selection: choose a flower color, land, acquire nectar volume, and decide whether to switch or sample the same color again. The question at hand was simple. How do honeybees value uncertainty in reward magnitude that follows a decision to sample one of the two flower types? In the actual experiment, both flower types yielded the same mean reward (2 μl of nectar). Initially, all of the blue flowers gave 2 μl of nectar, while 1/3 of yellow flowers gave 6 μl of nectar and the remaining 2/3 yielded 0 μl. Both colors predicted a mean return of 2 μl, but as a predictor of nectar volume, yellow had high variance (yielding many failures) and blue had zero variance. The bees were quite risk averse and chose blue flowers on greater than 80% of their flower visits. As a behavioral result, this finding agrees with the way that humans behave: they act to avoid uncertainty in reward magnitude.

Connecting the Reward Prediction-Error Signal to Action Selection (Bee)

It has been shown that reward predictor neurons in the bee (the octopaminergic neurons) can be used directly in an action-selection model to generate the same decision-making behavior observed in the real bees (Figure 8; Montague et al., 1995). Specifically, the reward prediction error is used to bias action selection through a probabilistic policy where the likelihood of selecting yellow (P_Y) is a function of the synaptic weight (W_Y) associated with yellow; likewise for blue. The weights are adjusted using the same reward prediction error signal as indicated in Figure 8. Through learning, the weights W_Y and W_B come to represent the current estimate of the average reward expected for choosing yellow (Y) or blue (B), respectively. This model shows clearly the dual use of the reward prediction error: (1) learning–it is used to update the current estimate (weight) of the value of a behavioral choice, and (2) action choice–it is used as the drive that biases action choice.

Weights Can Be Used in Dual Roles: Reward Predictions and Incentives to Act

For a sensory cue, the weight that develops in a predictive Hebbian rule (see legend, Figure 2) represents the future importance of that cue; hence, it encodes the degree to which an animal should want to process the cue. In situations where actions and their associated cues are followed immediately by reward, the weight that develops encodes the animal's willingness to take the action again in the future when presented with the choice to do so. This willingness takes the form of a probabilistic function, P, whose argument is the difference between the weights (Figure 8). In this sense, the weight, when used to bias actions, encodes the animal's incentive to take a particular action. These ideas show that for the limited cases considered here, our dual use of the prediction-error signal appears to take account, in a quantitative form, of psychological ideas like incentive salience (Berridge and Robinson, 1998).

As shown in Figure 9, the model matches the observed bee behavior quite well and suggests a substrate for how honeybees may compute and value the average uncertainty (variance in returns) associated with a reward predictor (flower color). Accordingly, the model suggests a physiological substrate and computational mechanism underlying risk aversion in bees.

Connecting the Reward Prediction-Error Signal to Action Selection (Human)

Inspired by this work in bees and the analogy with the dopamine-based reward prediction error, Egelman et al. (1998) used a modification of both the experimental task and the bee model to query risk aversion in human subjects. They used the same behavioral arrangement, that is, a two-choice behavioral task with immediate rewards following each choice (Figure 10). Instead of nectar volumes, monetary returns were used. In contrast with the bee experiment, instead of associating each choice with a fixed average uncertainty (variance) in reward, changes in reward for each choice were made into continuous functions of the history of choices made. The predicted behavior of the bee model guided the design of the reward functions (Figure 11). Specifically,

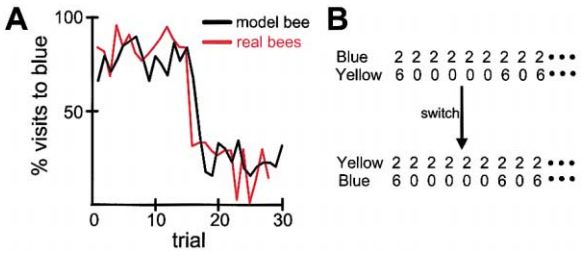

Figure 9. Prediction-Error Model of Sequential Decision Making in Bees

Honeybees were allowed to forage over an artificial field of blue and yellow flowers where the only predictor of reward was flower color (yellow or blue). Both flower types yielded the same mean reward (2 μl of nectar). Initially, all the blue flowers gave 2 μl of nectar, while 1/3 of yellow flowers gave 6 μl and the remaining 2/3 yielded 0 μl. That is, both colors predicted a mean return of 2 μl, but as a predictor of nectar volume, yellow had high variance (yielding many failures) and blue had zero variance. Bees possess octopaminergic neurons whose spike production is consistent with a reward prediction error signal and whose output is necessary for reward-dependent learning in the bee.

(A) Using this prediction-error signal in a computational model of action choice (model described in Figure 8), a model bee (black line) is shown to avoid the risky yellow flowers at the same rate as normal bees when the learning rate is high (see Montague et al., 1994, 1995; Figure 8). For both real bees and model bees, the fraction of visits to blue is an average over 40 flower visits. At trial 15, the statistics of nectar return for each flower color were switched so that yellow predicted zero variance in nectar returns (no uncertainty in magnitude) and blue became the variable predictor. Real bees switch in about three to five flower visits and at the high learning rate chosen ($\lambda = 0.95$ for the learning rule in Figure 8B); the model bees also switch their preference in approximately the same number of visits. These results show that real bees are averse to uncertainty in future reward magnitudes and can quickly learn to avoid risky flowers, that is, flowers that yield highly variable returns. The model shows how the prediction error signal available to the bee brain and generated by its octopaminergic neurons can be used to guide such behavior.

(B) Representative reward sequences for a series of samples from each flower type.

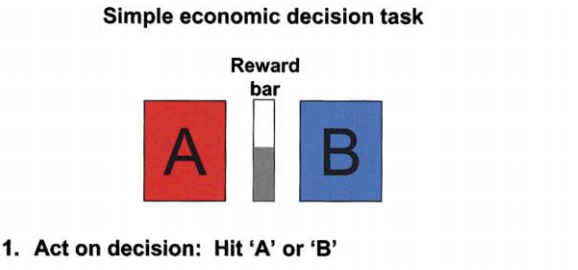

Figure 10. Sequential Decision-Making Task for Humans

A two-choice, decision-making task given to human subjects and a prediction error-based action model (Figure 8). Two buttons appear on a computer screen with a centrally placed slider bar, which indicates the magnitude of the immediate return after each selection. The slider stays at its last height until the next selection is made; that is, there is no memory requirement to execute the task. As indicated, the subject makes a choice, the slider bar changes its height from its last position, and the subject must choose to stay with current choice or switch to the other alternative. The subjects are told that: "each choice, A or B, results in a reward, which will be shown as a change in the slider bar height. Higher is more, lower is less. Try to make as much as you can by making choices that keep the slider bar high. There are no time limits; the computer will stop the task when it is complete."

The task was designed based on the success with the bee model but with one major difference. In the bee experiment and model, the tested parameter was the average uncertainty (variance) in reward magnitude associated with each of two choices (blue flower or yellow flower). Just as in the bee experiments, the monetary reward (change in slider bar height) was delivered immediately after each action. The difference in the human experiment is that the uncertainty in the reward magnitude was made to be a continuous function of the fractional allocation to button A (see Figure 11). The fractional allocation of choices to button A was computed as an average over the last 20 selections. It should be noted that the results reported below do not change if this window is extended to the last 40 selections.

the reward functions were built to trap the dynamics of the model in specific decision strategies.

There are three valuable outcomes of this approach. (1) It connects the prediction-error model of single-unit activity to biases in action choice without committing to any behavioral model of sequential decision-making strategies. (2) The action choice model makes quantitative predictions of how humans should perform under complex reward structures, if their actions are biased primarily by a dopamine prediction error signal. (3) The model provides insight into one possible neural substrate of matching behavior (Herrnstein, 1990; see below).

