48. H. S. Mayberg *et al., Ann. Neurol.* **28**, 57 (1990).
49. R. M. Cohen *et al.*, *Neuropsychopharmacology* **2**, 241 (1989).
50. J. E. LeDoux, *Sci. Am.* **6**, 50 (June 1994); M. Davis, *Annu. Rev. Neurosci.* **15**, 353 (1992).
51. J. E. LeDoux, *Curr. Opin. Neurobiol.* **2**, 191 (1992); L. M. Romanski and J. E. LeDoux, *J. Neurosci.* **12**, 4501 (1992); J. L. Armony, J. D. Cohen, D. Servan-Schreiber, J. E. LeDoux, *Behav. Neurosci.* **109**, 246 (1995).
52. K. P. Corodimas and J. E. LeDoux, *Behav. Neurosci.* **109**, 613 (1995).
53. M. J. D. Miserendino, C. B. Sananes, K. R. Melia, M. Davis, *Nature* **345**, 716 (1990); C. Farb *et al.*, *Brain Res.* **593**, 145 (1992); M. Davis, D. Rainne, M. Cassell, *Trends Neurosci.* **17**, 208 (1994).
54. M. E. P. Seligman, *J. Abnorm. Psychol.* **74**, 1 (1976); F. Schneider *et al., Am. J. Psychiatry* **153**, 206 (1996).
55. W. C. Drevets *et al.*, *J. Neuroscience* **12**, 3628 (1992).
56. Ssupported in part by National Instituite of Mental Health grants MH31593, MH40856, and MH-CRC43271; by a Research Scientist Award, MH00625; and by an Established Investigator Award from the National Association for Research in Schizophrenia and Affective Disorders.

# A Neural Substrate of Prediction and Reward

Wolfram Schultz, Peter Dayan, P. Read Montague*

**The capacity to predict future events permits a creature to detect, model, and manipulate the causal structure of its interactions with its environment. Behavioral experiments suggest that learning is driven by changes in the expectations about future salient events such as rewards and punishments. Physiological work has recently complemented these studies by identifying dopaminergic neurons in the primate whose fluctuating output apparently signals changes or errors in the predictions of future salient and rewarding events. Taken together, these findings can be understood through quantitative theories of adaptive optimizing control.**

An adaptive organism must be able to predict future events such as the presence of mates, food, and danger. For any creature, the features of its niche strongly constrain the time scales for prediction that are likely to be useful for its survival. Predictions give an animal time to prepare behavioral reactions and can be used to improve the choices an animal makes in the future. This anticipatory capacity is crucial for deciding between alternative courses of action because some choices may lead to food whereas others may result in injury or loss of resources.

Experiments show that animals can predict many different aspects of their environments, including complex properties such as the spatial locations and physical characteristics of stimuli (*1*). One simple, yet useful prediction that animals make is the probable time and magnitude of future rewarding events. "Reward" is an operational concept for describing the positive value that a creature ascribes to an object, a behavioral act, or an internal physical state. The function of reward can be described according to the behavior elicited (*2*). For example, appetitive or rewarding stimuli induce approach behavior that permits an animal to consume. Rewards may also play the role of positive reinforcers where they increase the frequency of behavioral reactions during learning and maintain well-established appetitive behaviors after learning. The reward value associated with a stimulus is not a static, intrinsic property of the stimulus. Animals can assign different appetitive values to a stimulus as a function of their internal states at the time the stimulus is encountered and as a function of their experience with the stimulus.

One clear connection between reward and prediction derives from a wide variety of conditioning experiments (*1*). In these experiments, arbitrary stimuli with no intrinsic reward value will function as rewarding stimuli after being repeatedly associated in time with rewarding objects—these objects are one form of unconditioned stimulus (US). After such associations develop, the neutral stimuli are called conditioned stimuli (CS). In the descriptions that follow, we call the appetitive CS the sensory cue and the US the reward. It should be kept in mind, however, that learning that depends on CS-US pairing takes many different forms and is not always dependent on reward (for example, learning associated with aversive stimuli). In standard conditioning paradigms, the sensory cue must consistently precede the reward in order for an association to develop. After conditioning, the animal's behavior indicates that the sensory cue induces a prediction about the likely time and magnitude of the reward and tends to elicit approach behavior. It appears that this form of learning is associated with a transfer of an appetitive or approach-eliciting component of the reward back to the sensory cue.

Some theories of reward-dependent learning suggest that learning is driven by the unpredictability of the reward by the sensory cue (*3*, *4*). One of the main ideas is that no further learning takes place when the reward is entirely predicted by a sensory cue (or cues). For example, if presentation of a light is consistently followed by food, a rat will learn that the light predicts the future arrival of food. If, after such training, the light is paired with a sound and this pair is consistently followed by food, then something unusual happens—the rat's behavior indicates that the light continues to predict food, but the sound predicts nothing. This phenomenon is called "blocking." The prediction-based explanation is that the light fully predicts the food that arrives and the presence of the sound adds no new predictive (useful) information; therefore, no association developed to the sound (*5*). It appears therefore that learning is driven by deviations or "errors" between the predicted time and amount of rewards and their actual experienced times and magnitudes [but see (*4*)].

Engineered systems that are designed to optimize their actions in complex environments face the same challenges as animals, except that the equivalent of rewards and punishments are determined by design goals. One established method by which artificial systems can learn to predict is called the temporal difference (TD) algorithm (*6*). This algorithm was originally inspired by behavioral data on how animals actually learn predictions (*7*). Real-world applications of TD models abound. The predictions learned by TD methods can also be used to implement a technique called dynamic programming, which specifies how a system can come to choose appropriate actions. In this article, we review how these computational methods provide an interpretation of the activity of dopamine neurons thought to mediate reward-processing and reward-dependent learning. The connection between the computational theory and the experimental results is striking and provides a quantitative framework for future experiments and theories on the computational roles of ascending monoaminergic systems (*8*–*13*).

W. Schultz is at the Institute of Physiology, University of Fribourg, CH-1700 Fribourg, Switzerland. E-mail: Wolfram.Schultz@unifr.ch P. Dayan is in the Department of Brain and Cognitive Sciences, Center for Biological and Computational Learning, E-25 MIT, Cambridge, MA 02139, USA. E-mail: dayan@ai.mit.edu P. R. Montague is in the Division of Neuroscience, Center for Theoretical Neuroscience, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA. E-mail: read@bcm.tmc.edu

*To whom correspondence should be addressed.

