Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis

Paul W. Glimcher[1]

Center for Neuroeconomics, New York University, New York, NY 10003

Edited by Donald W. Pfaff, The Rockefeller University, New York, NY, and approved December 9, 2010 (received for review October 14, 2010)

A number of recent advances have been achieved in the study of midbrain dopaminergic neurons. Understanding these advances and how they relate to one another requires a deep understanding of the computational models that serve as an explanatory framework and guide ongoing experimental inquiry. This intertwining of theory and experiment now suggests very clearly that the phasic activity of the midbrain dopamine neurons provides a global mechanism for synaptic modification. These synaptic modifications, in turn, provide the mechanistic underpinning for a specific class of reinforcement learning mechanisms that now seem to underlie much of human and animal behavior. This review describes both the critical empirical findings that are at the root of this conclusion and the fantastic theoretical advances from which this conclusion is drawn.

The theory and data available today indicate that the phasic activity of midbrain dopamine neurons encodes a reward prediction error used to guide learning throughout the frontal cortex and the basal ganglia. Activity in these dopaminergic neurons is now believed to signal that a subject's estimate of the value of current and future events is in error and indicate the magnitude of this error. This is a kind of combined signal that most scholars active in dopamine studies believe adjusts synaptic strengths in a quantitative manner until the subject's estimate of the value of current and future events is accurately encoded in the frontal cortex and basal ganglia. Although some confusion remains within the larger neuroscience community, very little data exist that are incompatible with this hypothesis. This review provides a brief overview of the explanatory synergy between behavioral, anatomical, physiological, and biophysical data that has been forged by recent computational advances. For a more detailed treatment of this hypothesis, refer to Niv and Montague (1) or Dayan and Abbot (2).

Features of Midbrain Dopamine Neurons

Three groups of dopamine secreting neurons send axons along long-distance trajectories that influence brain activity in many areas (3): the A8 and A10 groups of the ventral tegmental area (VTA) and the A9 group of the substantia nigra pars compacta (SNc). Two remarkable features of these neurons noted at the time of their discovery were their very large cell bodies and very long and complicated axonal arbors that include terminals specialized to release transmitter into the extracellular space, *en passant* synapses, through which dopamine achieves an extremely broad anatomical distribution (4). As Cajal (5) first pointed out, the length and complexity of axonal arbors are often tightly correlated with cell body size; large cell bodies are required to support large terminal fields, and dopaminergic cell bodies are about as large as they can be. The midbrain dopaminergic system, thus, achieves the largest possible distribution of its signal with the minimal possible number of neurons.

The A9 cluster connects to the caudate and putamen, and the A8 and A10 axons make contact with the ventral striatum and the fronto-cortical regions beyond (6, 7). There does, however, seem to be some intermixing of the three cell groups (8–10). Classic studies of these cells under conditions ranging from slice preparations to awake behaving primates, however, stress homogeneity in response patterns across these groups. Although knowing that one is actually recording from a dopamine neuron may be difficult in chronic studies (11), all cells that look like dopamine neurons in the core of the VTA and SNc seem to respond in the same way. Even the structure of the axons of these neurons supports the notion that activity is homogenous across this population of giant cells. Axons of adjacent neurons are electrically coupled to one another in this system (12, 13). Modeling studies suggest that this coupling makes it more difficult for individual neurons to fire alone, enforcing highly synchronous and thus, tightly correlated firing across the population (14).

A final note is that these neurons generate atypically long-duration action potentials, as long as 2–3 ms. This is relevant, because it places a very low limit on the maximal firing rates that these neurons can produce (15).

What emerges from these many studies is the idea that the dopamine neurons are structurally well-suited to serve as a specialized low-bandwidth channel for broadcasting the same information to large territories in the basal ganglia and frontal cortex. The large size of the cell bodies, the fact that the cells are electrically coupled, and the fact that they fire at low rates and distribute dopamine homogenously throughout a huge innervation territory—all these unusual things mean that they cannot say much to the rest of the brain but what they say must be widely heard. It should also be noted, however, that specializations at the site of release may well serve to filter this common message in ways that tailor it for different classes of recipients. Zhang et al. (16) have recently shown differences between the time courses of dopamine levels in the dorsal and ventral striata that likely reflect functional specializations for release and reuptake between these areas.

Dopaminergic Targets: Frontal Cortex and Basal Ganglia

It is also important to recognize that the dopamine neurons lie embedded in a large and well-described circuit. At the level of the cortex, the dopamine neurons send whatever signal they carry throughout territories anterior to the central sulcus and send little or no information to parietal, temporal, and occipital cortices (6). The outputs of the dopamine-innervated frontal cortices, however, also share another commonality; many of the major long-distance outputs of the frontal cortex pass in a topographic manner to the two main input nuclei of the basal ganglia complex, the caudate and the putamen (17). Both areas also receive dense innervation from the midbrain dopaminergic neurons.

Structurally, the caudate and putamen (and the ventral-most parts of the putamen, known as the ventral striatum) are largely a single nucleus separated during development by the incursion of the fibers of the corona radiata (2, 6, 18) that project principally to two output nuclei, the globus pallidus and the substantia nigra pars reticulata. These nuclei then, in turn, provide two basic outputs. The first and largest of these outputs returns information to the frontal cortex through a thalamic relay. Interestingly, this relay back to the cortex maintains a powerful topographic sorting (19, 20). The medial and posterior parts of the cortex that are concerned with planning skeletomuscular movements send their outputs to a specific subarea of the putamen, which sends signals back to this same area of the cortex through the globus pallidus

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Quantification of Behavior" held June 11–13, 2010, at the AAAS Building in Washington, DC. The complete program and audio files of most presentations are available on the NAS Web site at www.nasonline.org/quantification.

Author contributions: P.W.G. wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

[1]E-mail: glimcher@cns.nyu.edu.

and the ventrolateral thalamus. Together, these connections form a set of long feedback loops that seems to be serially interconnected (9) and ultimately, generates behavioral output through the skeletomuscular and eye-movement control pathways of the massive frontal-basal ganglia system.

The second principal class of output from the basal ganglia targets the midbrain dopamine neurons themselves and also forms a feedback loop. These outputs pass to the dendrites of the dopamine cells, where they combine with inputs from the brainstem that likely carry signals about rewards being currently consumed (21). In this way, the broadly distributed dopamine signals sent to the cortex and the basal ganglia likely reflect some combination of outputs from the cortex and places such as the tongue. The combined signals are then, of course, broadcast by the dopamine neurons throughout the basal ganglia and the frontal cortex.