One particular design feature of the reward functions deserves comment. If learning rates are set appropriately, the action-selection model (Figure 8) will adjust its weights in a manner that forces it to make allocations to each choice (button A or button B) that place it near crossing points in the reward functions. This attraction for the crossing point can be thought of as a kind of value illusion (see Appendix B). This predilection for crossing points can also be understood intuitively. In Figure 11, the red line shows the reward obtained for choosing A and the blue line shows the reward for choosing B for a particular fractional allocation to button A. If A is chosen, the fractional allocation to A increases, and the subject is moved rightward on the horizontal axis. If B is chosen, the fractional allocation to A decreases, and the subject is moved leftward on the horizontal axis. Notice the reward functions for the matching shoulders task (Figure 11). As the allocation to A increases, the subject is moved rightward on the horizontal axis, and the reward for choosing A goes down. Consequently, choosing B increases the reward and moves the subject leftward back toward the crossing point. However, if B continues to be selected, the subject moves left of the crossing point and the reward decreases for continued selections to B. Consequently, switching to A increases the reward and again moves the subject rightward back toward the crossing point.

Human Behavior on the Sequential Decision-Making Tasks

On the matching shoulders task (Figure 11), the optimal allocation to button A and the crossing point in the reward functions coincide. In this plot, each closed, green circle represents a single subject, the horizontal

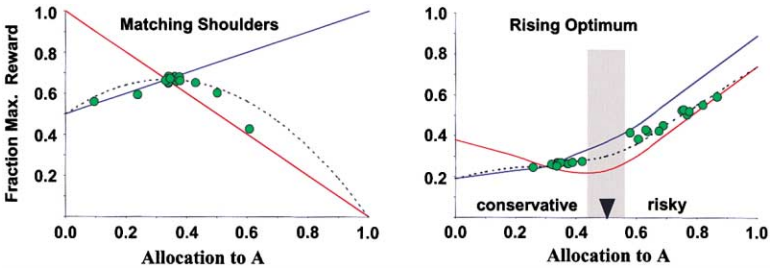

Figure 11. Using a Prediction-Error Signal to Make Economic Choices

The return (height of slider bar) for each choice is a function of the fraction of choices previously allocated to choice A (over the last 20 choices). The subjects do not know this fact and must learn how to act by making choices and receiving immediate returns. At each fractional allocation, *f*, to choice A, the red line gives the return (height of bar) for choosing A next, and the blue line gives the return for choosing B next. The dashed line in each panel represents the optimal curve, that is, the best that could be earned on average if the subject played at a fixed allocation to button A for the entire task. Each task consisted of 250 selections, but the subjects were not told this beforehand nor did they have knowledge of these reward functions.

Each subject was started with a fractional allocation to button of 0.5. This starting point was assigned stochastically, providing some wobble in the exact starting point. This effect can be seen in Figure 12. Other initial conditions were tried and the overall results did not change (data not shown). The reward functions still separated the subjects into at least two distinct groups. The prediction-error model of action choice (Figure 8) experiences a kind of "value illusion," which expresses itself as a strong tendency to play at crossing points in the reward functions (explained in Appendix B). The overall consequence of this value illusion is that the model will give up better long-term returns to stay near crossing points.

Matching shoulders task. Here, the optimal curve (dashed line) and the crossing point in the reward functions coincide (explained in Appendix B). Rising optimum task. The optimal solution is to choose A every time; hence, the optimal curve is an almost monotonically increasing function of the allocation to A. In both panels, each green dot is a single subject; its horizontal position encodes the average allocation to A over 250 selections, and its vertical height encodes the average return per selection earned over the 250 selections. Subjects' equilibrium behavior in the matching shoulders task is nearly optimal, a fact likely to be an artifact of the coincidence of the maximum of the optimal curve and the crossing point in the reward functions. In the rising optimum task, subjects' equilibrium behavior separates them into two distinct groups: conservative (stick near crossing point) and risky (nearly optimize). The black arrowhead indicates the fractional allocation to A that defines the two groups: less than 0.5 allocation to A (conservatives) and greater than 0.5 allocation to A (risky). The gray bar shows the separation of these two groups based on their equilibrium allocations to choice A. This separation into two groups is not merely an equilibrium behavior but is also reflected by detailed choice dynamics on the rising optimum task (see Figure 12).

coordinate encodes the average fractional allocation to button A over the entire task, and the vertical component encodes the average reward received per choice. Except for a few outliers, all subjects choose on average to stay near the crossing point in the reward functions. The model plays almost exactly at the crossing point (data not shown). This is an adaptation of a task used by Herrnstein to address rational choice theory and matching law behavior (Herrnstein, 1990). Here, we see that the matching law behavior emerges near crossing points from the bias in action selection imposed by the prediction-error signal. The apparent optimality of this matching strategy near the crossing point of the matching shoulders task is most likely an artifact of the specific ratio chosen for the slopes of the linear reward functions and is not evidence that all the humans are acting optimally.

In the rising optimum task, the optimal curve (dashed black line) is a nearly monotonic increasing function of the subjects' allocation to button A. As with the matching shoulders task, there is a crossing point in the reward functions, and this crossing point is aligned at the same fractional allocation to button A ($\sim$0.32). As before, each closed, green circle is a single subject. These reward functions separating the subjects fall neatly into two groups, which we have labeled risky and conservative (Figure 11). The conservatives play just as predicted by the model–just to the right of the crossing point in the reward functions (Appendix B). However, the model did not anticipate the behavior of the risky subjects who express a nearly optimal selection strategy. These results would be unremarkable if the groups differed only in their equilibrium behavior; however, as shown in Figure 12, these two groups are vastly different in their choice-by-choice dynamics. In this figure, each point is a mean over subjects or instances of the model, and the error bars are the standard errors of the means.

The difference in these two groups is also evident in a switching task where the subject begins with the matching shoulders reward functions, which are then secretly switched to the rising optimum reward functions midway through the task (selection 125). Figure 12B shows the behavior of these two groups on such a switching task. Each group is categorized by their equilibrium behavior on the rising optimum task and later brought back to perform the switching task. Although none of the subjects are informed of the initial reward functions or the switch, only the subjects categorized as risky sense the switch in reward functions and respond by changing their fractional allocation to button A. As shown, the model anticipates well the behavior of the conservative subjects.

Taken together, the equilibrium behavior and the choice-by-choice dynamics show that the two categories of subjects is a real, measurable distinction. It's clear that both groups can sense the crossing point in the reward function; both risky and conservative are capable of playing at the crossing point in the matching shoulders task. However, the rising optimum task exposes their apparent differential sensitivity to decreases in returns. In this task, as the crossing point is passed from left to right (as allocation to A increases), there is a dip in the return received from choosing button A. In the simplest case, if the past models the future, then such a dip might well be interpreted by some neural mechanism as increased risk for continued selections to

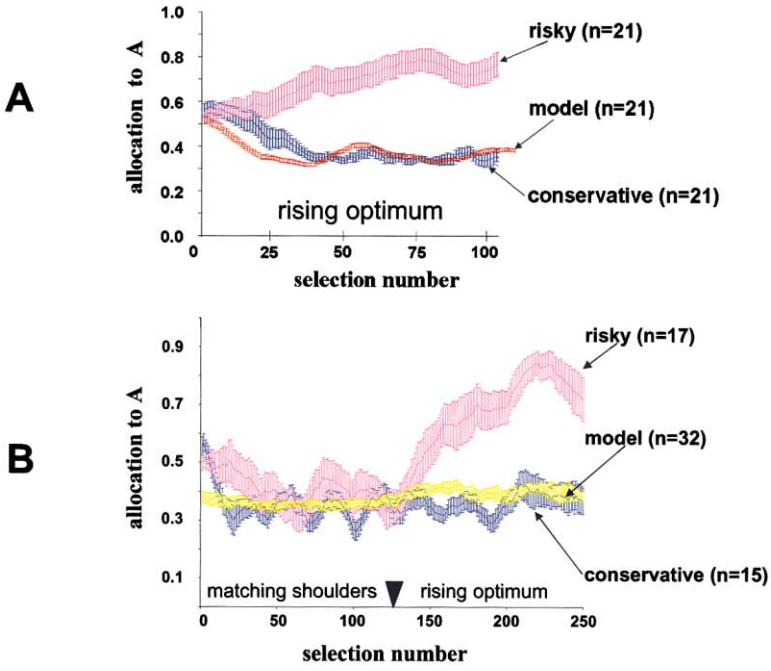

Figure 12. Choice Dynamics Also Distinguish Conservative and Risky Subjects

Subjects were categorized as conservative or risky based on their equilibrium behavior on the rising optimum task illustrated in Figure 11. Here, we plot the choice-by-choice dynamics on the rising optimum task (A) and on a switching task (B) where reward functions starting as matching shoulders were secretly switched to rising optimum after selection number 125.

