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).

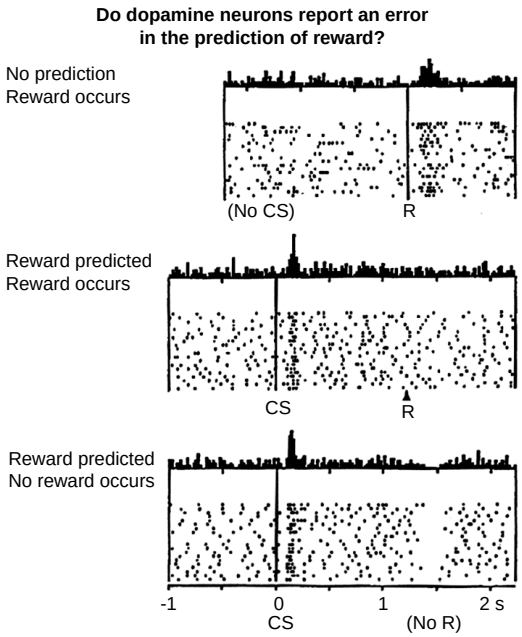


Fig. 1. Changes in dopamine neurons' output code for an error in the prediction of appetitive events. (**Top**) Before learning, a drop of appetitive fruit juice occurs in the absence of prediction—hence a positive error in the prediction of reward. The dopamine neuron is activated by this unpredicted 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 α_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

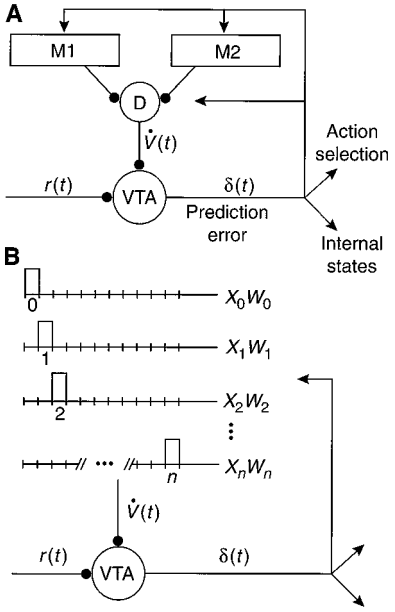


Fig. 2. Constructing and using a prediction error. (**A**) Interpretation of the anatomical arrangement of inputs and outputs of the ventral tegmental area (VTA). M1 and M2 represent two different cortical modalities whose output is assumed to arrive at the VTA in the form of a temporal derivative (surprise 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 δ 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

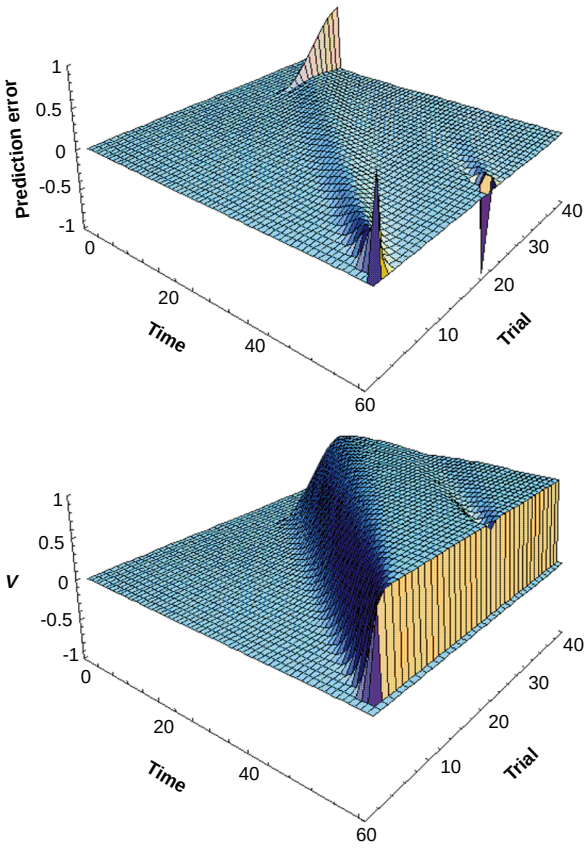


Fig. 3. Development of prediction error signal through training. (**Top**) Prediction error (changes in dopamine neuron output) as a function of time and trial. On each trial, a sensory cue is presented at time step 10 and time step 20 followed by reward delivery [$r(t) = 1$] at time step 60. On trial 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