Theory of Reinforcement Learning

From Pavlov to Rescorla and Wagner. Understanding the functional role of dopamine neurons, however, requires more than a knowledge of brain circuitry; it also requires an understanding of the classes of computational algorithms in which dopamine neurons seem to participate. Pavlov (22) observed, in his famous experiment on the salivating dog, that if one rings a bell and follows that bell with food, dogs become conditioned to salivate after the bell is rung. This process, where an unconditioned response comes to be elicited by a conditioned stimulus, is one of the core empirical regularities around which psychological theories of learning have been built. Pavlov (22) hypothesized that this behavioral regularity emerges because a preexisting anatomical connection between the sight of food and activation of the salivary glands comes to be connected to bell-detecting neurons by experience.

This very general idea was first mathematically formalized when Bush and Mosteller (23, 24) proposed that the probability of Pavlov's (22) dog expressing the salivary response on sequential trials could be computed through an iterative equation where (Eq. **1**)

$$A_{next\_trial} = A_{last\_trial} + \alpha(R_{current\_trial} - A_{last\_trial}). \quad [1]$$

In this equation, A_{next_trial} is the probability that the salivation will occur on the next trial (or more formally, the associative strength of the connection between the bell and salivation). To compute A_{next_trial}, one begins with the value of A on the previous trial and adds to it a correction based on the animal's experience during the most recent trial. This correction, or error term, is the difference between what the animal actually experienced (in this case, the reward of the meat powder expressed as $R_{current\_trial}$) and what he expected (simply, what A was on the previous trial). The difference between what was obtained and what was expected is multiplied by α, a number ranging from 0 to 1, which is known as the learning rate. When $\alpha = 1$, A is always immediately updated so that it equals R from the last trial. When $\alpha = 0.5$, only one-half of the error is corrected, and the value of A converges in half steps to R. When the value of α is small, around 0.1, then A is only very slowly incremented to the value of R.

What the Bush and Mosteller (23, 24) equation does is compute an average of previous rewards across previous trials. In this average, the most recent rewards have the greatest impact, whereas rewards far in the past have only a weak impact. If, to take a concrete example, $\alpha = 0.5$, then the equation takes the most recent reward, uses it to compute the error term, and multiplies that term by 0.5. One-half of the new value of A is, thus, constructed from this most recent observation. That means that the sum of all previous error terms (those from all trials in the past) has to count for the other one-half of the estimate. If one looks at that older one-half of the estimate, one-half of that one-half comes from what was observed one trial ago (thus, 0.25 of the total estimate) and one-half (0.25 of the estimate) comes from the sum of all trials before that one. The iterative equation reflects a weighted sum of previous rewards. When the learning rate (α) is 0.5, the weighting rule effectively being carried out is (Eq. **2**)

$$A_{now} = 0.5\ R_{now} + 0.25\ R_{t-1} + 0.125\ R_{t-2} + 0.0625\ R_{t-3} + \ldots, \quad [2]$$

an exponential series, the rate at which the weight declines being controlled by α.

When α is high, the exponential function declines rapidly and puts all of the weight on the most recent experiences of the animal. When α is low, it declines slowly and averages together many observations, which is shown in Fig. 1.

The Bush and Mosteller (23, 24) equation was critically important, because it was the first use of this kind of iterative error-based rule for reinforcement learning; additionally, it forms the basis of all modern approaches to this problem. This is a fact often obscured by what is known as the Rescorla–Wagner model of classical conditioning (25). The Rescorla–Wagner model was an important extension of the Bush and Mosteller approach (23, 24) to the study of what happens to associative strength when two cues predict the same event. Their findings were so influential that the basic Bush and Mosteller rule is now often mistakenly attributed to Rescorla and Wagner by neurobiologists.

Learning Value Instead of Associations. The next important point to make is that these early psychological theories were about associative strengths between classically conditioned stimuli and conditioned automatic responses. These models were about learning but not about concepts like value and choice that figure prominently in modern discussions of the function of dopamine. In the worlds of dynamic programming, computer science, and economics, however, these basic equations were easily extended to include a more explicit notion of value. Consider an animal trying to learn the value of pressing a lever that yields four pellets of food with a probability of 0.5. Returning to the Bush and Mosteller (23, 24) (or Rescorla–Wagner) (25) equation (Eq. **3**),

$$A_{next\_trial} = A_{last\_trial} + \alpha(Number\_of\_Pellets_{current\_trial} - A_{last\_trial}). \quad [3]$$

Because, in one-half of all trials, the animal is rewarded and in one-half, he is not and because all rewards have a value of four, we know exactly what this equation will do. If α has a value of one, A will bounce up and down between zero and four; if α is infinitely small, A will converge to two. That is striking, because two is the long-run average, or expected, value of pressing the lever. Therefore, in an environment that does not change, when α is small, this equation converges to the expected value of the action (Eq. **4**):

$$EV_{next\_trial} = EV_{last\_trial} + \alpha(Number\_of\_Pellets_{current\_trial} - EV_{last\_trial}). \quad [4]$$

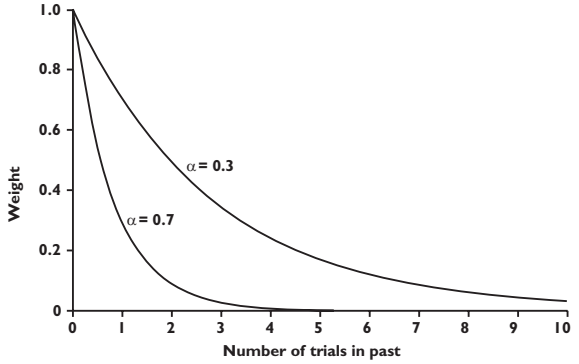

Fig. 1. "Weights determining the effects of previous rewards on current associative strength effectively decline as an exponential function of time" (65). [Reproduced with permission from Oxford University Press from ref. 65 (Copyright 2010, Paul W. Glimcher).]

Today, the Bush and Mosteller (23, 24) equation forms the core of how most people think about learning values. The equation provides a way for us to learn expected values. If we face a stable environment and have lots of time, we can even show that this equation is guaranteed to converge to expected value (26).