(A) Choice dynamics on rising optimum task. Fractional allocation to choice A versus raw selection number. Three traces are shown for the first 100 choices on the rising optimum task: (1) conservative human subjects (less than 0.5 average allocation to A on the rising optimum task; blue), (2) risky human subjects (greater than 0.5 average allocation to A on the rising optimum task; magenta), and (3) dopamine prediction-error choice model (Figure 8; $\lambda = 0.93$; red). At each selection number, the central point is the mean, and the error bars represent the standard error of the means. Notice that by 25–30 selections, the two groups are separated and remain so. The model is stable near the crossing points and plays like the conservative human subjects. (humans, $n = 42$; model, $n = 21$).

(B) Switching task for humans. In the bee decision task, the average uncertainty in reward magnitude for blue and yellow were switched. Both real bees and the model bee switched their behavioral allocations in response to this switch (Figure 9). In this switching task, the rewards functions began as matching shoulders and were switched to rising optimum at trial 125. The risky and conservative subjects behave differently after the switch. The risky subjects sense the switch and change their allocation strategy to near optimal while the conservative subjects continue to choose A in a fashion that keeps them near the crossing point in the reward functions. Likewise, the model also chooses to stay near the crossing point in the reward functions. These results show that the risky subjects are capable of playing at the crossing point when this is optimal, but unlike the conservative subjects, they sense and respond readily to unanticipated changes in reward structure.

A. The willingness to play through the dip and continue exploring is one feature that characterizes the capacity of the risky group to discover the nearly optimal strategy. However, this is not a complete explanation since individual conservatives often displayed large excursions into the optimal allocation to A range but after a few selections were driven back toward the crossing point.

Possible Linkage between Brain Response to Predictability Changes and Riskiness

The tight connection of the reward prediction error and the action-selection mechanism suggested to us that there might be a connection between the subjects' expressed risk profile on the rising optimum task and their brain response to changes in predictability for the sequential gustatory stimulus. We had no expectation about the polarity of such a relationship. Remarkably, a regionally specific difference in brain response to changes in predictability paralleled these behaviorally defined labels. This difference can be seen by correlating the allocation to choice A and the brain response to changes in predictability for the sequential gustatory experiment described above (Figure 5). This analysis showed that the response of the left nucleus accumbens to changes in predictability correlates strongly with risky behavior on the rising optimum task (Figure 13; $n = 14$; $p < 0.0005$).

Orbitofrontal-Striatal (OFS) Circuit as a Valuation System

To summarize, a prediction error-based model of how reward expectancy should influence decision-making in humans produced a sequential behavioral task that separated subjects into two groups. These groups could be characterized by their brain response to changes in predictability for sequential gustatory stimuli and by their dynamic and equilibrium performance on a simple two-choice behavioral task (rising optimum). Altogether, the constellation of results reviewed above suggested to us and others (O'Doherty et al., 2001) that a more general function than simple expectation violation was being carried out by the ventral striatum, dorsal striatum, and orbitofrontal circuit (OFS circuit). We strongly suspected the existence of a more generalized valuation function.

We propose that the OFS circuit computes an ongoing valuation of potential payoffs (including rewards), losses, and their proxies (predictors) across a broad domain of stimuli. This is a different proposal from the prediction-error signal discussed above for midbrain dopamine neurons (Montague and Sejnowski, 1994; Montague et al., 1996) and proposed for many other brain regions (Schultz and Dickinson, 2000). The prediction error signal guides the system to learn the time and amount of future rewards, and may, as reviewed above, direct some forms of simple decision-making. Our specific proposal for one function of the OFS is that it computes a valuation of rewards, punishments, and their predictors. By providing a common valuation scale for diverse stimuli, this system emits a signal useful for comparing and contrasting the value of future events that have not yet happened–a signal required for deci-

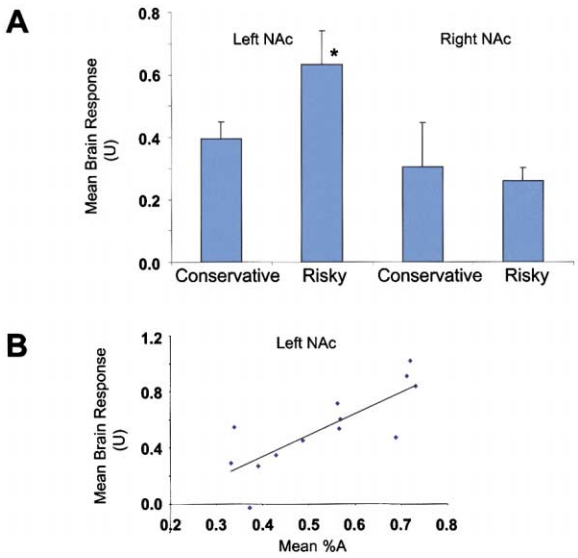

Figure 13. Multiplying Two Experiments: Brain Response that Correlates with Risky Choices

The optimum strategy in the rising optimum task (Figure 11) is to choose A every time; a strategy pursued by half the subjects. This strategy represents risk-taking because a subject must endure a decrease in the returns from A in order to discover the domain of maximal returns (Figure 11; notice the dip in A's reward function just to the right of the crossing point). As demonstrated by their different choice-by-choice dynamics, the risky and conservative subjects were different from the very start of the task, suggesting the hypothesis that their brain response to changes in predictability might also differ (Figure 12). Consequently, we tested the hypothesis that there would be a correlation between fractional allocation to A (higher = riskier) and brain response to changes in predictability for sequential gustatory stimuli.

(A) Brain response in nucleus accumbens (left and right, separately) versus assignment to conservative or risky groups based on performance on rising optimum task (Figure 11). In the left nucleus accumbens, the risky group was significantly different from the conservative group, as indicated by the asterisk ($p < 0.0005$).

(B) Brain response in left nucleus accumbens versus fractional allocation to choice A ($n = 14$). Each point is a subject.

sion-making algorithms that assign attention, plan actions, and compare disparate stimuli (see O'Doherty et al., 2001). Below, we consider generally how this working hypothesis can be converted into quantitative predictions of neural signals that participate in ongoing valuation.

The Predictor-Valuation Model

Common Scales through Internal Currencies

Any economic system possesses three basic real-world features: (1) markets–goods and services desired by consumers, (2) currency–some way to represent the value of goods and services, and (3) limited resources. In biology, the markets for most creatures are pretty clear. Creatures must obtain adequate food and rest in order to acquire the most important resource–mates, and hence, offspring. The idea of limited resources is also a clear constraint in the biological world. Creatures that take an excessive amount of time or effort acquiring food, mates, and safety will be less successful than creatures that carry out these functions quickly and efficiently. The most interesting connection to neural systems is the idea of a currency.

A currency is an abstract way to represent the value of a good or service. For our purposes in this paper, it possesses an important property: it provides a common scale to value fundamentally incommensurable stimuli and behavioral acts. For example, suppose we want to understand the relative value of 17 coconuts and 41 sips of water. There is no natural way to combine coconuts and sips of water; however, each can be converted to their valuation in some currency, and the values can be combined in any number of ways. This kind of abstraction is so common in our everyday world that its biological substrates go virtually unnoticed.

Without internal currencies in the nervous system, a creature would be unable to assess the relative value of different events like drinking water, smelling food, scanning for predators, sitting quietly in the sun, and so forth. To decide on an appropriate behavior, the nervous system must estimate the value of each of these potential actions, convert it to a common scale, and use this scale to determine a course of action. This idea of a common scale can also be used to value both predictors and rewards.