## Information Encoded in Dopaminergic Activity

Dopamine neurons of the ventral tegmental area (VTA) and substantia nigra have long been identified with the processing of rewarding stimuli. These neurons send their axons to brain structures involved in motivation and goal-directed behavior, for example, the striatum, nucleus accumbens, and frontal cortex. Multiple lines of evidence support the idea that these neurons construct and distribute information about rewarding events.

First, drugs like amphetamine and cocaine exert their addictive actions in part by prolonging the influence of dopamine on target neurons (*14*). Second, neural pathways associated with dopamine neurons are among the best targets for electrical self-stimulation. In these experiments, rats press bars to excite neurons at the site of an implanted electrode (*15*). The rats often choose these apparently rewarding stimuli over food and sex. Third, animals treated with dopamine receptor blockers learn less rapidly to press a bar for a reward pellet (*16*). All the above results generally implicate midbrain dopaminergic activity in reward-dependent learning. More precise information about the role played by midbrain dopaminergic activity derives from experiments in which activity of single dopamine neurons is recorded in alert monkeys while they perform behavioral acts and receive rewards.

In these latter experiments (*17*), dopamine neurons respond with short, phasic activations when monkeys are presented with various appetitive stimuli. For example, dopamine neurons are activated when animals touch a small morsel of apple or receive a small quantity of fruit juice to the mouth as liquid reward (Fig. 1). These phasic activations do not, however, discriminate between these different types of rewarding stimuli. Aversive stimuli like air puffs to the hand or drops of saline to the mouth do not cause these same transient activations. Dopamine neurons are also activated by novel stimuli that elicit orienting reactions; however, for most stimuli, this activation lasts for only a few presentations. The responses of these neurons are relatively homogeneous—different neurons respond in the same manner and different appetitive stimuli elicit similar neuronal responses. All responses occur in the majority of dopamine neurons (55 to 80%).

Surprisingly, after repeated pairings of visual and auditory cues followed by reward, dopamine neurons change the time of their phasic activation from just after the time of reward delivery to the time of cue onset. In one task, a naïve monkey is required to touch a lever after the appearance of a small light. Before training and in the initial phases of training, most dopamine neurons show a short burst of impulses after reward delivery (Fig. 1, top). After several days of training, the animal learns to reach for the lever as soon as the light is illuminated, and this behavioral change correlates with two remarkable changes in the dopamine neuron output: (i) the primary reward no longer elicits a phasic response; and (ii) the onset of the (predictive) light now causes a phasic activation in dopamine cell output (Fig. 1, middle). The changes in dopaminergic activity strongly resemble the transfer of an animal's appetitive behavioral reaction from the US to the CS.

In trials where the reward is not delivered at the appropriate time after the onset of the light, dopamine neurons are depressed markedly below their basal firing rate exactly at the time that the reward should have occurred (Fig. 1, bottom). This well-timed decrease in spike output shows that the expected time of reward delivery based on the occurrence of the light is also encoded in the fluctuations in dopaminergic activity (*18*). In contrast, very few dopamine neurons respond to stimuli that predict aversive outcomes.

The language used in the foregoing description already incorporates the idea that dopaminergic activity encodes expectations about external stimuli or reward. This interpretation of these data provides a link to an established body of computational theory (*6, 7*). From this perspective, one sees that dopamine neurons do not simply report the occurrence of appetitive events. Rather, their outputs appear to code for a deviation or error between the actual reward received and predictions of the time and magnitude of reward. These neurons are activated only if the time of the reward is uncertain, that is, unpredicted by any preceding cues. Dopamine neurons are therefore excellent feature detectors of the "goodness" of environmental events relative to learned predictions about those events. They emit a positive signal (increased spike production) if an appetitive event is better than predicted, no signal (no change in spike production) if an appetitive event occurs as predicted, and a negative signal (decreased spike production) if an appetitive event is worse than predicted (Fig. 1).

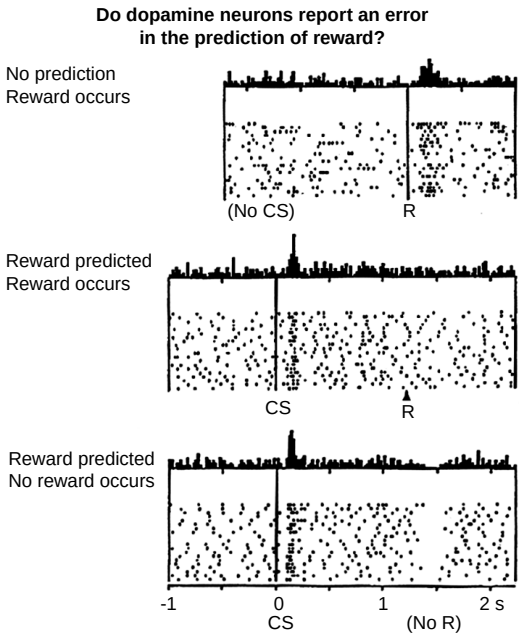


**Fig. 1.** Changes in dopamine neurons' output code for an error in the prediction of appetitive events. (**Top**) Before learning, a drop of appetitive fruit juice occurs in the absence of prediction—hence a positive error in the prediction of reward. The dopamine neuron is activated by this unpredicted occurrence of juice. (**Middle**) After learning, the conditioned stimulus predicts reward, and the reward occurs according to the prediction—hence no error in the prediction of reward. The dopamine neuron is activated by the reward-predicting stimulus but fails to be activated by the predicted reward (right). (**Bottom**) After learning, the conditioned stimulus predicts a reward, but the reward fails to occur because of a mistake in the behavioral response of the monkey. The activity of the dopamine neuron is depressed exactly at the time when the reward would have occurred. The depression occurs more than 1 s after the conditioned stimulus without any intervening stimuli, revealing an internal representation of the time of the predicted reward. Neuronal activity is aligned on the electronic pulse that drives the solenoid valve delivering the reward liquid (top) or the onset of the conditioned visual stimulus (middle and bottom). Each panel shows the peri-event time histogram and raster of impulses from the same neuron. Horizontal distances of dots correspond to real-time intervals. Each line of dots shows one trial. Original sequence of trials is plotted from top to bottom. CS, conditioned, reward-predicting stimulus; R, primary reward.

## Computational Theory and Model

The TD algorithm (*6, 7*) is particularly well suited to understanding the functional role played by the dopamine signal in terms of the information it constructs and broadcasts (*8, 10, 12*). This work has used fluctuations in dopamine activity in dual roles (i) as a supervisory signal for synaptic weight changes (*8, 10, 12*) and (ii) as a signal to influence directly and indirectly the choice of behavioral actions in humans and bees (*9–11*). Temporal difference methods have been used in a wide spectrum of engineering applications that seek to solve prediction

problems analogous to those faced by living creatures (*19*). Temporal difference methods were introduced into the psychological and biological literature by Richard Sutton and Andrew Barto in the early 1980s (*6*, *7*). It is therefore interesting that this method yields some insight into the output of dopamine neurons in primates.