Sutton and Barto: Temporal Difference Model. The story of reinforcement learning described up to this point is a story largely from psychology and mostly focused on associative learning. That story changed abruptly in the 1990s when computer scientists Sutton and Barto (26) began to think seriously about these preexisting theories and noticed two key problems with them:

i) These theories all treated time as passing in unitary fixed epochs usually called trials. In Bush and Mosteller (23, 24), trials pass one after another, and updates to the values of actions occur only between trials. In the real world, time is more continuous. Different events in a trial might mean different things or might indicate different things about value.

ii) The second key problem was that these theories dealt in only a rudimentary way with how to link sequential cues (for example, a tone followed by a bell) with a later event of positive or negative value. The theories were good at learning that a tone or a lever predicted a reward but not so good at learning that a light that perfectly predicted the appearance of a lever meant that the later appearance of the lever told you nothing new.

To address these issues, Sutton and Barto (26) developed what has come to be known as temporal difference (TD) learning. That model has been presented in detail in elsewhere (26). Here, we review the most important advances that they achieved that are critical for understanding dopamine.

Sutton and Barto (26) began by arguing that, in essence, the Bush and Mosteller (23, 24) approach stated the problem that learning systems were trying to solve incorrectly. The Bush and Mosteller (23, 24) equation learns the values of previous events. Sutton and Barto (26) argued that the goal of a learning system should instead be to predict the value of future events. Of course, predictions have to be based on previous experience, and therefore, these two ideas are closely related; however, TD learning was designed with a clear goal in mind: predict the value of the future.

That is an important distinction, because it changes how one has to think about the reward prediction error at the heart of these reinforcement learning models. In Bush and Mosteller (23, 24) class models, reward prediction error is the difference between a weighted average of past rewards and the reward that has just been experienced. When those are the same, there is no error, and the system does not learn. Sutton and Barto (26), by contrast, argued that the reward prediction error term should be viewed as the difference between one's rational expectations of all future rewards and any information (be it an actual reward or a signal that a reward is coming up) that leads to a revision of expectations. If, for example, we predict that we will receive one reward every 1 min for the next 10 min and a visual cue indicates that, instead of these 10 rewards, we will receive one reward every 1 min for 11 min, then a prediction error exists when the visual cue arrives, not 11 min later when the final (and at that point, fully expected) reward actually arrives. This is a key difference between TD class and Bush and Mosteller (23, 24) class models.

To accomplish the goal of building a theory that both could deal with a more continuous notion of time and could build a rational (or near-rational) expectation of future rewards, Sutton and Barto (26) switched away from simple trial-based representations of time to a representation of time as a series of discrete moments extending from now into the infinite future. They then imagined learning as a process that occurred not just at the end of each trial but at each of these discrete moments.

To understand how they did this, consider a simple version of TD learning in which each trial can be thought of as made up of 20 moments. What the TD model is attempting to accomplish is to build a prediction about the rewards that can be expected in each of those 20 moments. The sum of those predictions is our total expectation of reward. We can represent this 20-moment expectation as a set of 20 learned values, one for each of the 20 moments. This is the first critical difference between TD class and Bush and Mosteller (23, 24) class models. The second difference lies in how these 20 predictions are generated. In TD, the prediction at each moment indicates not only the reward that is expected at that moment but also the sum of (discounted) rewards available in each of the subsequent moments.

To understand this critical point, consider the value estimate, V_1, that is attached to the first moment in the 20-moment-long trial. That variable needs to encode the value of any rewards expected at that moment, the value of any reward expected at the next moment decremented by the discount factor, the value of the next moment further decremented by the discount factor, and so on. Formally, that value function at time tick number one is (Eq. **5**)

$$V_1 = r_1 + \gamma^1 r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \gamma^4 r_{t+4} + \ldots + \gamma^{19} r_{t+19}, \quad \textbf{[5]}$$

where γ, the discount parameter, captures the fact that each of us prefers (derives more utility from) sooner rather than later rewards; the size of γ depends on the individual and the environmental context. Because this is a reinforcement learning system, it also automatically takes probability into account as it builds these estimates of r at each time tick. This means that the r values shown here are really expected rewards or average rewards observed at that time tick. Two kinds of events can, thus, lead to a positive prediction error: the receipt of an unexpected reward or the receipt of information that allows one to predict a later (and previously unexpected) reward.

To make this important feature clear, consider a situation in which an animal sits for 20 moments, and at any unpredictable moment, a reward might be delivered with a probability of 0.01. Whenever a reward is delivered, it is almost completely unpredictable, which leads to a large prediction error at the moment that the reward is delivered. This necessarily leads to an increment in the value of that moment. On subsequent trials, however, it is usually the case that no reward is received (because the probability is so low), and thus, on subsequent trials, the value of that moment is repeatedly decremented. If learning rates are low, the result of this process of increment and decrement is that the value of that moment will fluctuate close to zero, and we will observe a large reward prediction error signal after each unpredicted reward. Of course this is, under these conditions, true of all of the 20 moments in this imaginary trial.

Next, consider what happens when we present a tone at any of the first 10 moments that is followed 10 moments later by a reward. The first time that this happens, the tone conveys no information about future reward, no reward is expected, and therefore, we have no prediction error to drive learning. At the time of the reward, in contrast, a prediction error occurs that drives learning in that moment. The goal of TD, however, is to reach a point at which the reward delivered 10 moments after the tone is unsurprising. The goal of the system is to produce no prediction error when the reward is delivered. Why is the later reward unsurprising? It is unsurprising because of the tone. Therefore, the goal of TD is to shift the prediction error from the reward to the tone.

TD accomplishes this goal by attributing each obtained reward not just to the value function for the current moment in time but also to a few of the preceding moments in time (exactly how many is a free parameter of the model). In this way, gradually over time, the unexpected increment in value associated with the reward effectively propagates backward in time to the tone. It stops there simply because there is nothing before the tone that predicts the future reward. If there had been a light fixed before that tone in time, then the prediction would have propagated backward to that earlier light. In exactly this way, TD uses patterns of stimuli and experienced rewards to build an expectation about future rewards.

Theory and Physiology of Dopamine

With a basic knowledge of both the anatomy of dopamine and the theory of reinforcement learning, consider the following classic experiment by Schultz et al. (27). A thirsty monkey is seated before two levers. The monkey has been trained to perform a simple instructed choice task. After the illumination of a centrally located start cue, the monkey will receive an apple juice reward if he reaches out and presses the left but not the right lever. While the animal is performing this task repeatedly, Schultz et al. (27) record the activity of midbrain dopamine neurons. Interestingly, during the early phases of this process, the monkeys behave somewhat erratically, and the neurons are silent when the start cue is presented but respond strongly whenever the monkey receives a juice reward. As the monkey continues to perform the task, however, both the behavior and the activity of the neurons change systematically. The monkey comes to focus all of his lever pressing on the lever that yields a reward, and as this happens, the response of the neurons to the juice reward dies out. This is shown in Fig. 2.