Cost of Believing and Acting on a Predictor

A predictor of future reward acts as a promise to the nervous system that a certain amount of reward will be delivered at a specified future time. In the models that we have reviewed above, only the amount and time of the reward were important for driving learning. However, a behavioral act or fixed amount of some rewarding substance does not possess a fixed value to the organism; rather, the value of a reward can change dramatically as new, unexpected information arrives. Suppose that a red light predicted 10 ml of water 1 min in the future. If everything goes as expected, then 1 min after the light the system can expect 10 ml of water–the system can plan actions accordingly. However, suppose that 5 ml is unexpectedly delivered 30 s after the red light and suppose that this is a rare event, not something that will systematically happen in the future. What happens to the value of the red light as a predictor of 10 ml of water? Surely, the value of the 10 ml is less because of this unexpected event. Should the same actions continue to be planned for 1 min based on the red light? Should the decreased value of the reward also cause a decrease in the value of the predictor for that reward?

To value a predictor, a neural system must have a way to compute the predictor's value before the reward that it promises actually arrives. One difficulty with this proposition derives from the uncertainty associated with the time interval extending from predictor onset to the expected future time of reward delivery. However, this uncertainty must be handled correctly because predictions about future reward represent a real cost to the creature. Believing a predictor means that processing time is tied up and behavioral resources committed as actions are prepared. In a system with finite processing capacity, finite resources for planning potential actions, and finite resources for output behaviors, the continued belief in the promise of a predictor is a potentially costly commitment. It follows that there must exist neural signals that provide an ongoing valuation of both predictors and potential future rewards. Below, we develop a sim-

ple model of such valuation, and show how it may actually be represented at the single cell level as changes in spike production that occur prior to the arrival of a predicted rewarding stimulus. This model should apply equally well to both rewards and punishments; however, for convenience we refer only to predictors of reward.

Diffuse and Discount Strategy Produces Predictor-Valuation Model

Any valuation scheme for reward predictors must take account of two important principles. Principle 1: any estimate of future reward is not exact. Uncertainty accrues with time and more uncertainty will accumulate for reward estimates in the distant future than for those in the near future. Principle 2: there is risk associated with the future time that separates the predictor from future reward; therefore, there must be some discounting of time.

The two principles are meant to distinguish two different effects. (1) A dynamical estimate of the future is not exact. As time passes, the uncertainty (error) in the estimate will accumulate. (2) The value to most creatures of a fixed return diminishes as a function of the time to payoff, that is, the time from now until the reward arrives. Two examples below may help illustrate these principles.

Principle (1) example: I build a dynamical model of weather. My prediction 2 days hence is more certain than my prediction 10 days hence because error in my initial estimate builds up with time. The only question now is the nature of the model that captures quantitatively the build-up in uncertainty. Below, we choose a simple diffusion approach to the accumulation of uncertainty in future reward.

Principle (2) example: "I'll give you $100 in 1 min or $100 in 1 year." Which do you take? The answer for most humans is clear, and we have established the relative value of the two choices based only on a difference in the promised time of delivery. The only issue now is the value of the time to reward. Just as the bee's valuation of reward variability can be measured behaviorally, the human can likewise be queried. $100 in 1 min or $1000 in 1 year, $100 in 1 min or $10,000 in 1 year, $100 in 1 min or $100,000 in 1 year, $100 in 1 min or $1,000,000 in 1 year, and so on. We could arrive at a person's valuation of 1 year (scaling factor and offset) quite quickly.

Diffusing the Reward Estimate

The first point to make is to reemphasize that the valuation of a reward predictor is in units of the value of future predicted reward and is not simply related to the amount of future reward. As indicated in Figure 14, n is perceptual or experiential time, that is, the animal's internal index for the present (now). Let $S(\hat{r}(x))$ be the value of the estimated reward $\hat{r}(x)$ defined for time x. In this paper, we simply assume that value $S(\hat{r}(x))$ is proportional to $\hat{r}(x)$, and leave our development in terms of $\hat{r}(x)$. Assume that some sensory cue, which has formerly acted as a reward predictor, is experienced at perceptual time n. This cue is associated with a stored estimate of the likely reward for all time (future and past of n). In this sense, the cue evokes a function that expresses what likely has happened (past of n) and what will likely happen (future of n).

Since $\hat{r}(x)$ is an estimate of the true reward function, it will not be exact. Now we implement principle 1 from above, that is, more time before experiencing a promised fluctuation in reward means more uncertainty in the estimate of that fluctuation. We capture this formally by letting the estimate $\hat{r}(x)$ diffuse, but differentially as a function of the time it would take from time n ('now') to reach the future time x. This maneuver generates a new uncertainty-adjusted estimate $\hat{R}(x,n)$ that has scaled the uncertainty in the reward associated with future times according to the time it will take to reach those future times:

$$\hat{R}(x,n;D) = \int_{-\infty}^{+\infty} dy\, G(x-y,(x-n)D)\hat{r}(y), \qquad (1)$$

where $G(z,b) = (2\pi b)^{-1/2}\exp\{-z^2/2b\}$ and D is a constant.

The dependence of the diffusion on the time to reach a future reward can be seen by delivering three equal-sized fluctuations in reward at evenly spaced times to the future of n (Figure 14). For example, these fluctuations could represent water squirts in the mouth of a thirsty creature; that is, the vertical axis would represent volume of water. These fluctuations were Gaussian and had a width (standard deviation) of 0.05 s. There are two main effects to notice in the uncertainty-adjusted estimate $\hat{R}(x,n;D)$ (red trace) of the initial reward estimate $\hat{r}(x)$ (black trace). (1) The uncertainty grows with distance into the future, and (2) the amplitude necessarily decreases.

Discounting the Future

The development above imposed a simple model for how uncertainties in reward fluctuations accrue as a

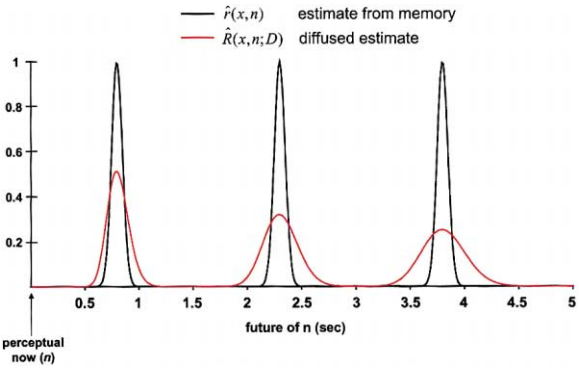

Figure 14. Uncertainty-Adjusted Reward Estimate

Example of how the uncertainty-adjusted estimate of future reward ranks identical rewards that arrive at evenly spaced times to the future of n (perceptual now). The solid black trace is a plot of the initial reward estimate $\hat{r}(x)$, composed of three Gaussian fluctuations expected to arrive at different, but evenly spaced, future times. As an example, the trace could represent the amount of water delivered as a function of time. In this case, the Gaussian fluctuations are squirts of water. Each peak has a standard deviation of 0.05 s. Using this initial estimate, the uncertainty-adjusted reward estimate $\hat{R}(x,n;D)$ is shown as a solid red trace (equation 1; D = 1/10, n = 0). Two features are evident. (1) The uncertainty in the reward estimate is seen to increase with increasing time in the future. (2) The amplitude of $\hat{R}(x,n;D)$ diminishes with increasing time. The uncertainty accrues at a rate proportional to the time it will take to reach each future point starting at perceptual time n. This representation captures the requirements of principle 1 (text), that the distant future will add more uncertainty than the near future. Simulated curve courtesy of Dr. Phillip Baldwin (unpublished data).

function of the time that will elapse before the promised fluctuations are experienced. In our case, we allowed the stored estimate $\hat{r}(x)$ (from memory) to diffuse differentially as indicated in equation 1. This adjusts our model of future reward to account for principle 1. Once our new uncertainty-adjusted estimate $\hat{R}(x,n;D)$ is generated, the system must now assign a value to the expected reward for all times to the future of n. We must choose a way to discount the value of the expected reward at future times.