There are two main assumptions in TD. First, the computational goal of learning is to use the sensory cues to predict a discounted sum of all future rewards $V(t)$ within a learning trial:

$$V(t) = E[\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \cdots] \quad (1)$$

where $r(t)$ is the reward at time $t$ and $E[\cdot]$ denotes the expected value of the sum of future rewards up to the end of the trial. $0 \leq \gamma \leq 1$ is a discount factor that makes rewards that arrive sooner more important than rewards that arrive later. Predicting the sum of future rewards is an important generalization over static conditioning models like the Rescorla-Wagner rule for classical conditioning (*1*–*4*). The second main assumption is the Markovian one, that is, the presentation of future sensory cues and rewards depends only on the immediate (current) sensory cues and not the past sensory cues.

As explained below, the strategy is to use a vector describing the presence of sensory cues $\mathbf{x}(t)$ in the trial along with a vector of adaptable weights $\mathbf{w}$ to make an estimate $\hat{V}(t)$ of the true $V(t)$. The reason that the sensory cue is written as a vector is explained below. The difficulty in adjusting weights $\mathbf{w}$ to estimate $V(t)$ is that the system (that is, the animal) would have to wait to receive all its future rewards in a trial $r(t+1)$, $r(t+2)$, . . . 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 $V(t)$ because it satisfies a condition of consistency through time:

$$V(t) = E[r(t) + \gamma V(t+1)] \quad (2)$$

An error in the estimated predictions can now be defined with information available at successive time steps:

$$\delta(t) = r(t) + \gamma\hat{V}(t+1) - \hat{V}(t) \quad (3)$$

This $\delta(t)$ is called the TD error and acts as a surrogate prediction error signal that is instantly available at time $t + 1$. As described below, $\delta(t)$ is used to improve the estimates of $V(t)$ and also to choose appropriate actions.

*Representing a stimulus through time*. We suggested above that a set of sensory cues along with an associated set of adaptable weights would suffice to estimate $V(t)$ (the discounted sum of future rewards). It is, however, not sufficient for the representation of each sensory cue (for example, a light) to have only one associated adaptable weight because such a model would not account for the data shown above—it would not be able to represent both the time of the cue and the time of reward delivery. These experimental data show that a sensory cue can predict reward delivery at arbitrary times into the near future. This conclusion holds for both the monkeys' behavior and the output of the dopamine neurons. If the time of reward delivery is changed relative to the time of cue onset, then the same cue will come to predict the new time of reward delivery. The way in which such temporal labels are constructed in neural tissue is not known, but it is clear that they exist (*20*).

Given these facts, we assume that each sensory cue consists of a vector of signals $\mathbf{x}(t) = \{x_1(t), x_2(t), \cdots\}$ that represent the light for variable lengths of time into the future, that is, $x_i(t)$ is 1 exactly $i$ time steps after the presentation of the light in the trial and 0 otherwise (Fig. 2B). Each component of $\mathbf{x}(t)$, $x_i(t)$, has its own prediction weight $w_i$ (Fig. 2B). This representation means that if the light comes on at time $s$, $x_1(s+1) = 1$, $x_2(s+2) = 1$, . . . represent the light at 1, 2, . . . time steps into the future and $w_1$, $w_2$, . . . are the respective weights. The net prediction for cue $\mathbf{x}(t)$ at time $t$ takes the simple linear form

$$\hat{V}(t) \equiv \hat{V}(\mathbf{x}(t)) = \Sigma_i w_i x_i(t) \quad (4)$$

This form of temporal representation is what Sutton and Barto (*7*) call a complete serial-compound stimulus and is related to Grossberg's spectral timing model (*21*). Unfortunately, virtually nothing is known about how the brain represents a stimulus for substantial periods of time into the future; therefore, all temporal representations are underconstrained from a biological perspective.

As in trial-based models like the Rescorla-Wagner rule, the adaptable weights $\mathbf{w}$ are improved according to the correlation between the stimulus representations and the prediction error. The change in weights from one trial to the next is

$$\Delta w_i = \alpha_x \Sigma_t x_i(t)\delta(t) \quad (5)$$

where $\alpha_x$ is the learning rate for cue $\mathbf{x}(t)$ and the sum over $t$ is taken over the course of a trial. It has been shown that under certain conditions this update rule (Eq. 5) will cause $\hat{V}(t)$ to converge to the true $V(t)$ (*22*). If there were many different sensory cues, each would have its own vector representation and its own vector of weights, and Eq. 4 would be summed over all the cues.

*Comparing model and data*. We now turn this apparatus toward the neural and behavioral data described above. To construct and use an error signal similar to the TD error above, a neural system would need to possess four basic features: (i) access to a measure of reward value $r(t)$; (ii) a signal measuring the temporal derivative of the ongoing prediction of reward $\gamma\hat{V}(t+1) - \hat{V}(t)$; (iii) a site where these signals could be summed; and (iv) delivery of the error signal to areas constructing the prediction in such a way that it can control plasticity.

It has been previously proposed that midbrain dopamine neurons satisfy features

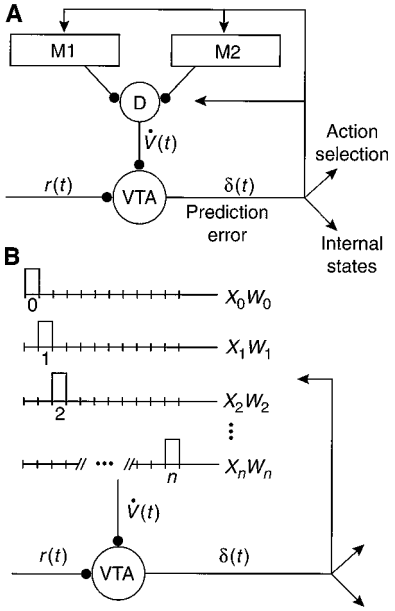


**Fig. 2.** Constructing and using a prediction error. (**A**) Interpretation of the anatomical arrangement of inputs and outputs of the ventral tegmental area (VTA). M1 and M2 represent two different cortical modalities whose output is assumed to arrive at the VTA in the form of a temporal derivative (surprise signal) $\dot{V}(t)$, which reflects the degree to which the current sensory state differs from the previous sensory state. The high degree of convergence forces $\dot{V}(t)$ to arrive at the VTA as a scalar signal. Information about reward $r(t)$ also converges on the VTA. The VTA output is taken as a simple linear sum $\delta(t) = r(t) + \dot{V}(t)$. The widespread output connections of the VTA make the prediction error $\delta(t)$ simultaneously available to structures constructing the predictions. (**B**) Temporal representation of a sensory cue. A cue like a light is represented at multiple delays $\mathbf{x}_n$ from its initial time of onset, and each delay is associated with a separate adjustable weight $\mathbf{w}_n$. These parameters $\mathbf{w}_n$ are adjusted according to the correlation of activity $\mathbf{x}_n$ and $\delta$ and through training come to act as predictions. This simple system stores predictions rather than correlations.