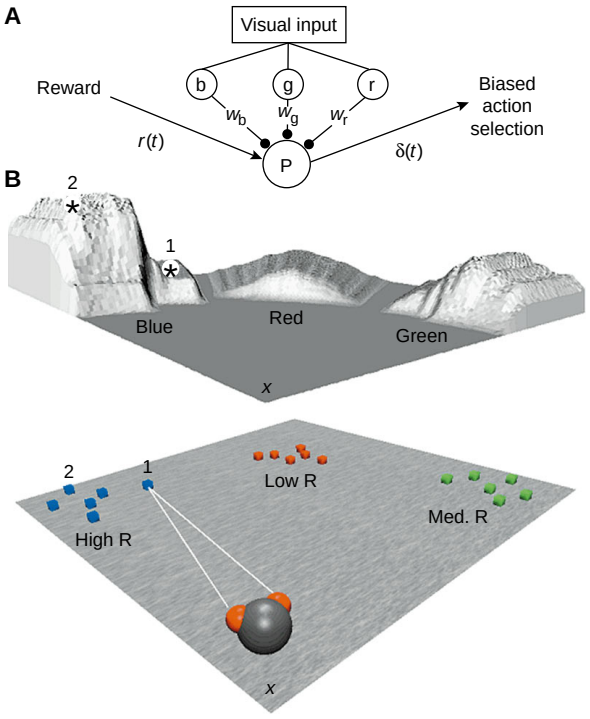


Fig. 4. Simple cognitive maps can be easily built and used. (**A**) Architecture of the TD model. Three color-sensitive units (b, g, r) report, respectively, the percentage of blue, green, and red in the visual field. Each unit influences neuron P (VTA analog) through a single weight. The colored blocks 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

Vocal Experimentation in the Juvenile Songbird Requires a Basal Ganglia Circuit

Bence P. Ölveczky[1,2], Aaron S. Andalman[1], Michale S. Fee[1*]

1 McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Harvard Society of Fellows, Harvard University, Cambridge, Massachusetts, United States of America

Songbirds learn their songs by trial-and-error experimentation, producing highly variable vocal output as juveniles. By comparing their own sounds to the song of a tutor, young songbirds gradually converge to a stable song that can be a remarkably good copy of the tutor song. Here we show that vocal variability in the learning songbird is induced by a basal-ganglia-related circuit, the output of which projects to the motor pathway via the lateral magnocellular nucleus of the nidopallium (LMAN). We found that pharmacological inactivation of LMAN dramatically reduced acoustic and sequence variability in the songs of juvenile zebra finches, doing so in a rapid and reversible manner. In addition, recordings from LMAN neurons projecting to the motor pathway revealed highly variable spiking activity across song renditions, showing that LMAN may act as a source of variability. Lastly, pharmacological blockade of synaptic inputs from LMAN to its target premotor area also reduced song variability. Our results establish that, in the juvenile songbird, the exploratory motor behavior required to learn a complex motor sequence is dependent on a dedicated neural circuit homologous to cortico-basal ganglia circuits in mammals.

Citation: Ölveczky BP, Andalman AS, Fee MS (2005) Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. PLoS Biol 3(5): e153.

Introduction

The acquisition of complex motor sequences, such as swinging a golf club or playing the piano, can be thought of as reinforcement learning. This learning process requires the exploration of a range of motor actions and the concomitant evaluation of the resulting performance, reinforcing motor programs that lead to improved outcomes [1]. Similarly, juvenile songbirds explore a large range of vocalizations by continuously varying their song [2], utilizing auditory feedback to improve their performance [3]. Thus, song learning encompasses the two ingredients of reinforcement learning: exploratory motor behavior, and performance evaluation.

In the songbird, two main neural pathways are involved in song production and song learning (Figure 1A). The "motor pathway" controls the vocal motor program through the hierarchical organization of several premotor nuclei [4]. A key nucleus in the motor pathway is the robust nucleus of the arcopallium (RA), which projects to brainstem nuclei controlling the vocal and respiratory muscles [5]. During singing, RA neurons in adult birds generate a highly stereotyped sequence of bursts [6,7], which appear to be driven by precisely timed inputs from a higher premotor vocal area, nucleus HVC [8]. RA also receives input from the "anterior forebrain pathway" (AFP), a circuit homologous to the basal ganglia thalamo-cortical loops [9,10] that may be involved in controlling motor behavior and stereotypy in mammals [11]. Lesions of the AFP in juvenile zebra finches have devastating effects on song development, whereas the same manipulations in adults have few short-term consequences for song production [12,13].

While the critical importance of the AFP for song learning has been established, its specific role remains unknown [14]. It has been proposed that the AFP may be involved in comparing the auditory feedback of the bird's vocal output with a stored auditory template of the desired song—an evaluation process that could provide a corrective signal to the motor pathway needed for learning [15]. However, recent results showing that the firing patterns of neurons in the lateral magnocellular nucleus of the nidopallium (LMAN) of adult birds are insensitive to distorted auditory feedback have called this idea into question [16,17]. Here we test the alternative hypothesis that, in juvenile songbirds, LMAN is involved in generating vocal variability [18]—the other important ingredient of reinforcement learning.