There is no global time discount rate in the brain; however, one can make a reasonable argument that exponentially discounting the future is a straightforward idea that could be implemented locally through time. The basic idea is to treat the reward predictor as something that consumes resources and try to decide, in each small instant, whether it is worth continued processing. The argument for exponential discounting can be made with an example.

Assume that there is a predictor that promises an impulse of reward at some relatively long time, t^*, to the future of n. During a small time step to the future of n, a reward more immediately valuable than the reward promised by the predictor may present itself. In order to know whether to continue processing the current predictor-reward pair, the system must be able to value the predictor before its expected reward arrives. This perspective suggests a simple strategy for valuing a predictor: if the current value of the predictor is greater than the value of what could otherwise be gained immediately, then stay with processing the predictor, otherwise switch to processing the immediate return. Switching to the more valuable immediate return means that the system forgoes the expected future return promised by the predictor. This strategy translates directly into a method to value the predictor during the time interval between its onset at any time, n, and reward delivery at some later time, t^*. Divide the interval $t^* - n$ into N equal intervals of size $\Delta t = (t^* - n)/N$. Let q be the probability per unit time that an event occurs in Δt that is more valuable than the current value of the predictor (which we have not yet specified). Now implement the strategy above. The system should stay with processing the predictor if no event occurs in the first time interval, Δt_1, that is more valuable than the current value of the predictor. The probability that no such event occurs during Δt_1 is 1 minus the probability that it does occur, that is, $1 - q\Delta t_1$. The probability of staying with the predictor through the entire interval is the probability that no such event occurs in all N subintervals:

$$\begin{aligned} P_{stay}(N) &= \begin{pmatrix}\text{Prob. that event does}\\ \text{not occur in } \Delta t_1\end{pmatrix} \cdot \begin{pmatrix}\text{Prob. that event does}\\ \text{not occur in } \Delta t_2\end{pmatrix} \cdots \\ &\quad \begin{pmatrix}\text{Prob. that event does}\\ \text{not occur in } \Delta t_N\end{pmatrix} \\ &= (1 - q\Delta t_1)(1 - q\Delta t_2) \cdots (1 - q\Delta t_N) \\ &= (1 - q\Delta t)^N \\ &= (1 - q(t^* - n)N^{-1})^N. \end{aligned} \quad (2)$$

In the limit of large N, P_{stay} becomes $e^{-q(t^*-n)}$, the probability that nothing more valuable arrived during the interval between predictor and expected reward at time t^*.

Predictor-Valuation Model

We can now combine the diffuse and discount steps described above. The diffusion part captured the fact that less uncertainty will accrue with less time to wait into the future. The discounting argument showed that the system can easily accomplish exponential discounting if it simply monitors a signal that tells it the highest value of immediately available rewards. The diffusion part produced a new uncertainty-adjusted estimate, $\hat{R}(x,n)$, of reward that should be discounted exponentially through time according to the time it will take to reach each point in time x to the future of the present experienced time n, and all of these contributions from the future must be added up to produce the current value $F(n)$ of the predictor at perceptual time n:

$$\begin{aligned} F(n) &= \int_n^{+\infty} dx\, e^{-q(x-n)} \int_{-\infty}^{+\infty} dy\, G(x - y, (x - n)D)\, \hat{r}(y) \\ &= \int_n^{+\infty} dx\, \{e^{-q(x-n)}\} \cdot \{\hat{R}(x,n;D)\} \\ &= \int_n^{+\infty} dx\, \{\text{discount future time } x \text{ relative to perceptual time } n\} \cdot \\ &\quad \{\text{diffused version of reward estimate } \hat{r}(x) \text{ for same } x \text{ and } n\} \end{aligned} \quad (3)$$

This integral equation is the predictor-valuation model and expresses the value of a predictor as a function of current perceptual time n. It can also be expressed as a differential equation that produces a number of interesting solutions related to current electrophysiological experiments on reward prediction and valuation (Montague and Baldwin, unpublished data).

Neural Economics: Applying the Predictor-Valuation Model to Neural Data

The stay or switch argument above results in a scheme for continuously deciding whether the current value of a predictor is worth continued processing (investment). For reward prediction, the predictor-valuation model predicts an escalating increase in some signal up until the time of significant future fluctuations in reward. Near the perceptual present, this escalation should be approximately exponential, especially in simple experimental scenarios. Near the time of future reward delivery, the shape of the function is more complicated because it depends on the exact reward estimate, the value of D, and time to reach the future point. The model generates a fairly rich set of predictions given its simple diffuse and discount derivation. As described below, it may indeed provide insight into neural responses measured in prefrontal cortex and dorsal striatum and may also provide a biological substrate, or rather justification, for economic models for the valuation of market options (Montague and Baldwin, unpublished data).

Our working hypothesis, as expressed above based on fMRI data, is that the orbitofrontal cortex and striatum are the likely sites to participate in such an important valuation function. The measured single-unit responses to reward processing in orbitofrontal cortex are extremely heterogeneous. Some neurons respond only to reward-predicting cues and others only to delivery of liquid or food reward (Schultz et al., 2000). Nevertheless, there is one major class of neural response anticipated very well by the predictor-valuation model described above. Figure 15B shows the response of a neuron from orbitofrontal cortex in an alert primate during the period

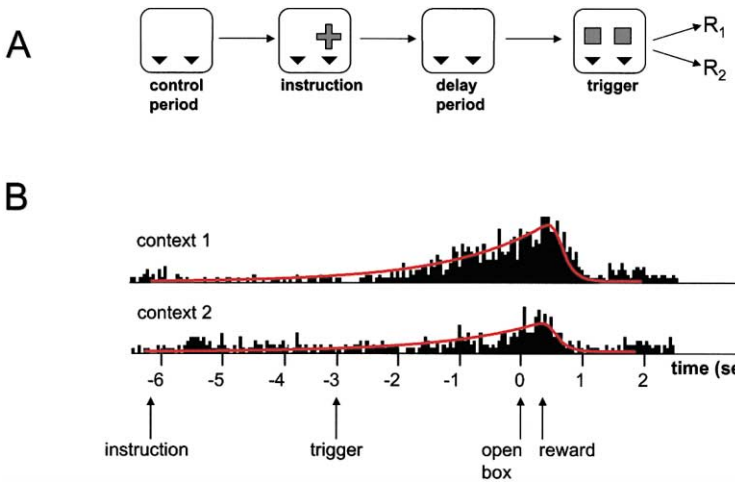

Figure 15. Escalating Activity Increase during Delay Period

(A) Spatial delayed-response task given to alert monkeys. Monkey holds down a press key and is presented with a control screen for 2 s, after which an instruction screen delivers two pieces of information: (1) which lever to press (right or left) on subsequent trigger stimulus and (2) which reward could be expected. Second delay is 2.5–3.5 s. Trigger: two identical squares are presented, monkey releases press key and selects (if correct) the right or left lever (black triangle), as indicated by the instruction.

(B) Changes in amplitude of exponential build-up during delay period depend on the relative value of the terminal reward. Activity in a single orbitofrontal neuron during spatial delayed response task where reward is a piece of apple, but the apple is presented in two separate contexts. In context one, the two possible rewards are cereal and apple, and in context two, they are raisin and apple. Similar to (B), both contexts produce an escalating increase in activity during the period from instruction to reward, but when apple is the more preferred reward (context one; apple versus cereal), the amplitude of this increase is larger than when apple is the less preferred reward (context two; apple versus raisin).

The influence of this change in valuation of the reward could be expressed in the prediction valuation model either through changes in q, changes in the function, S, that values the reward, r, or in some combination of variables on which F depends (equation 3 in text). The red line is a fit of the predictor-valuation model to the spike rate histogram. Fits courtesy of Dr. Phillip Baldwin (unpublished data) (adapted from Tremblay and Schultz, 1999 [copyright 1999 by Nature, www.nature.com]).

between an instruction cue and the subsequent delivery of reward at the time indicated. The animal was carrying out a spatial delayed-response task (Figure 15A) in which an instruction cue is illuminated and delivers two pieces of information: (1) which of two targets the animal should select (right or left) and (2) which reward will be delivered for a correct response. In the figure, two traces are shown for the same neuron. These traces illustrate the relative difference in the neuronal response to the same reward (apple) but in comparison to other rewards (raisin, cereal). In behavioral assays, the animal's preference was raisin > apple > cereal. The response in Figure 15B (top) is for apple versus cereal, but the same stimulus (apple) causes a smaller overall response during the delay period for apple versus raisin ([B], bottom). In both cases, the increase in activity occurs in an exponential fashion throughout the trial and decays back to baseline (at a different rate) from the time of reward delivery. The figure legend discusses a possible substrate for this change using the predictor valuation model.