(i), (ii), and (iii) listed above (Fig. 2A) (*8*, *10*, *12*). As indicated in Fig. 2, the dopamine neurons receive highly convergent input from many brain regions. The model represents the hypothesis that this input arrives in the form of a surprise signal that measures the degree to which the current sensory state differs from the last sensory state. We assume that the dopamine neurons' output actually reflects $\delta(t) + b(t)$, where $b(t)$ is a basal firing rate (*12*). Figure 3 shows the training of the model on a task where a single sensory cue predicted the future delivery of a fixed amount of reward 20 time steps into the future. The prediction error signal (top) matches the activity of the real dopamine neurons over the course of learning. The pattern of weights that develops (bottom) provide the model's explanations for two well-described behavioral effects—blocking and secondary conditioning (*1*). The model accounts for the behavior of the dopamine neurons in a variety of other experiments in monkeys (*12*). The model also accounts for changes in dopaminergic activity if the time of the reward is changed (*18*).

The model makes two other testable predictions: (i) in the presence of multiple sensory cues that predict reward, the phasic activation of the neurons will transfer to the earliest consistent cue. (ii) After training on multiple sensory cues, omission of an intermediate cue will be accompanied by a phasic decrease in dopaminergic activity at the time that the cue formerly occurred. For example, after training a monkey on the temporal sequence light 1→light 2→reward, the dopamine neurons should respond phasically only to the onset of light 1. At this point, if light 2 is omitted on a trial, the activity in the neurons will depress at the time that light 2 would have occurred.

*Choosing and criticizing actions*. We showed above how the dopamine signal can be used to learn and store predictions; however, these same responses could also be used to influence the choice of appropriate actions through a connection with a technique called dynamic programming (*23*). We discuss below the connection to dynamic programming.

We introduce this use with a simple example. Suppose a rat must move through a maze to gain food. In the hallways of the maze, the rat has two options available to it: go forward a step or go backward a step. At junctions, the rat has three or four directions from which to choose. At each position, the rat has various actions available to it, and the action chosen will affect its future prospects for finding its way to food. A wrong turn at one point may not be felt as a mistake until many steps later when the rat runs into a dead end. How is the rat to know which action was crucial in leading it to the dead end? This is called the temporal credit assignment problem: Actions at one point in time can affect the acquisition of rewards in the future in complicated ways.

One solution to temporal credit assignment is to describe the animal as adopting and improving a "policy" that specifies how its actions are assigned to its states. Its state is the collection of sensory cues associated with each maze position. To improve a policy, the animal requires a means to evaluate the value of each maze position. The evaluation used in dynamic programming is the amount of summed future reward expected from each maze position provided that the animal follows its policy. The summed future rewards expected from some state [that is, $V(t)$] is exactly what the TD method learns, suggesting a connection with the dopamine signal.

As the rat above explores the maze, its predictions become more accurate. The predictions are considered "correct" once the average prediction error $\underline{\delta}(t)$ is 0. At this point, fluctuations in dopaminergic activity represent an important "economic evaluation" that is broadcast to target structures: Greater than baseline dopamine activity means the action performed is "better than expected" and less than baseline means "worse than expected." Hence, dopamine responses provide the information to implement a simple behavioral strategy—take [or learn to take (*24*)] actions correlated with increased dopamine activity and avoid actions correlated with decreases in dopamine activity.

A very simple such use of $\delta(t)$ as an evaluation signal for action choice is a form of learned klinokinesis (*25*), choosing one action while $\delta(t) > 0$, and choosing a new random action if $\delta(t) \leq 0$. This use of $\delta(t)$ has been shown to account for bee foraging behavior on flowers that yield variable returns (*9*, *11*). Figure 4 shows the way in which TD methods can construct for a mobile "creature" a useful map of the value of certain actions.

A TD model was equipped with a simple visual system (two, 200 by 200 pixel retinae) and trained on three different sensory cues (colored blocks) that differed in the amount of reward each contained (blue > green > red). The model had three neurons, each sensitive only to the percentage of one color in the visual field. Each color-sensitive neuron provides input to the prediction unit P (analog of VTA unit in Fig. 2) through a single weight. Dedicating only

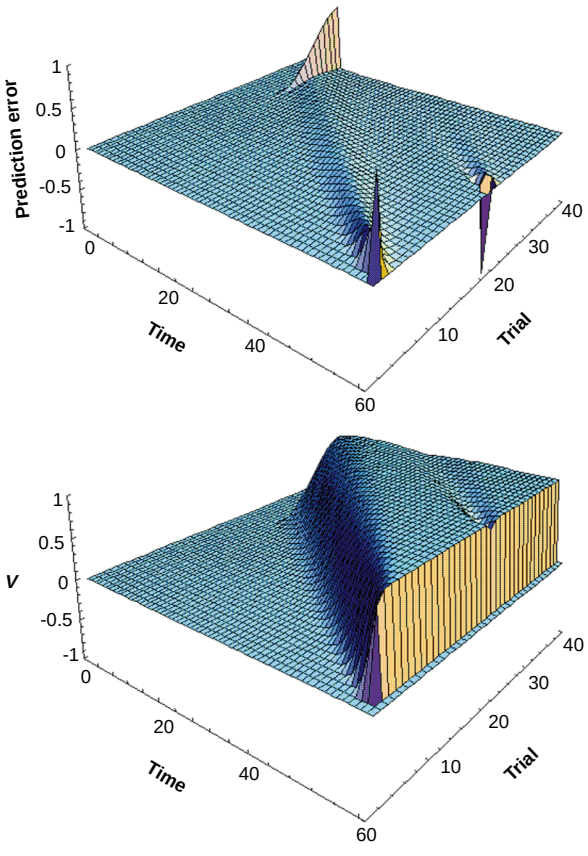


**Fig. 3.** Development of prediction error signal through training. (**Top**) Prediction error (changes in dopamine neuron output) as a function of time and trial. On each trial, a sensory cue is presented at time step 10 and time step 20 followed by reward delivery [$r(t) = 1$] at time step 60. On trial 0, the presentation of the two cues causes no change because the associated weights are initially set to 0. There is, however, a strong positive response (increased firing rate) at the delivery of reward at time step 60. By repeating the pairing of the sensory cues followed in time by reward, the transient response of the model shifts to the time of the earliest sensory cue (time step 10). Failure to deliver the reward during an intermediate trial causes a large negative fluctuation in the model's output. This would be seen in an experiment as a marked decrease in spike output at the time that reward should have been delivered. In this example, the timing of reward delivery is learned well before any response transfers to the earliest sensory cue. (**Bottom**) The value function $V(t)$. The weights are all initially set to 0 (trial 0). After the large prediction error occurs on trial 0, the weights begin to grow. Eventually they all saturate to 1 so that the only transient is the unpredicted onset of the first sensory cue. The depression in the surface results from the error trial where the reward was not delivered at the expected time.