At the same time, however, the neurons begin to respond whenever the start cue is illuminated. When Schultz et al. (27) first observed these responses, they hypothesized that "dopamine neurons are involved with transient changes of impulse activity in basic attentional and motivational processes underlying learning and cognitive behavior" (27).

Shortly after this report had been published, Montague et al. (28, 29) had begun to examine the activity of octopamine neurons in honey bees engaged in learning. They had hypothesized that the activity of these dopamine-related neurons in these insects encoded a reward prediction error of some kind (28, 29). When they became aware of the results of Schultz et al. (27), they realized that it was not simply the reward prediction error (RPE) defined by Bush and Mosteller (23, 24) class models, but it was exactly the RPE signal predicted by a TD class model. Recall that the TD model generates an RPE whenever the subject's expected reward changes. For a TD class model, this means that, after an unpredictable visual cue comes to predict a reward, it is the unexpected visual cue that tells you that the world is better than you expected. The key insight here is that the early burst of action potentials after the visual cue is what suggested to Montague et al. (28, 29) that Schultz et al. (27) were looking at a TD class system.

Subsequently, these two groups collaborated (26) to examine the activity of primate midbrain dopamine neurons during a conditioning task of exactly the kind that Pavlov (22) had originally studied. In that experiment, thirsty monkeys sat quietly under one of two conditions. In the first condition, the monkeys received, at unpredictable times, a squirt of water into their mouths. They found that, under these conditions, the neurons responded to the juice with a burst of action potentials immediately after any unpredicted water was delivered. In the second condition, the same monkey sat while a visual stimulus was delivered followed by a squirt of juice. The first time that this happened to the monkey, the neurons responded as before: they generated a burst of action potentials after the juice delivery but were silent after the preceding visual stimulus. With repetition, however, two things happened. First, the magnitude of the response to the water declined until, after dozens of trials, the water came to evoke no response in the neurons. Second and with exactly the same time course, the dopamine neurons began responding to the visual stimulus. As the response to the reward itself diminished, the response to the visual stimulus grew. What they had observed were two classes of responses, one to the reward and one to the tone, but both were responses predicted by the TD models that Montague et al. (28, 29) had been exploring.

Two Dopamine Responses and One Theory. This is a point about which there has been much confusion, and therefore, we pause for a moment to clarify this important issue. Many scientists who are familiar only with Bush and Mosteller (23, 24) class models (like the Rescorla–Wagner model) (25) have looked at these data (or others like them) and been struck by these two different responses—one at the reward delivery, which happens only early in the session, and a second at the visual cue, which happens only late in the session. The Bush and Mosteller (23, 24) algorithm predicts only the responses synchronized to the reward itself, and therefore, these scholars often conclude that dopamine neurons are doing two different things, only one of which is predicted by theory. If, however, one considers the TD class of models (which was defined more than a decade before these neurons were studied), then this statement is erroneous. The insight of Sutton and Barto (31) in the early 1980s was that reinforcement learning systems should use the reward prediction error signal to drive learning whenever something changes expectations about upcoming rewards. After a monkey has learned that a tone indicates a reward is forthcoming, then hearing the tone at an unexpected time is as much a positive reward prediction error as is an unexpected reward itself. The point here is that the early and late bursts observed in the Schultz et al. (27, 30) experiment described above are really the same thing in TD class models. This means that there is no need to posit that dopamine neurons are doing two things during these trials: they seem to be just encoding reward prediction errors in a way well-predicted by theory.

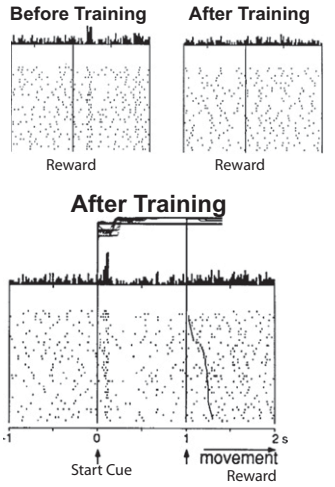

Fig. 2. "Raster plot of dopamine neuron activity. Upper panel shows response of dopamine neuron to reward before and after training. Lower panel shows response of dopamine neuron to start cue after training" (26). [Reproduced with permission from ref. 26 (Copyright 1993, Society for Neuroscience).]

Negative Reward Prediction Errors. In the same paper mentioned above, Schultz et al. (30) also examined what happens when an expected reward is omitted and the animal experiences a negative prediction error. To examine this, monkeys were first trained to anticipate a reward after a visual cue as described above and then, on rare trials, they simply omitted the water reward at the end of the trial. Under these conditions, Schultz et al. (30) found that the neurons responded to the omitted reward with a decrement in their firing rates from baseline levels (Fig. 3).

Montague et al. (28, 29) realized that this makes sense from the point of view of a TD class—and in this case, a Bush and Mosteller (23, 24) class—reward prediction error. In this case, an unexpected visual cue predicted a reward. The neurons produced a burst of action potentials in response to this prediction error. Then, the predicted reward was omitted. This yields a negative prediction error, and indeed, the neurons respond after the omitted reward with a decrease in firing rates. One interesting feature of this neuronal response, however, is that the neurons do not respond with much of a decrease. The presentation of an unexpected reward may increase firing rates to 20 or 30 Hz from their 3- to 5-Hz baseline. Omitting the same reward briefly decreases firing rates to 0 Hz, but this is a decrease of only 3–5 Hz in total rate.

If one were to assume that firing rates above and below baseline were linearly related to the reward prediction error in TD class models, then one would have to conclude that primates should be less influenced in their valuations by negative prediction errors than by positive prediction errors, but we know that primates are much more sensitive to losses below expectation than to gains above expectation (32–35). Thus, the finding of Schultz et al. (27, 30) that positive prediction errors shift dopamine firing rates more than negative prediction errors suggests

either that the relationship between this firing rate and actual learning is strongly nonlinear about the zero point or that dopamine codes positive and negative prediction errors in tandem with a second system specialized for the negative component. This latter possibility was first raised by Daw et al. (36), who specifically proposed that two systems might work together to encode prediction errors, one for coding positive errors and one for coding negative errors.