Notice the increase in spike production beginning somewhere near the instruction and increasing (roughly) exponentially up to the time of reward, after which it decays back to baseline levels. In addition, there are different time constants for the activity build-up up to reward delivery and the decay after reward delivery. The solution to the predictor-valuation model (equation 3 above; red lines in Figure 15B) anticipates both the exponential build-up and the different time constants (the fit to these data was generously provided by Dr. Phillip Baldwin as a personal communication).

The escalating response is remarkable for a cortical pyramidal neuron, which is well known to show strong spike rate adaptation to inputs. It is not a trivial biophysical or circuit level task to make the neuron produce spikes in this escalating fashion. In our view, it is this basic trend that may well represent the valuation of the predictor throughout the trial and up until the time of actual reward delivery.

Discussion

We have reviewed neuroimaging, neurophysiological, and behavioral data and their correlations with respect to both reward and valuation. Together, these results suggest that the OFS circuits act to generate a common internal currency (scale) for the valuation of payoffs, losses, and their proxies (predictors of payoffs and losses; see O'Doherty et al., 2001). Our focus was narrow and primarily neural, using behavioral evidence primarily to emphasize the neural responses on which our review concentrated. We showed that these qualitative observations may be upgraded to a quantitative biological model of reward predictor valuation. The two fundamental components of this model were: (1) diffusion in time to account for future uncertainty and (2) discounting in time to allow for the possibility that better alternatives may intervene in the future. Through this model, we were able to account in detail for the functional form of electrophysiological activity in orbitofrontal. We now discuss the connection of the predictor-valuation model to both behavior and finance.

Neurons or Behavior?

The predictor valuation model, while inspired by economic ideas, is meant to explain the response of a specific set of circuits in the brain–the OFS. Indeed, we showed that the model anticipates the functional form of neural activity in the OFS during delay periods, separating reward-predictors and their promised rewards. Economists, however, are interested in understanding the behavior of people (microeconomics) or markets (macroeconomics). A question naturally arises: How are the neural data and economic constructs connected? The notion of framing economic behavior in terms of conflicting behavioral tendencies within an individual is

not new (Ainslie, 1992). Here, we are proposing a simple scaling principle. Economies are composed of people acting according to their internal valuation of events, goods, services, and so forth. The argument is simple: a model of neuronal valuation in individuals should have explanatory power at both the micro- and macroeconomic level and especially in narrowly defined markets.

Reward versus Valuation

We have made a critical distinction between the concepts of reward and valuation. Reward refers to the actual stimulus, be it food, money, drugs, etc., but as noted earlier, there is no straightforward way to equate these very different entities. Similarly, the types of stimuli that predict reward may also take on vastly different forms. Both neurophysiological and brain imaging data strongly implicate the OFS as a common pathway for the representation of both rewards and predictors. Indeed, even faces of attractive women can activate this circuit (Aharon et al., 2001), as well as nonrewarding events like noxious thermal stimuli (Becerra et al., 2001). If all these different stimuli evoke activity in the OFS, then what is the nature of the representation? As proposed here, we believe it to be value. The notion of value is well known to economists–for example, a guaranteed $10 today is more valuable than $10 promised for 5 years from now. Indeed, this example illustrates one form of valuation–discounting, but this is only one aspect of the model proposed here. In addition to discounting, one needs both a common currency and an accounting for uncertainty associated with the future. By converting into such a currency, the value of reading a book can be weighed against working a few more hours. Both have very different "rewards," but converting them to a common valuation scale allows them to be compared. The data reviewed here strongly suggests that part of this representation is present within the OFS.

Faster-than Exponential Discounting

One behavioral construct of valuation that has captivated economists is the problem of intertemporal choice (Lowenstein and Elster, 1992). It has been observed across a wide range of species, ranging from pigeons to humans, that animals prefer a smaller, immediate reward over a larger, delayed one. For example, given the hypothetical choice of receiving $100 immediately or $200 in two years, most people will choose the $100 now; the same people do not prefer $100 in 6 years to $200 in 8 years (Lowenstein and Elster, 1992; Thaler, 1981). Aspects of these findings in humans may be attributable to an element of trust, i.e., the belief that such payoffs will actually occur in the distant future, but this trust component is indistinguishable from the uncertainty that time itself imposes on the belief of a payoff. The replication of similarly nonrational choices in other animals suggests that time value is represented fundamentally as part of any neural encoding scheme.

A remarkably simple assay of valuation is the rate at which pigeons will peck for a reward (Herrnstein, 1961). By varying the amount of the reward and the time of delivery, one can estimate the internal valuation structure, namely the value of time. Even in pigeons, valuations are said to be dynamically inconsistent, and this observation has been used to explain everything from rate of savings (Laibson, 1997) to alcoholism (Heyman, 2000). Early economic models of so-called discounted utility posited a decaying exponential relationship between delay and value, but subsequent experiments in both pigeons (Mazur, 1988) and humans (Lowenstein and Prelec, 1992) suggested that this relationship is more concave than an exponential, possibly hyperbolic. To our knowledge, no biological explanation for this relationship in terms of identified neural circuits has been offered. Evolving the concept of reward into one of valuation, however, offers one explanation for such apparently "irrational" behavior. From first principles, the two-step diffuse-and-discount formulation leads to an overall discounting that is faster than a single exponential and will lead to the same choice reversals that have been observed across a variety of species.

Diffuse and Discount Produces a Connection to the Black-Scholes Equation

The predictor-valuation model was derived above by imposing a diffusion model of how to adjust the initial reward estimate $\hat{r}(x)$ by the uncertainty that will accrue in the future. The value of the new, uncertainty-adjusted estimate $\hat{R}(x,n)$, was then exponentially discounted backward through time to psychological now (n). This exponential discounting was based on the idea that processing the predictor consumes system resources, and its processing should therefore be continuously reevaluated. These two components produced the predictor-valuation model, which anticipated several important features of single-unit responses from orbitofrontal cortex and dorsal striatum. Detailed experimental tests of this model await future work. However, it should be noted that the predictor-valuation model has a remarkable relationship to economic theories (Black and Scholes, 1973) that seek to value hedged portfolios in an efficient marketplace (Montague and Baldwin, unpublished data).

Although no arguments about hedging and efficient markets were marshaled in support of the predictor-valuation model, it can be shown that the model in equation 3 is analogous to the formulation first proposed by Black and Scholes as a method to set a fair price for options on securities (Montague and Baldwin, unpublished data). This odd connection may simply be a coincidence; however, it is possible that the connection between the two approaches is symptomatic of a more fundamental biological connection. The Black-Scholes class of equations was initially developed in an effort to provide a principled approach to the way that options should be priced; that is, they sought a normative solution that matched real market data. The eventual price of options in a real market is set by the behavior of lots of individual brains expressing their valuations through a propensity to buy or sell at specific prices. These brains have long been equipped with rapid valuation mechanisms crafted to deal with vast number of stimuli and possible behavioral acts available to them. We derived the predictor-valuation model based on the need to extend the reward prediction-error model to include the way that the system should value future promises of reward; in particular, promises made by reward predictors. Our assumptions were based on the simplest model through which uncertainty adjusted the system's current estimate of reward. We suspect that the equation discovered initially by Black and Scholes and extended by Merton (1990) may have hit upon a form of solution long-ago embedded in the valuation systems

present in hominid brains. In this sense, their derivation in terms of a hedged portfolio simply led them to a class of equation describing the valuations carried out by the individual brains that compose any marketplace. Such biological connections suggest that brain science may well provide constraints that can help stabilize certain markets.