a single weight to each cue limits this "creature" to a one time step prediction on the basis of its current state. After experiencing each type of object multiple times, the weights reflect the relative amounts of reward in each object, that is, $w_b > w_g > w_r$. These three weights equip the creature with a kind of cognitive map or "value surface" with which to assay its possible actions (Fig. 4B).

The value surface above the arena is a plot of the value function $V(x, y)$ (height) when the creature is placed in the indicated corner and looks at every position $(x, y)$ in the arena. The value $V(x, y)$ of looking at each position $(x, y)$ is computed as a linear function of the weights $(w_b, w_g, w_r)$ associated with activity induced in the color-sensitive units. As this "creature" changes its direction of gaze from one position $(x_0, y_0)$ at time $t$ to another position $(x_1, y_1)$ at time $t + 1$, the difference in the values of these two positions $V(t + 1) - V(t)$ is available as the output $\delta(t)$ of the prediction neuron P. In this example, when the creature looks from point 1 to point 2, the percentage of blue in its visual field increases. This increase is available as a positive fluctuation ("things are better than expected") in the output $\delta(t)$ of neuron P. Similarly, looking from point 2 to point 1 causes a large negative fluctuation in $\delta(t)$ ("things are worse than expected"). As discussed above, these fluctuations could be used by some target structure to decide whether to move in the direction of sight. Directions associated with a positive prediction error are likely to yield increased future returns.

This example illustrates how only three stored quantities (weights associated with each color) and the capacity to look at different locations endow this simple "creature" with a useful map of the quality of different directions in the arena. This same model has been given simple card-choice tasks analogous to those given to humans (*26*), and the model matches well the human behavior. It is also interesting that humans develop a predictive galvanic skin response that predicts appropriately which card decks are good and which are bad (*26*).

## Summary and Future Questions

We have reviewed evidence that supports the proposal that dopamine neurons in the VTA and the substantia nigra report ongoing prediction errors for reward. The output of these neurons is consistent with a scalar prediction error signal; therefore, the delivery of this signal to target structures may influence the processing of predictions and the choice of reward-maximizing actions. These conclusions are supported by data on the activity changes of these neurons during the acquisition and expression of a range of simple conditioning tasks. This representation of the experimental data raises a number of important issues for future work.

The first issue concerns temporal representations, that is, how is any stimulus represented through time? A large body of behavioral data show that animals can keep track of the time elapsed from the presentation of a CS and make precise predictions accordingly. We adopted a very simple model of this capacity, but experiments have yet to suggest where or how the temporal information is constructed and used by the brain. It is not yet clear how far into the future such predictions can be made; however, one suspects that they will be longer than the predictions made by structures that mediate cerebellar eyeblink conditioning and motor learning displayed by the vestibulo-ocular reflex (*27*). The time scales that are ethologically important to a particular creature should provide good constraints when searching for mechanisms that might construct and distribute temporal labels in the cerebral cortex.

A second issue is information about aversive events. The experimental data suggest that the dopamine system provides information about appetitive stimuli, not aversive stimuli. It is possible however that the absence of an expected reward is interpreted as a kind of "punishment" to some other system to which the dopamine neurons send their output. It would then be the responsibility of these targets to pass out information about the degree to which the nondelivery of reward was "punishing." It was long ago proposed that rewards and punishments represent opponent processes and that the dynamics of opponency might be responsible for many puzzling effects in conditioning (*28*).

A third issue raised by the model is the relation between scalar signals of appetitive values and vector signals with many components, including those that represent primary rewards and predictive stimuli. Simple models like the one presented above may be able to learn with a scalar signal only if the scope of choices is limited. Behavior in more realistic environmental situations requires vector signaling of the type of rewards and of the various physical components of the predictive stimuli. Without the capacity to discriminate which stimuli are responsible for fluctuations in a broadcast scalar error signal, an agent may learn inappropriately, for example, it may learn to approach food when it is actually thirsty.

Dopamine neurons emit an excellent appetitive error (teaching) signal without indicating further details about the appetitive event. It is therefore likely that other reward-processing structures subserve the analysis and discrimination of appetitive events without constituting particularly efficient teaching signals. This putative division of labor between the analysis of physical and functional attributes and scalar

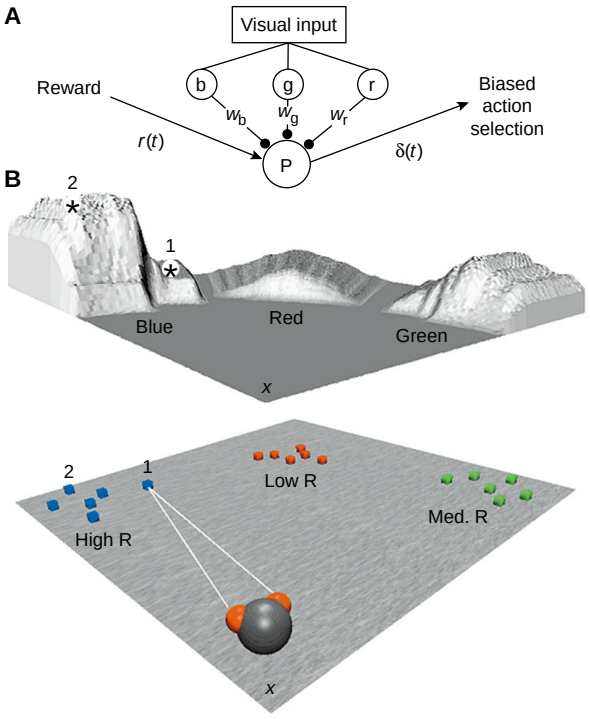


**Fig. 4.** Simple cognitive maps can be easily built and used. (**A**) Architecture of the TD model. Three color-sensitive units (b, g, r) report, respectively, the percentage of blue, green, and red in the visual field. Each unit influences neuron P (VTA analog) through a single weight. The colored blocks contain varying amounts of reward with blue > green > red. After training, the weights $(w_b, w_g, w_r)$ reflect this difference in reward content. Using only a single weight for each sensory cue, the model can make only one-time step predictions; however, combined with its capacity to move its head or walk about the arena, a crude "value-map" is available in the output $\delta(t)$ of neuron P. (**B**) Value surface for the arena when the creature is positioned in the corner as indicated. The height of the surface codes for the value $V(x, y)$ of each location when viewed from the corner where the "creature" is positioned. All the creature needs to do is look from one location to another (or move from one position to another), and the differences in value $V(t + 1) - V(t)$ are coded in the changes in the firing rate of P (see text).

evaluation signals raises a fourth issue—attention.