TD Models and Dopamine Firing Rates. The TD class models, however, predict much more than simply that some neurons must respond positively to positive prediction errors and negatively to negative prediction errors. These iterative computations also tell us about how these neurons must combine recent rewards in their reward prediction. Saying a system recursively estimates value by computing (Eq. **6**)

$$EV_{next\_trial} = EV_{last\_trial} + \alpha(Number\_of\_Pellets_{current\_trial} - EV_{last\_trial}) \quad [6]$$

is mathematically equivalent to saying that the computation of value averages recent rewards using an exponential weighting function of (Eq. **7**)

$$EV_{now} = \alpha^1 Pellets_{now} + \alpha^2 Pellets_{t-1} + \alpha^3 Pellets_{t-2} + \alpha^4 Pellets_{t-3} + \ldots, \quad [7]$$

where α, the learning rate, is a number between one and zero. If, for example, α has a value of 0.5, then (Eq. **8**)

$$EV_{now} = 0.5\ Pellets_{now} + 0.25\ Pellets_{t-1} + 0.125\ Pellets_{t-2} + 0.0625\ Pellets_{t-3} + \ldots. \quad [8]$$

If the dopamine neurons really do encode an RPE, they encode the difference between expected and obtained rewards. In a simple conditioning or choice task, that means that they encode something like (Eq. **9**)

$$RPE = R_{obtained} - [0.5\ Pellets_{now} + 0.25\ Pellets_{t-1} + 0.125\ Pellets_{t-2} + 0.0625\ Pellets_{t-3} + \ldots]. \quad [9]$$

The TD model presented by Sutton and Barto (26) tells us little about the value α should take under any specific set of conditions (here, it is arbitrarily set to 0.5), but we do know that the decay rate for the weights in the bracketed part of the equation above should decline exponentially for any stationary environment. We also know something else: when the prediction equals the obtained reward, then the prediction error should equal zero. That means that the actual value of $R_{obtained}$ should be exactly equal to the sum of the exponentially declining weights in the bracketed part of the equation.

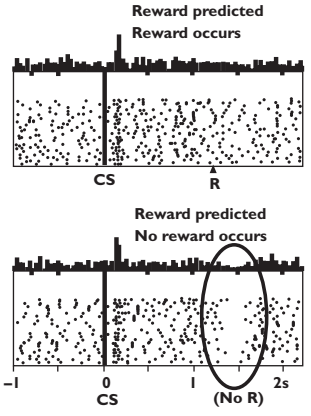

Fig. 3. "When a reward is cued and delivered, dopamine neurons respond only to the cue. When an expected reward is omitted after a cue the neuron responds with a suppression of activity as indicated by the oval" (29). [Reproduced with permission from ref. 29 (Copyright 1997, American Association for the Advancement of Science).]

Bayer and Glimcher (37) tested these predictions by recording from dopamine neurons while monkeys engaged in a learning and choice task. In their experiment, monkeys had to precisely time when in a trial they would make a response for a reward. One particular response time would yield the most reward but that best time shifted unexpectedly (with a roughly flat hazard function) across large blocks of trials. On each trial, the monkey could cumulate information from previous trials to make a reward prediction. Then, the monkey made his movement and received his reward. The difference between these two should have been the reward prediction error and thus, should be correlated with dopamine firing rates.

To test that prediction, Bayer and Glimcher (37) performed a linear regression between the history of rewards given to the monkey and the firing rates of dopamine neurons. The linear regression determines the weighting function that combines information about these previous rewards in a way that best predicts dopamine firing rates. If dopamine neurons are an iteratively computed reward prediction error system, then increasing reward on the current trial should increase firing rates. Increasing rewards on trials before that should decrease firing rates and should do so with an exponentially declining weight. Finally, the regression should indicate that the sum of old weights should be equal (and opposite in sign) to the weight attached to the current reward. In fact, this is exactly what Bayer and Glimcher (37) found (Fig. 4).

The dopamine firing rates could be well-described as computing an exponentially weighted sum of previous rewards and subtracting from that value the magnitude of the most recent reward. Furthermore, they found, as predicted, that the integral of the declining exponential weights was equal to the weight attributed to the most recent reward. It is important to note that this was not required by the regression in any way. Any possible weighting function could have come out of this analysis, but the observed weighting function was exactly that predicted by the TD model.

A second observation that Bayer and Glimcher (37) made, however, was that the weighting functions for positive and negative prediction errors (as opposed to rewards) were quite different. Comparatively speaking, the dopamine neurons seemed fairly insensitive to negative prediction errors. Although Bayer et al. (15) later showed that, with a sufficiently complex nonlinearity, it was possible to extract positive and negative reward prediction errors from dopamine firing rates, their data raise again the possibility that negative prediction errors might well be coded in tandem with another unidentified system.

Dopamine Neurons and Probability of Reward. Following on these observations, Schultz et al. (27, 30) observed yet another interesting feature of the dopamine neurons well-described by the TD model. In a widely read paper, Fiorillo et al. (38) showed that dopamine neurons in classical conditioning tasks seem to show a ramp of activity between cue and reward whenever the rewards are delivered probabilistically, as shown in Fig. 5.

Recall that TD class models essentially propagate responsibility for rewards backward in time. This is how responses to unexpected rewards move through time and attach to earlier stimuli that predict those later rewards. Of course, the theory predicts that both negative and positive prediction errors should propagate backward in time in the same way.

Now, with that in mind, consider what happens when a monkey sees a visual cue and receives a 1-mL water reward with a probability of 0.5 1 s after the tone. The average value of the tone is, thus, 0.5 mL. In one-half of all trials, the monkey gets a reward (a positive prediction error of 0.5), and in one-half of all trials, it gets does not get a reward (a negative prediction error of 0.5). One would imagine that these two signals would work their way backward in trial time to the visual cue. Averaging across many trials, one would expect to see these two propagating signals cancel out each other. However, what would happen if the dopamine neurons responded more strongly to positive than negative prediction errors (37)? Under that set of conditions, the TD class models would predict that average dopaminergic activity would show the much larger positive pre-

diction error propagating backward in time as a ramp—exactly what Schultz et al. (27, 30) observed.

This observation of the ramp has been quite controversial and has led to a lot of confusion. Schultz et al. (27, 30) said two things about the ramp: that the magnitude and shape of the ramp carried information about the history of previous rewards and that this was a feature suggesting that the neurons encoded uncertainty in a way not predicted by theory. The first of these observations is unarguably true. The second is true only if we assume that positive and negative prediction errors are coded as precise mirror images of one another. If instead, as the Bayer and Glimcher (37) data indicate, negative and positive prediction errors are encoded differentially in the dopamine neurons, then the ramp is not only predicted by existing theory, it is required. This is a point first made in print by Niv et al. (39).