Conclusion

In summary, the predictor-valuation model suggests identifiable neural substrates that may support sophisticated economic evaluations of diverse stimuli. This interpretation is strengthened by the kind of learning and adaptation displayed by dopaminergic circuits during reward-dependent learning and the possible influence of this signal in sequential decision making. Altogether, we strongly suspect that a new generation of electrophysiological results in animals and neuroimaging results in humans may well forge a connection between neural responses and direct measures of economic behaviors. A connection that should provide insights into the valuations carried out by individual nervous systems and their quantitative relationship to valuations carried out by real markets.

Appendix A: Reinforcement-Learning and the TD Prediction-Error Model of Dopamine Function

There are three basic components to every reinforcement-learning system: (1) a reward function, (2) a value function, and (3) a policy. These relatively abstract terms capture the idea of immediate evaluation (reward function), long-term judgment (value function), and action selection (policy):

The reward function formalizes the idea of a goal for a reinforcement learning system. It assigns to each state of the agent a single numerical quantity–the reward. The reward function defines what is good right now and can be viewed as a built-in assessment of each state available to the agent (learner). It is also used to define the agent's goal: to maximize the total reward.

The value function formalizes the notion of longer-term assessments (judgements) about each state of the agent. It provides a valuation of the current state of the agent taking into account the succession of states that could follow. Formally, for each state, value is defined as the total amount of reward the agent can expect from that state forward into the distant future. These values would have to be stored in some fashion within the agent. In practice, the learner uses the reward function to improve its internal estimate of the value function.

In short hand, rewards are immediate and values are long-term. For example, a rat may take many steps across an electrified grid (low reward) to reach food (high reward). All those intermediate states (steps on the grid) have very low reward but possess high value because they directly lead to future states with food (high reward).

A policy formalizes exactly what the word implies: "given this, do that." Formally, a policy maps states to actions. In both biological and machine-learning examples, a policy is usually probabilistic. For a given state, a policy defines the probability of taking one of many possible actions to end up in one of many succeeding states.

These three components are used in combination with some model of the environment to produce a reinforcement learning system. They are clearly present in the decision-making model presented in the text.

The computational goal of learning is to use a set of sensory cues $x(t) = \{x_1(t), x_2(t), x_3(t), \ldots\}$ (e.g., characterizing the current state of an organism) to fit a "value" function $V^*(\mathbf{x}(t))$ that values the current state as the average discounted sum of all future rewards from time t onward:

$$V^*(\mathbf{x}(t)) = E\{\gamma^0 r(t+0) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \cdots\}. \quad (4)$$

E is the expected value operator (the average). r(t) is the reward at time t, r(t + 1) is the reward at time t + 1, and so on. γ is a discount factor that ranges between zero and one and captures the idea that rewards in the near future are more valuable than rewards in the distant future. If the true (optimal) $V^*(\mathbf{x}(t))$ could be estimated by a system, then the system could use such an estimate to update its internal model of future rewards and future actions predicated on the expected receipt of those rewards. This would give the system a way to simulate possible future action sequences and value them according to their expected long-term returns.

Adjusting the Predictions (Weights)

The strategy of TD learning is to use a set of sensory cues $\mathbf{x}(t) = \{x_1(t), x_2(t), x_3(t), \ldots\}$ present in a learning trial along with a set of adaptable weights $\mathbf{w}(t) = \{w_1(t), w_2(t), w_3(t), \ldots\}$ to make an estimate $V(\mathbf{x}(t))$ of the true $V^*(\mathbf{x}(t))$. In this formulation, the weights act as predictions of future reward. For completeness, we add here a remark about the weights. The weight associated with each sensory cue, e.g., $w_1(t)$ associated with sensory cue 1, is actually a collection of weights, one for each time point following the appearance of sensory cue 1.

Local Data Anticipate Long-Term Reward

The difficulty in actually adjusting weights to estimate $V(\mathbf{x}(t))$ is that the system (i.e., the animal) would have to wait to receive all its future rewards in a trial r(t + 1), r(t + 2), r(t + 3) . . . to assess its predictions. This latter constraint would require the animal to remember over time which weights need changing and which weights do not. Fortunately, there is information available at each instant in time that can act as a surrogate prediction error. This possibility is implicit in the definition of $V^*(\mathbf{x}(t))$ since it satisfies a condition of consistency through time:

$$V^*(\mathbf{x}(t)) = E\{r(t) + \gamma V^*(\mathbf{x}(t+1)) - V^*(\mathbf{x}(t))\}. \quad (5)$$

Since the estimate V satisfies the same condition, an error, δ, in the estimated value function V (estimated predictions) can now be defined using information available at successive timesteps, i.e., taking the difference between both sides of the above equation and ignoring the expected value operator E for clarity.

$$\delta(t) = r(t) + \gamma V(\mathbf{x}(t+1)) - V(\mathbf{x}(t)). \quad (6)$$

δ is called the TD error and acts as a surrogate prediction error signal which is instantly available at time $t + 1$. If the estimated predictions are correct then $V^*(\mathbf{x}(t)) = V(\mathbf{x}(t))$, and the average prediction error is zero, i.e.,

$E[\delta] = 0$. In other words, if the system can adjust its weights (predictions) appropriately, then it can learn to expect future rewards predicted by the collection of sensory cues.

Appendix B: Decision Tasks

Linear Reward Functions

In the matching shoulders task illustrated and described in Figure 11, the reward functions, r_A and r_B, are linear in f (fractional allocation to choice A), that is,

$$r_A(f) = k_A + m_A f \quad \text{and} \quad r_B(f) = k_B + m_B f. \tag{7}$$

The crossing point of the linear reward functions occurs when $\Delta r = r_B - r_A = 0$, that is, at the allocation f_c where:

$$f_c = \frac{k_B - k_A}{m_A - m_B}. \tag{8}$$

The Optimal Return is Quadratic in Average Allocation to A

The linearity of the reward functions also implies that the average return is equal to the return on the average allocation, that is,

$$\langle r_A(f)\rangle = \langle k_A + m_A f\rangle = k_A + m_A\langle f\rangle = r_A(\langle f\rangle). \tag{9}$$

Therefore, the expected value of the return R for an average allocation $\langle f\rangle$ to choice A is:

$$\begin{aligned} E(R) &= \langle f\rangle r_A(\langle f\rangle) + (1 - \langle f\rangle) r_B(\langle f\rangle) \\ &= (m_A - m_B)\langle f\rangle^2 + \\ &\quad (k_A + m_B - k_B)\langle f\rangle + k_B. \end{aligned} \tag{10}$$

Expression 10 is found by plugging in the linear approximations from equation 7, and it illustrates why the optimal average return, as a function of f, should possess a quadratic dependence on the average allocation to choice A. This fact is illustrated by the dashed line in Figure 11 and represents the best average return that could be achieved at each allocation to A.

Matching Shoulders Task: Why Do the Optimum and Crossing Point Coincide?

We show below that f_c, the allocation to A at the crossing point, coincides with the optimal allocation to A for the matching shoulders task. This is also a useful result for cases where the reward functions are nearly linear in the vicinity of the crossing point.

We seek $\langle f\rangle_e$, the average allocation to A that extremizes $E(R)$, the expected value of the total reward (equation 10 from above). Differentiate $E(R)$ with respect to $\langle f\rangle_e$, set the result to 0, and solve for $\langle f\rangle_e$:

$$\frac{\partial E(R)}{\partial\langle f\rangle} = 2(m_A - m_B)\langle f\rangle + (k_A + m_B - k_B). \tag{11}$$

Setting this derivative to 0 yields $\langle f\rangle_e$:

$$\langle f\rangle_e = \frac{k_B - k_A - m_B}{2(m_A - m_B)}. \tag{12}$$

The essential condition that defines the matching shoulders reward functions for choices A and B is $r_A(0) = r_B(1)$, which implies that $k_A - k_B = m_B$ (see equation 7 above). Substituting this value into equation 12 yields:

$$\langle f\rangle_e = \frac{k_B - k_A - k_A + k_B}{2(m_A - m_B)} = \frac{k_B - k_A}{m_A - m_B} - f_c \tag{13}$$

This shows that $\langle f\rangle_e = f_c$ (see equation 8 above). Since the optimal allocation curve is concave downward, this extremum is a maximum.