The model does not address the attentional functions of some of the innervated structures, such as the nucleus accumbens and the frontal cortex. Evidence suggests that these structures are important for cases in which different amounts of attention are paid to different stimuli. There is, however, evidence to suggest that the required attentional mechanisms might also operate at the level of the dopamine neurons. Their responses to novel stimuli will decrement with repeated presentation and they will generalize their responses to nonappetitive stimuli that are physically similar to appetitive stimuli (*29*). In general, questions about attentional effects in dopaminergic systems are ripe for future work.

The suggestions that a scalar prediction-error signal influences behavioral choices receives support from the preliminary work on human decision-making and from the fact that changes in dopamine activity fluctuations parallel changes in the behavioral performance of the monkeys (*30*). In the mammalian brain, the striatum is one site where this kind of scalar evaluation could have a direct effect on action choice, and activity relating to conditioned stimuli is seen in the striatum (*31*). The widespread projection of dopamine axons to striatal neurons gives rise to synapses at dendritic spines that are also contacted by excitatory inputs from cortex (*32*). This may be a site where the dopamine signal influences behavioral choices by modulating the level of competition in the dorsal striatum. Phasic dopamine signals may lead to an augmentation of excitatory influences in the striatum (*33*), and there is evidence for striatal plasticity after pulsatile application of dopamine (*34*). Plasticity could mediate the learning of appropriate policies (*24*).

The possibilities in the striatum for using a scalar evaluation signal carried by changes in dopamine delivery are complemented by interesting possibilities in the cerebral cortex. In prefrontal cortex, dopamine delivery has a dramatic influence on working memory (*35*). Dopamine also modulates cognitive activation of anterior cingulate cortex in schizophenic patients (*36*). Clearly, dopamine delivery has important cognitive consequences at the level of the cerebral cortex. Under the model presented here, changes in dopaminergic activity distribute prediction errors to widespread target structures. It seems reasonable to require that the prediction errors be delivered primarily to those regions most responsible for making the predictions; otherwise, one cortical region would have to deal with prediction errors engendered by the bad guesses of another region. From this point of view, one could expect there to be a mechanism that coupled local activity in the cortex to an enhanced sensitivity of nearby dopamine terminals to differences from baseline in spike production along their parent axon. There is experimental evidence that supports this possibility (*37*).

Neuromodulatory systems like dopamine systems are so named because they were thought to modulate global states of the brain at time scales and temporal resolutions much poorer than other systems like fast glutamatergic connections. Although this global modulation function may be accurate, the work discussed here shows that neuromodulatory systems may also deliver precisely timed information to specific target structures to influence a number of important cognitive functions.