Axiomatic Approaches. How sure are we that dopamine neurons encode a reward prediction error? It is certainly the case that the average firing rates of dopamine neurons under a variety of conditions conform to the predictions of the TD model, but just as the TD class succeeded the Bush and Mosteller (23, 24) class, we have every reason to believe that future models will improve on the predictions of TD. Therefore, can there ever be a way to say conclusively that the activity of dopamine neurons meets some absolute criteria of necessity and sufficiency with regard to reinforcement learning? To begin to answer that question, Caplin and Dean (40) used a standard set of economic tools for the study of dopamine. Caplin and Dean (40) asked whether there was a compact, testable, mathematically axiomatic way to state the current dopamine hypothesis.

After careful study, Caplin and Dean (40) were able to show that the entire class of reward prediction error-based models could be reduced to three compact and testable mathematical statements called axioms—common mathematical features that all reward prediction error-based models must include irrespective of their specific features.

i) Consistent prize ordering. When the probabilities of obtaining specific rewards are fixed and the magnitudes of those rewards are varied, the ordering of obtained reward outcomes by neural activity (e.g., which reward produces more activity, regardless of how much more) must be the same regardless of the environmental conditions under which the rewards were received.

ii) Consistent lottery ordering. When rewards are fixed and the probabilities of obtaining specific rewards are varied, the ordering of rewards by neural activity (e.g., which reward outcome produces more activity) should be the same for all of the reward outcomes that can occur under a given set of probabilities.

iii) No surprise equivalence. The final criterion of necessity and sufficiency identified by Caplin and Dean (41) was that RPE signals must respond identically to all fully predicted outcomes (whether good or bad), conditions under which the reward prediction error is zero.

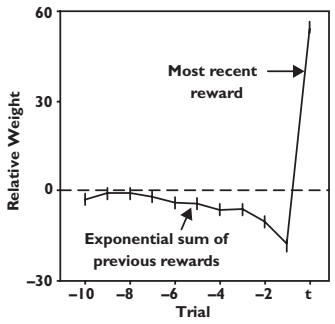

Fig. 4. "The linear weighting function which best relates dopamine activity to reward history" (65). [Reproduced with permission from Oxford University Press from ref. 65 (Copyright 2011, Paul W. Glimcher).]

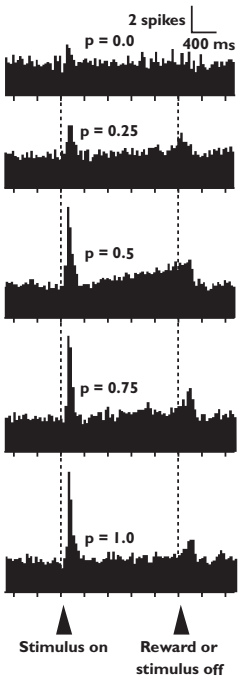

Fig. 5. "Peri-stimulus time histogram of dopamine neuron activity during a cued and probabilistically rewarded task" (37). [Reproduced with permission from ref. 37 (Copyright 2003, American Association for the Advancement of Science).]

Caplin and Dean (40, 41) showed that any RPE system, whether a Bush and Mosteller (23, 24) class or TD class model, must meet these three axiomatic criteria. Saying that an observed system violated one or more of these axioms, they showed, was the same as saying that it could not, in principle, serve as a reward prediction error system. Conversely, they showed that, for any system that obeyed these three rules, neuronal activity could without a doubt be accurately described using at least one member of the reward prediction error model class. Thus, what was important about the axiomatization of the class of all RPE models by Caplin and Dean (40, 41) is that it provided a clear way to test this entire class of hypotheses.

In a subsequent experiment, Caplin et al. (42) then performed an empirical test of the axioms on brain activations (measured with functional MRI) in areas receiving strong dopaminergic inputs by constructing a set of monetary lotteries and having human subjects play those lotteries for real money. In those experiments, subjects either won or lost $5 on each trial, and the probabilities of winning or losing were systematically manipulated. The axioms indicate that for a reward prediction error encoding system under these conditions, three things will occur.

i) Winning $5 must always give rise to more activity than losing $5, regardless of the probability (from consistent prize ordering).

ii) The more certain you are that you will win, the lower must be the neural activation to winning, and conversely, the more certain you are that you will lose, the higher must be the activity to losing (from consistent lottery ordering).

iii) If you are certain of an outcome, whether it be winning or losing, neural activity should be the same, regardless of whether you win or lose $5 (from no surprise equivalence).

What they found was that activations in the insula violated the first two axioms of the reward prediction error theory. This was an unambiguous indication that the blood oxygen level-dependent (BOLD) activity in the insula could not, in principle, serve as an RPE signal for learning under the conditions that they studied. In contrast, activity in the ventral striatum obeyed all three axioms and thus, met the criteria of both necessity and sufficiency for serving as an RPE system. Finally, activity in the medial prefrontal cortex and the amygdala yielded an intermediate result. Activations in these areas seemed to weakly violate one of the axioms, raising the pos-

sibility that future theories of these areas would have to consider the possibility that RPEs either were not present or were only a part of the activation pattern here.

The paper by Caplin et al. (42) was important, because it was, in a real sense, the final proof that some areas activated by dopamine, the ventral striatum in particular, can serve as a reward prediction error encoder of the type postulated by TD models. The argument that this activation only looks like an RPE signal can now be entirely dismissed. The pattern of activity that the ventral striatum shows is both necessary and sufficient for use in an RPE system. That does not mean that it has to be such a system, but it draws us closer and closer to that conclusion.

Cellular Mechanisms of Reinforcement Learning

In the 1940s and 1950s, Hebb (43) was among the first to propose that alterations of synaptic strength based local patterns of activation might serve to explain how conditioned reflexes operated at the biophysical level. Bliss and Lomo (44) succeeded in relating these two sets of concepts when they showed long-term potentiation (LTP) in the rabbit hippocampus. Subsequent biophysical studies have shown several other mechanisms for altering synaptic strength that are closely related to both the theoretical proposal of Hebb (43) and the biophysical mechanism of Bliss and Lomo (44). Wickens (45) and Wickens and Kotter (46) proposed the most relevant of these for our discussion, which is often known as the three-factor rule. What Wickens (45) and Wickens and Kotter (46) proposed was that synapses would be strengthened whenever presynaptic and postsynaptic activities co-occurred with dopamine, and these same synapses would be weakened when presynaptic and postsynaptic activities occurred in the absence of dopamine. Indeed, there is now growing understanding at the biophysical level of the many steps by which dopamine can alter synaptic strengths (47).