A Value Illusion: Action-Choice Model Driven by Prediction Error Is Strongly Attracted to the Crossing Point in Reward Functions

In the matching shoulders task, the reward functions are linear in f, the fractional allocation to choice A. In the rising optimum task, the reward functions are approximately linear near the crossing point; therefore, they inherit the conditions from above (see equations 7–13). In these cases, the attraction of the model for the crossing point can be understood directly by examining how the probability for choosing A changes as a function of f.

The likelihood, P_A, for choosing A is a sigmoidal function of Δw (see text), the difference $(w_B - w_A)$ in the weights associated with choices A and B. At each selection, the weight association with the selection is updated by a simple delta rule:

$$\Delta w_i = \lambda\delta_i = \lambda(r_i - w_i), \tag{14}$$

where $i = A, B$. If the learning rate, λ, is set appropriately, the weights will (on average) track the reward functions so that

$$\Delta r \cong \Delta w. \tag{15}$$

Substituting Δr for Δw in the equation for the likelihood of choosing A yields P'_A, a useful approximation for P_A:

$$\begin{aligned} P_A(f) &= (1 + e^{\mu\Delta w)-1} \cong (1 + e^{\mu\Delta r})^{-1} \\ &= (1 + e^{\mu(k_d + m_d f)})^{-1} = P'_A(f), \end{aligned} \tag{16}$$

where $k_d = k_B - k_A$ and $m_d = m_B - m_A$. Now ask how P'_A changes as a function of f by simply differentiating it with respect to f. For a fixed μ, we obtain

$$\begin{aligned} \frac{\partial}{\partial f} P'_A(f) &= \frac{\partial}{\partial f}(1 + e^{\mu\Delta r})^{-1} \\ &= -\frac{a m_d e^{m_d f}}{(1 + b e^{m_d f})^2} \text{ for positive constants a,b .} \end{aligned} \tag{17}$$

To the right of the crossing point, $m_d > 0$, $P_A(f)$, $m_d < 0$, $P'_A(f)$ strictly decreases ($\partial P'_A/\partial f$ is negative) and approaches zero very rapidly for increasing f. To the left of the crossing point, $m_d < 0$, $P'_A(f)$ strictly increases ($\partial P'_A/\partial f$ is positive) and approaches $(1 + e^{\mu k_d})^{-1}$ for decreasing f. In the matching shoulders task, $k_d = -1/2$, making $P_A(0; \mu = 2) \cong 0.73$. This shows why the decision model gets stuck near the crossing point.

Acknowledgments

This work was supported by grants R01 MH52797 and R01 DA11723, the Kane Family Foundation (P.R.M.), and grants KO8 DA00367 and RO1 MH61010 (G.S.B.). We also wish to thank Phillip Baldwin, Peter Dayan, Ron Fisher, and Sam McClure for comments on an earlier version of this manuscript.

References

Aharon, I., Etcoff, N., Ariely, D., Chabris, C.F., O'Connor, E., and Breiter, H.C. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. Neuron *32*, 537–551.

Ainslie, G. (1992). Picoeconomics. The Strategic Interaction of Successive Motivational States within the Person (Cambridge: Cambridge University Press).

Becerra, L., Breiter, H.C., Wise, R., Gonzales, R.G., and Borsook, D. (2001). Reward circuitry activation by noxious thermal stimuli. Neuron *32*, 927–946.

Berns, G.S., Cohen, J.D., and Mintun, M.A. (1997). Brain regions responsive to novelty in the absence of awareness. Science *276*, 1272–1275.

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

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

Bischoff-Grethe, A., Proper, S.M., Mao, K., Daniels, K.A., and Berns, G.S. (2000). Conscious and unconscious processing of nonverbal predictability in Wernicke's area. J. Neurosci. *20*, 1975–1981.

Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. J. Pol. Econ. *81*, 637–654.

Breiter, H.C., Gollub, R.L., Weisskoff, R.M., Kennedy, D.N., Makris, N., Berke, J.D., Goodman, J.M., Kantor, H.L., Gastfriend, D.R., Riorden, J.P., et al. (1997). Acute effects of cocaine on human brain activity and emotion. Neuron *19*, 591–611.

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

Dayan, P., and Abbott, L.F. (2001). Theoretical Neuroscience. Computational and Mathematical Modeling of Neural Systems (Cambridge, MA: MIT Press).

Dayan, P., and Balleine, B.W. (2002). Reward, motivation, and reinforcement learning. Neuron *36*, this issue, 285–298.

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

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

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

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

Garris, P.A., Kilpatrick, M., Bunin, M.A., Michael, D., Walker, O.D., and Wightman, R.M. (1999). Dissociation of dopamine release in the nucleus accumbens from intracranial self-stimulation. Nature *398*, 67–69.

Hammer, M. (1993). An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. Nature *336*, 59–63.

Herrnstein, R.J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. J. Exp. Anal. Behav. *4*, 267–272.

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

Heyman, G.M. (2000). An economic approach to animal models of alcoholism. Alcohol Res. Health *24*, 132–139.

Hollerman, J.R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. *1*, 304–309.

Kahneman, D., and Tverksy, A. (1979). Prospect theory: an analysis of decision under risk. Econometrica *47*, 263–291.

Kilpatrick, M.R., Rooney, M.B., Michael, D.J., and Wightman, R.M. (2000). Extracellular dopamine dynamics in rat caudate-putamen during experimenter-delivered and intracranial self-stimulation. Neuroscience *96*, 697–706.

Knutson, B., Westdorp, A., Kaiser, E., and Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. Neuroimage *12*, 20–27.

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

Krimer, L.S., Muly, E.C., III, Williams, G.V., and Goldman-Rakic, P.S. (1998). Dopaminergic regulation of cerebral cortical microcirculation. Nat. Neurosci. *1*, 286–289.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. Quarterly Journal of Economics *112*, 443–478.

Lowenstein, G., and Elster, J. eds. (1992). Choice Over Time (New York: Russell Sage Foundation).

Lowenstein, G., and Prelec, D. (1992). Anomalies in intertemporal choice: evidence and an interpretation. Quarterly Journal of Economics *107*, 573–597.

Mazur, J.E. (1988). Estimation of indifference points with an adjusting-delay procedure. J. Exp. Anal. Behav. *49*, 37–47.

Merton, R. (1990). Continuous Time Finance (Cambridge: Blackwell).

Montague, P.R., and Sejnowksi, T.J. (1994). The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. Learn. Mem. *1*, 1–33.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1994). Foraging in an uncertain environment using predictive hebbian learning. In Advances in Neural Information Processing Systems, Volume 6, J.D. Dowan, G. Tesauro, and J. Alspector, eds. (San Francisco: Morgan Kaufmann), pp. 598–605.

Montague, P.R., Dayan, P., Person, C., and Sejnowski, T.J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. Nature *376*, 725–728.

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

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

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

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

Real, L.A. (1991). Animal choice behavior and the evolution of cognitive architecture. Science *253*, 980–986.

Rolls, E.T. (2000). The orbitofrontal cortex and reward. Cereb. Cortex *10*, 284–294.

Romo, R., and Schultz, W. (1990). Dopamine deurons of the monkey midbrain: contingencies of response to active touch during self-initiated arm movements. J. Neurophysiol. *63*, 592–606.

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

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron *36*, this issue, 241–263.

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

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

Schultz, W., Tremblay, L., and Hollerman, J.B. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. Cereb. Cortex *10*, 272–283.

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

Sutton, R., and Barto, A.G. (1998). Reinforcement Learning (Cambridge, MA: MIT Press).

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. Economic Letters *8*, 201–207.

Tremblay, L., and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. Nature *398*, 704–708.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. Nature *412*, 43–48.

Zald, D.H., Hagen, M.C., and Pardo, J.V. (2002). Neural correlates of tasting concentrated quinine and sugar solutions. J. Neurophysiol. *87*, 1068–1075.