## REFERENCES AND NOTES


1. A. Dickinson, *Contemporary Animal Learning Theory* (Cambridge Univ. Press, Cambridge, 1980); N. J. Mackintosh, *Conditioning and Associative Learning* (Oxford Univ. Press, Oxford, 1983); C. R. Gallistel, *The Organization of Learning* (MIT Press, Cambridge, MA, 1990); L. A. Real, *Science* **253**, 980 (1991).
2. I. P. Pavlov, *Conditioned Reflexes* (Oxford Univ. Press, Oxford, 1927); B. F. Skinner, *The Behavior of Organisms* (Appleton-Century-Crofts, New York, 1938); J. Olds, *Drives and Reinforcement* (Raven, New York 1977); R. A. Wise, in *The Neuropharmacological Basis of Reward*, J. M. Liebeman and S. J. Cooper, Eds. (Clarendon Press, New York, 1989); N. W. White and P. M. Milner, *Annu. Rev. Psychol.* **43**, 443 (1992); T. W. Robbins and B. J. Everitt, *Curr. Opin. Neurobiol.* **6**, 228 (1996).
3. R. A. Rescorla and A. R. Wagner, in *Classical Conditioning II: Current Research and Theory*, A. H. Black and W. F. Prokasy, Eds. (Appleton-Century-Crofts, New York, 1972), pp. 64–69.
4. N. J. Mackintosh, *Psychol. Rev.* **82**, 276 (1975); J. M. Pearce and G. Hall, *ibid.* **87**, 532 (1980).
5. L. J. Kamin, in *Punishment and Aversive Behavior*, B. A. Campbell and R. M. Church, Eds. (Appleton-Century-Crofts, New York (1969), pp. 279–296.
6. R. S. Sutton and A. G. Barto, *Psychol. Rev.* **88** (no. 2), 135 (1981); R. S. Sutton, *Mach. Learn.* **3**, 9 (1988).
7. R. S. Sutton and A. G. Barto, *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (Seattle, WA, 1987); in *Learning and Computational Neuroscience*, M. Gabriel and J. Moore, Eds. (MIT Press, Cambridge, MA, 1989). For specific application to eyeblink conditioning, see J. W. Moore *et al*., *Behav. Brain Res.* **12**, 143 (1986).
8. S. R. Quartz, P. Dayan, P. R. Montague, T. J. Sejnowski, *Soc. Neurosci. Abstr.* **18**, 1210 (1992); P. R. Montague, P. Dayan, S. J. Nowlan, A. Pouget, T. J. Sejnowski, in *Advances in Neural Information Processing Systems 5*, S. J. Hanson, J. D. Cowan, C. L. Giles, Eds. (Morgan Kaufmann, San Mateo, CA, 1993), pp. 969–976.
9. P. R. Montague, P. Dayan, T. J. Sejnowski, in *Advances in Neural Information Processing Systems 6*, G. Tesauro, J. D. Cowan, J. Alspector, Eds. (Morgan Kaufmann, San Mateo, CA, 1994), pp. 598–605.
10. P. R. Montague and T. J. Sejnowski, *Learn. Mem.* **1**, 1 (1994); P. R. Montague, *Neural-Network Approaches to Cognition—Biobehavioral Foundations*, J. Donahoe, Ed. (Elsevier, Amsterdam, in press); P. R. Montague and P. Dayan, *A Companion to Cognitive Science*, W. Bechtel and G. Graham, Eds. (Blackwell, Oxford, in press).
11. P. R. Montague, P. Dayan, C. Person, T. J. Sejnowski, *Nature* **377,** 725 (1995).
12. P. R. Montague, P. Dayan, T. J. Sejnowski, *J. Neurosci.* **16**, 1936 (1996).
13. Other work has suggested an interpretation of monoaminergic influences similar to that taken above (*8–12*) [K. J. Friston, G. Tononi, G. N. Reeke, O. Sporns, G. M. Edelman, *Neuroscience* **59**, 229 (1994); J. C. Houk, J. L. Adams, A. G. Barto, in *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, D. G. Beiser, Eds. (MIT Press, Cambridge, MA, 1995)], pp. 249–270. Other models of monoaminergic influences have considered what could be called attention-based accounts (*4*) rather than prediction error–based explanations [D. Servan-Schreiber, H. Printz, J. D. Cohen, *Science* **249**, 892 (1990)].
14. G. F. Koob, *Semin. Neurosci.* **4**, 139 (1992); R. A. Wise and D. C. Hoffman, *Synapse* **10**, 247 (1992); G. DiChiara, *Drug Alcohol Depend.* **38**, 95 (1995).
15. A. G. Phillips, S. M. Brooke, H. C. Fibiger, *Brain Res.* **85**, 13 (1975); A. G. Phillips, D. A. Carter, H. C. Fibiger, *ibid.* **104**, 221 (1976); F. Mora and R. D. Myers, *Science* **197**, 1387 (1977); A. G. Phillips, F. Mora, E. T. Rolls, *Psychopharmacology* **62**, 79 (1979); D. Corbett and R. A. Wise, *Brain Res.* **185**, 1 (1980); R. A. Wise and P.-P. Rompre, *Annu. Rev. Psychol.* **40**, 191 (1989).
16. R. A. Wise, *Behav. Brain Sci.* **5**, 39 (1982); R. J. Beninger, *Brain Res. Rev.* **6**, 173 (1983); _____ and B. L. Hahn, *Science* **220**, 1304 (1983); R. J. Beninger, *Brain Res. Bull.* **23**, 365 (1989); M. LeMoal and H. Simon, *Physiol. Rev.* **71**, 155 (1991); T. W. Robbins and B. J. Everitt, *Semin. Neurosci.* **4**, 119 (1992).
17. W. Schultz, *J. Neurophysiol.* **56**, 1439 (1986); R. Romo and W. Schultz, *ibid.* **63**, 592 (1990); W. Schultz and R. Romo, *ibid.*, p. 607; T. Ljungberg, P. Apicella, W. Schultz, *ibid.* **67**, 145 (1992); W. Schultz, P. Apicella, T. Ljungberg, *J. Neurosci.* **13**, 900 (1993); J. Mirenowicz and W. Schultz, *J. Neurophysiol.* **72**, 1024 (1994); W. Schultz *et al*., in *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, D. G. Beiser, Eds. (MIT Press, Cambridge, MA, 1995), pp. 233–248; J. Mirenowicz and W. Schultz, *Nature* **379**, 449 (1996).
18. Recent experiments showed that the simple displacement of the time of reward delivery resulted in dopamine responses. In a situation in which neurons were not driven by a fully predicted drop of juice, activations reappeared when the juice reward occurred 0.5 s earlier or later than predicted. Depressions were observed at the normal time of juice reward only if reward delivery was late [J. R. Hollerman and W. Schultz, *Soc. Neuroci. Abstr.* **22**, 1388 (1996)].
19. G. Tesauro, *Commun. ACM* **38**, 58 (1995); D. P. Bertsekas and J. N. Tsitsiklis, *Neurodynamic Programming* (Athena Scientific, Belmont, NJ, 1996).
20. R. M. Church, in *Contemporary Learning Theories: Instrumental Conditioning Theory and the Impact of Biological Constraints on Learning*, S. B. Klein and R. R. Mowrer, Eds. (Erlbaum, Hillsdale, NJ, 1989), p. 41; J. Gibbon, *Learn. Motiv.* **22**, 3 (1991).
21. S. Grossberg and N. A. Schmajuk, *Neural Networks* **2**, 79 (1989); S. Grossberg and J. W. L. Merrill, *Cognit. Brain Res.* **1**, 3 (1992).
22. P. Dayan, *Mach. Learn.* **8**, 341 (1992); P. Dayan and T. J. Sejnowski, *ibid.* **14**, 295 (1994); T. Jaakkola, M. I. Jordan, S. P. Singh, *Neural Computation* **6**, 1185 (1994).
23. R. E. Bellman, *Dynamic Programming* (Princeton Univ. Press, Princeton, NJ, 1957); R. A. Howard, *Dynamic Programming and Markov Processes* (MIT Press, Cambridge, MA, 1960).
24. A. G. Barto, R. S. Sutton, C. W. Anderson, *IEEE Trans. Syst. Man Cybernetics* **13**, 834 (1983).
25. Bacterial klinokinesis has been described in great detail. Early work emphasized the mechanisms required for bacteria to climb gradients of nutrients. See R. M. Macnab and D. E. Koshland, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 2509 (1972); N. Tsang, R. Macnab, D. E. Koshland Jr., *Science* **181**, 60 (1973); H. C. Berg and R. A. Anderson, *Nature* **245**, 380 (1973); H. C. Berg *ibid.* **254**, 389 (1975); J. L. Spudich and D. E. Koshland, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 710 (1975). The klinokinetic action-selection mechanism causes a TD model to climb hills