Why is this important for models of reinforcement learning? An animal experiences a large positive reward prediction error: he just earned an unexpected reward. The TD model tells us that, under these conditions, we want to increment the value attributed to all actions or sensations that have just occurred. Under these conditions, we know that the dopamine neurons release dopamine throughout the frontocortical–basal ganglia loops and do so in a highly homogenous manner. The three-factor rule implies that any dopamine receptor-equipped neuron, active because it just participated in, for example, a movement to a lever, will have its active synapses strengthened. Thus, whenever a positive prediction error occurs and dopamine is released throughout the frontal cortices and the basal ganglia, any segment of the frontocortical–basal ganglia loop that is already active will have its synapses strengthened.

To see how this would play out in behavior, consider that neurons of the dorsal striatum form maps of all possible movements into the extrapersonal space. Each time that we make one of those movements, the neurons associated with that movement are active for a brief period and that activity persists after the movement is complete (48, 49). If any movement is followed by a positive prediction error, then the entire topographic map is transiently bathed in the global prediction error signal carried by dopamine into this area. What would this combination of events produce? It would produce a permanent increment in synaptic strength only among those neurons associated with recently produced movements. What would that synapse come to encode after repeated exposure to dopamine? It would come to encode the expected value (or perhaps, more precisely, the expected subjective value) of the movement.

What is critical to understand here is that essentially everything in this story is a preexisting observation of properties of the nervous system. We know that neurons in the striatum are active after movements as required of (the eligibility traces of) TD models. We know that a blanket dopaminergic prediction error is broadcast throughout the frontocortical–basal ganglia loops. We know that dopamine produces LTP-like phenomena in these areas when correlated with underlying activity. In fact, we even know that, after conditioning, synaptically driven action potential rates in these areas encode the subjective values of actions (48–51). Therefore, all of these biophysical components exist, and they exist in a configuration that could implement TD class models of learning.

We even can begin to see how the prediction error signal coded by the dopamine neurons could be produced. We know that neurons in the striatum encode, in their firing rates, the learned values of actions. We know that these neurons send outputs to the dopaminergic nuclei—a reward prediction. We also know that the dopaminergic neurons receive fairly direct inputs from sensory areas that can detect and encode the magnitudes of consumed rewards. The properties of sugar solutions encoded by the tongue, for example, have an almost direct pathway through which these signals can reach the dopaminergic nuclei. Given that this is true, constructing a prediction error signal at the dopamine neurons simply requires that excitatory and inhibitory synapses take the difference between predicted and experienced reward in the voltage of the dopamine neurons themselves or their immediate antecedents.

Summary and Conclusion

The basic outlines of the dopamine reward prediction error model seem remarkably well-aligned with both biological level and behavioral data; a wide range of behavioral and physiological phenomena seem well-described in a parsimonious way by this hypothesis. The goal of this presentation has been to communicate the key features of that alignment, which has been mediated by rigorous computational theory. It is important to note, however, that many observations do exist that present key challenges to the existing dopamine reward prediction error model. Most of these challenges are reviewed in Dayan and Niv (52).* It is also true that the reward prediction error hypothesis has focused almost entirely on the phasic responses of the dopamine neurons. It is unarguably true that the tonic activity of these neurons is also an important clinical and physiological feature (55) that is only just beginning to receive computational attention (56, 57).

One more recent challenge that deserves special mention arises from the work of Matsumoto and Hikosaka (58), who have recently documented the existence of neurons in the ventro-lateral portion of the SNc that clearly do not encode a reward prediction error. They hypothesize that these neurons form a second physiologically distinct population of dopamine neurons that plays some alternative functional role. Although it has not yet been established that these neurons do use dopamine as their neurotransmitter (which can be difficult) (11), this observation might suggest the existence of a second group of dopamine neurons whose activity lies outside the boundaries of current theory.

In a similar way, Ungless et al. (59) have shown that, in anesthetized rodents, some dopamine neurons in the VTA respond positively to aversive stimuli. Of course, for an animal that predicts a very aversive event, the occurrence of an only mildly aversive event would be a positive prediction error. Although it is hard to know what predictions the nervous system of an anesthetized rat might make, the observation that some dopamine neurons respond to aversive stimuli poses another important challenge to existing theory that requires further investigation.

Despite these challenges, the dopamine reward prediction error has proven remarkably robust. Caplin et al. (42) have shown axiomatically that dopamine-related signals in the ventral striatum can, by definition, be described accurately with models of this class. Montague et al. (29) have shown that the broad features of dopamine activity are well-described by TD class (26) models. More detailed analyses like those by Bayer and Glimcher (37) have shown quantitative agreement between dopamine firing rates and key structural features of the model. Work in humans (60, 61) has shown that activity in dopaminergic target areas is also well-accounted for by the general features of the model in this species. Similar work in rats also reveals the exis-

*It is important to acknowledge that there are alternative views of the function of these neurons. Berridge (53) has argued that dopamine neurons play a role closely related to the one described here that is referred to as incentive salience. Redgrave and Gurney (54) have argued that dopamine plays a central role in processes related to attention.

tence of a reward prediction error-like signal in midbrain dopamine neurons of that species (62). Additionally, it is also true that many of the components of larger reinforcement learning circuits in which the dopamine neurons are believed to be embedded have also now been identified (48–51, 63–65). Although it is always true that existing scientific models turn out to be incorrect at some point in the future with new data, there can be little doubt that the quantitative and computational study of dopamine neurons is a significant accomplishment of contemporary integrative neuroscience.