defined by the sensory weights, that is, the model will climb the surface defined by the value function *V*.
26. A. R. Damasio, *Descartes' Error* (Putnam, New York, 1994); A. Bechara, A. R. Damasio, H. Damasio, S. Anderson, *Cognition* **50**, 7 (1994).
27. S. P. Perrett, B. P. Ruiz, M. D. Mauk *J. Neurosci.* **13**, 1708 (1993); J. L. Raymond, S. G. Lisberger, M. D. Mauk *Science* **272**, 1126 (1996).
28. S. Grossberg, *Math. Biosci.* **15**, 253 (1972); R. L. Solomon and J. D. Corbit, *Psychol. Rev.* **81**, 119 (1974); S. Grossberg, *ibid.* **89**, 529 (1982).
29. W. Schultz and R. Romo, *J. Neurophysiol.* **63**, 607 (1990); T. Ljungberg, P. Apicella, W. Schultz, *ibid.* **67**, 145 (1992); J. Mirenowicz and W. Schultz, *Nature* **379**, 449 (1996).
30. W. Schultz, P. Apicella, T. Ljungberg, *J. Neurosci.* **13**, 900 (1993).
31. T. Aosaki *et al*., *ibid.* **14**, 3969 (1994); A. M. Graybiel *Curr. Opin. Neurobiol.* **5**, 733 (1995); *Trends Neurosci.* **18**, 60 (1995). Recent models of sequence generation in the striatum use fluctuating dopamine input as a scalar error signal [G. S. Berns and T. J. Sejnowski, in *Neurobiology of Decision Making*, A. Damasio, Ed. (Springer-Verlag, Berlin, 1996), pp. 101–113.
32. T. F. Freund, J. F. Powell, A. D. Smith, *Neuroscience* **13**, 1189 (1984); Y. Smith, B. D. Bennett, J. P. Bolam, A. Parent, A. F. Sadikot, *J. Comp. Neurol.* **344**, 1 (1994).
33. C. Cepeda, N. A. Buchwald, M. S. Levine, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 9576 (1993).
34. J. R. Wickens, A. J. Begg, G. W. Arbuthnott, *Neuroscience* **70**, 1 (1996).
35. P. S. Goldman-Rakic, C. Leranth, M. S. Williams, N. Mons, M. Geffard, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9015 (1989); T. Sawaguchi and P. S. Goldman-Rakic, *Science* **251**, 947 (1991); G. V. Williams and P. S. Goldman-Rakic, *Nature* **376**, 572 (1995).
36. R. J. Dolan *et al*., *Nature*, **378** 180 (1995).
37. P. R. Montague, C. D. Gancayco, M. J. Winn, R. B. Marchase, M. J. Friedlander, *Science* **263**, 973 (1994). The mechanistic suggestion requires that local cortical activity (presumably glutamatergic) increases the sensitivity of nearby dopamine terminals to differences from baseline in spike production along their parent axon. This may result from local increases in nitric oxide production. In this manner, baseline dopamine release remains constant in inactive cortical areas while active cortical areas feel strongly the effect of increases and decreases in dopamine delivery due to increases and decreases in spike production along the parent dopamine axon.
38. We thank A. Damasio and T. Sejnowski for comments and criticisms, and C. Person for help in generating figures. The theoretical work received continuing support from the Center for Theoretical Neuroscience at Baylor College of Medicine and the National Institutes of Mental Health (NIMH) (P.R.M.). P.D. was supported by Massachusetts Institute of Technology and the NIH. The primate studies were supported by the Swiss National Science Foundation, the McDonnell-Pew Foundation (Princeton), the Fyssen Foundation (Paris), the Fondation pour la Recherche Midicale (Paris), the United Parkinson Foundation (Chicago), the Roche Research Foundation (Basel), the NIMH (Bethesda), and the British Council.

# Language Acquisition and Use: Learning and Applying Probabilistic Constraints

Mark S. Seidenberg

What kinds of knowledge underlie the use of language and how is this knowledge acquired? Linguists equate knowing a language with knowing a grammar. Classic "poverty of the stimulus" arguments suggest that grammar identification is an intractable inductive problem and that acquisition is possible only because children possess innate knowledge of grammatical structure. An alternative view is emerging from studies of statistical and probabilistic aspects of language, connectionist models, and the learning capacities of infants. This approach emphasizes continuity between how language is acquired and how it is used. It retains the idea that innate capacities constrain language learning, but calls into question whether they include knowledge of grammatical structure.

Modern thinking about language has been dominated by the views of Noam Chomsky, who created the generative paradigm within which most research has been conducted for over 30 years (*1*). This approach continues to flourish (*2*), and although alternative theories exist, they typically share Chomsky's assumptions about the nature of language and the goals of linguistic theory (*3*). Research on language has arrived at a particularly interesting point, however, because of important developments outside of the linguistic mainstream that are converging on a different view of the nature of language. These developments represent an important turn of events in the history of ideas about language.

## The Standard Theory

The place to begin is with Chomsky's classic questions (*4*): (i) what constitutes knowledge of a language, (ii) how is this knowledge acquired, and (iii) how is it put to use? The standard theory provides the following answers (*1*–*5*).

In answer to the first question, what one knows is a grammar, a complex system of rules and constraints that allows people to distinguish grammatical from ungrammatical sentences. The grammar is an idealization that abstracts away from a variety of so-called performance factors related to language use. The Competence Hypothesis is that this idealization will facilitate the identification of generalizations about linguistic knowledge that lie beneath overt behavior, which is affected by many other factors. Many phenomena that are prominent characteristics of language use are therefore set aside. The clear cases that are often cited in separating competence from performance include dysfluencies and errors. In practice, however, the competence theory also excludes other factors that affect language use, including the nature of the perceptual and motor systems that are used; memory capacities that limit the complexity of utterances that can be produced or understood; and reasoning capacities used in comprehending text or discourse. The competence theory also excludes information about statistical and probabilistic aspects of language—for example, the fact that verbs differ in how often they occur in transitive and intransitive sentences ("John ate the candy" versus "John ate," respectively), or the fact that when the subject of the verb "break" is animate, it is typically the agent of the action, but when it is inanimate, it is typically the entity being broken (compare "John broke the glass" with "The glass broke"). That this information should be excluded was the point of Chomsky's famous sentence "Colorless green ideas sleep furiously" and the accompanying observation that, "I think that we are forced to conclude that . . . probabilistic models give no particular insight into some of the basic problems of syntactic structure" (*6*). Finally, the competence theory also disregards the communicative functions of language and how they are achieved. These aspects of language are acknowledged as important but considered separable from core grammatical knowledge.

The grammar's essential properties include generativity (it can be used to produce and comprehend an essentially infinite number of sentences); abstractness of structure (it uses representations that are not overtly marked in the surface forms of utterances); modularity (the grammar is organized into components with different types of representations governed by different principles); and domain specificity (language exhibits properties that are not seen in other aspects of cognition; therefore, it cannot be an expression of general capacities to think and to learn).

The second question regarding language

Neuroscience Program, University of Southern California, Los Angeles, CA 90089–2520, USA. E-mail: marks@gizmo.usc.edu

# A Neural Substrate of Prediction and Reward

Wolfram Schultz, Peter Dayan and P. Read Montague

*Science* **275** (5306), 1593-1599.
DOI: 10.1126/science.275.5306.1593