1. Niv Y, Montague PR (2008) Theoretical and empirical studies of learning. *Neuroeconomics: Decision Making and the Brain*, eds Glimcher PW, et al. (Academic Press, New York), pp 329–249.
2. Dayan P, Abbot LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge, MA).
3. Carlsson A (2000) *A Half-Century of Neurotransmitter Research: Impact on Neurology and Psychiatry. Nobel Prize Lecture Presented at Karolinska Institutet, Stockholm.* Available at http://nobelprize.org/nobel_prizes/medicine/laureates/2000/carlsson-lecture.html. Accessed November 2008.
4. Fuxe K, et al. (2010) The discovery of central monoamine neurons gave volume transmission to the wired brain. *Prog Neurobiol* 90:82–100.
5. Cajal SR (1909–1911) Histologie du système nerveux de l'homme & des vertébrés . . . Édition français revue & mise à jour par l'auteur, Translated by L. Azoulay. Paris, E. Arrault et Cie for A. Maloine.
6. Dahlström AB, Fuxe K (1964) Evidence for the existence of monoamine-containing neurons in the central nervous system. I: Demonstration of monoamines in the cell bodies of brain stem neurons. *Acta Physiol Scand* 62:1–55.
7. Lindvall O, Björklund A, Moore RY, Stenevi U (1974) Mesencephalic dopamine neurons projecting to neocortex. *Brain Res* 81:325–331.
8. Fallon JH (1988) Topographic organization of ascending dopaminergic projections. *Ann N Y Acad Sci* 537:1–9.
9. Haber SN, Fudge JL, McFarland NR (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J Neurosci* 20: 2369–2382.
10. Williams SM, Goldman-Rakic PS (1998) Widespread origin of the primate mesofrontal dopamine system. *Cereb Cortex* 8:321–345.
11. Margolis EB, Lock H, Hjelmstad GO, Fields HL (2006) The ventral tegmental area revisited: Is there an electrophysiological marker for dopaminergic neurons? *J Physiol* 577:907–924.
12. Grace AA, Bunney BS (1983) Intracellular and extracellular electrophysiology of nigral dopaminergic neurons—1. Identification and characterization. *Neuroscience* 10: 301–315.
13. Vandecasteele M, Glowinski J, Venance L (2005) Electrical synapses between dopaminergic neurons of the substantia nigra pars compacta. *J Neurosci* 25: 291–298.
14. Komendantov AO, Canavier CC (2002) Electrical coupling between model midbrain dopamine neurons: Effects on firing pattern and synchrony. *J Neurophysiol* 87: 1526–1541.
15. Bayer HM, Lau B, Glimcher PW (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. *J Neurophysiol* 98:1428–1439.
16. Zhang L, Doyon WM, Clark JJ, Phillips PE, Dani JA (2009) Controls of tonic and phasic dopamine transmission in the dorsal and ventral striatum. *Mol Pharmacol* 76:396–404.
17. Middleton FA, Strick PL (2002) Basal-ganglia 'projections' to the prefrontal cortex of the primate. *Cereb Cortex* 12:926–935.
18. Holt DJ, Graybiel AM, Saper CB (1997) Neurochemical architecture of the human striatum. *J Comp Neurol* 384:1–25.
19. Kelly RM, Strick PL (2003) Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *J Neurosci* 23:8432–8444.
20. Kelly RM, Strick PL (2004) Macro-architecture of basal ganglia loops with the cerebral cortex: Use of rabies virus to reveal multisynaptic circuits. *Prog Brain Res* 143:449–459.
21. DeLong MR, Wichmann T (2007) Circuits and circuit disorders of the basal ganglia. *Arch Neurol* 64:20–24.
22. Pavlov IP (1927) *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Dover, New York).
23. Bush RR, Mosteller F (1951) A mathematical model for simple learning. *Psychol Rev* 58: 313–323.
24. Bush RR, Mosteller F (1951) A model for stimulus generalization and discrimination. *Psychol Rev* 58:413–423.
25. Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, eds Black AH, Prokasy WF (Appleton Century Crofts, New York), pp 64–99.
26. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
27. Schultz W, Apicella P, Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J Neurosci* 13:900–913.
28. Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive hebbian learning. *Nature* 377:725–728.
29. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936–1947.
30. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
31. Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol Rev* 88:135–170.
32. Chen MK, Lakshminarayanan V, Santos LR (2006) How basic are behavioral biases? Evidence from capuchin monkey trading behavior. *J Polit Econ* 114:517–537.
33. Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211:453–458.
34. Tversky A, Kahneman D (1986) Rational choice and the framing of decisions. *J Bus* 59: S251–S278.
35. Santos LR, Lakshminarayanan V (2008) Innate constraints on judgment and decision-making?: Insights from children and non-human primates. *The Innate Mind: Foundations and the Future*, eds Carruthers P, Laurence S, Stich S (Oxford University Press, Oxford), pp 293–310.
36. Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. *Neural Netw* 15:603–616.
37. Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129–141.
38. Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299:1898–1902.
39. Niv Y, Duff MO, Dayan P (2005) Dopamine, uncertainty and TD learning. *Behav Brain Funct* 1:6.
40. Caplin A, Dean M (2007) The neuroeconomic theory of learning. *Am Econ Rev* 97: 148–152.
41. Caplin A, Dean M (2008) Axiomatic neuroeconomics. *Neuroeconomics: Decision Making and the Brain*, eds Glimcher PW, Camerer CF, Fehr E, Poldrack RA (Academic, London), pp 21–31.
42. Caplin A, Dean M, Glimcher PW, Rutledge RB (2010) Measuring beliefs and rewards: A neuroeconomic approach. *Q J Econ* 125:923–960.
43. Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York).
44. Bliss TV, Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J Physiol* 232:331–356.
45. Wickens JR (1993) *A Theory of the Striatum* (Pergamon, Oxford), 1st Ed.
46. Wickens JR, Kotter R (1995) Cellular models of reinforcement. *Models of Information Processing in Basal Ganglia*, eds Houk JC, Davis JL, Beiser DG (MIT Press, Cambridge, MA), pp 187–214.
47. Surmeier DJ, Plotkin J, Shen W (2009) Dopamine and synaptic plasticity in dorsal striatal circuits controlling action selection. *Curr Opin Neurobiol* 19:621–628.
48. Lau B, Glimcher PW (2007) Action and outcome encoding in the primate caudate nucleus. *J Neurosci* 27:14502–14514.
49. Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. *Neuron* 58:451–463.
50. Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* 84:555–579.
51. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. *Science* 310:1337–1340.
52. Dayan P, Niv Y (2008) Reinforcement learning: The good, the bad and the ugly. *Curr Opin Neurobiol* 18:185–196.
53. Berridge KC (2007) The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology (Berl)* 191:391–431.
54. Redgrave P, Gurney K (2006) The short-latency dopamine signal: A role in discovering novel actions? *Nat Rev Neurosci* 7:967–975.
55. Grace AA (1991) Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience* 41:1–24.
56. Niv Y, Daw ND, Joel D, Dayan P (2007) Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology (Berl)* 191:507–520.
57. Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010) Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. *Neuron* 67:144–155.
58. Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* 459:837–841.
59. Ungless MA, Argilli E, Bonci A (2010) Effects of stress and aversion on dopamine neurons: Implications for addiction. *Neurosci Biobehav Rev* 35:151–156.
60. McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38:339–346.
61. O'Doherty JP, Dayan P, Friston KJ, Critchley HD, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38:329–337.
62. Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci* 10:1615–1624.
63. Tremblay L, Schultz W (2000) Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *J Neurophysiol* 83:1877–1885.
64. Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441:223–226.
65. Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK (2009) A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat Rev Neurosci* 10: 885–892.
66. Paul G (2010) *Foundations of Neuroeconomic Analysis* (Oxford University Press, London).