Results

Our approach was to transiently inactivate LMAN in juvenile zebra finches (n = 7 birds, see Materials and Methods), and observe whether and how their songs were affected. Birds were briefly head-restrained, and injections of a sodium channel blocker, tetrodotoxin (TTX, 30 nl, 50 μM), were made in LMAN in both hemispheres, inactivating the nucleus (see Figures S1 and S2). After injections, birds were returned to a sound-isolated chamber, where they typically began to sing after 0.5–1.5 h. In all birds probed, LMAN inactivation resulted in an immediate loss of acoustic variability across song renditions. The effect was particularly dramatic in birds at an early stage of song development (approximately 55 d post hatch [dph]) because these birds

Received February 4, 2005; Accepted March 1, 2005; Published March 29, 2005
DOI: 10.1371/journal.pbio.0030153

Copyright: © 2005 Ölveczky et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations: AFP, anterior forebrain pathway; AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid; AP5, 2-amino-5-phosphonovalerate; dph, days post hatch; LMAN, lateral magnocellular nucleus of the nidopallium; MMAN, medial magnocellular nucleus of the nidopallium; NMDA, N-methyl-D-aspartate; RA, robust nucleus of the arcopallium; TTX, tetrodotoxin

Academic Editor: Wolfram Schultz, University of Cambridge, United Kingdom

*To whom correspondence should be addressed. E-mail: fee@mit.edu

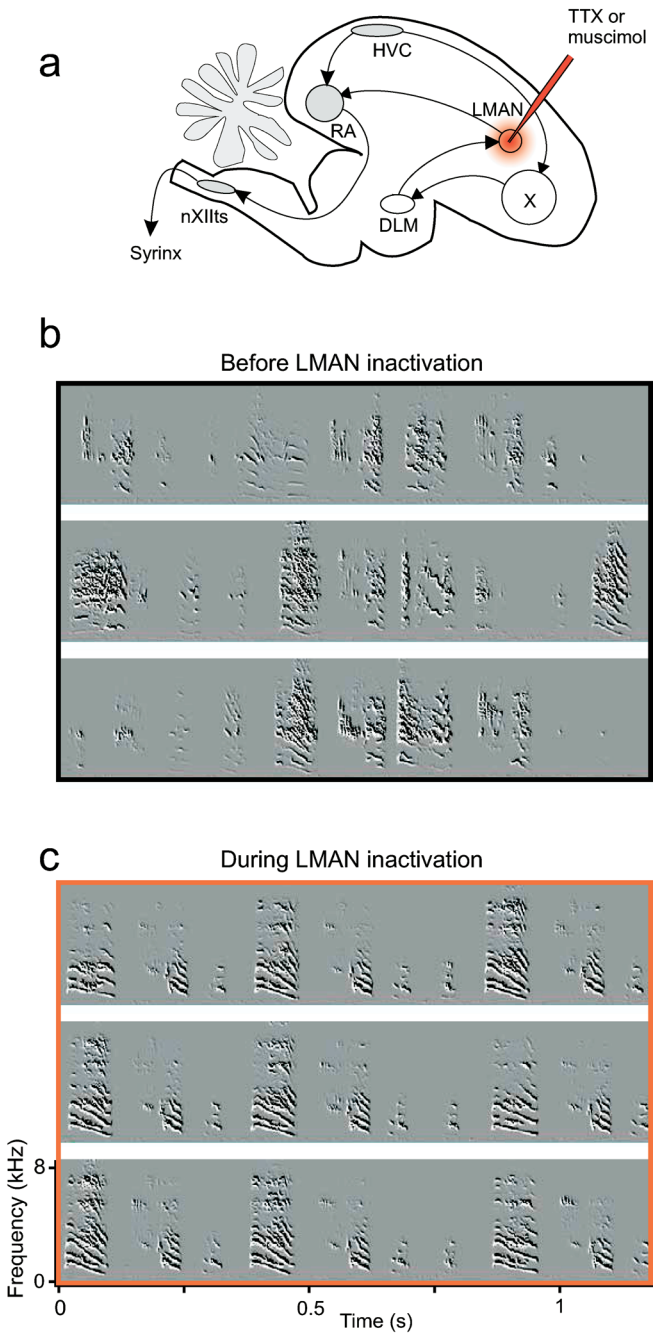


Figure 1. Inactivation of LMAN Significantly Reduces Vocal Experimentation, Making the Otherwise Variable Song of the Juvenile Zebra Finch Highly Stereotyped

(A) Two major pathways in the vocal control system of the songbird. The motor pathway (gray) includes motor cortex analogs HVC and RA, while the AFP (white), a basal ganglia thalamo-cortical circuit, consists of Area X, the dorsolateral anterior thalamic nucleus (DLM), and LMAN, which, in turn, projects to RA. To inactivate the output of the AFP, injections of TTX and muscimol (red bolus) were made into LMAN.

(B) Examples of a juvenile zebra finch song (57 dph) showing large variability in the sequence and acoustic structure of song syllables.

(C) Inactivating LMAN with TTX produces an immediate reduction of sequence and acoustic variability, revealing a highly stereotyped song produced by the motor pathway.

The song snippets shown in (B) and (C) are from consecutive song bouts, immediately before and 1 h after drug injection. Songs are displayed as spectral derivatives calculated as described [36]. The frequency range displayed is 0–8.6 kHz. For audio of song bouts before and during LMAN inactivation in this bird, refer to Audios S1 and S2, and S3 and S4, respectively.

DOI: 10.1371/journal.pbio.0030153.g001

normally exhibit greater song variability (Figures 1B, 1C, and S3; Audios S1–S4).

To quantify song variability, experiments were carried out in slightly older birds with less sequence and acoustic variability ($n = 6$ birds; age range, 59–72 dph) (Figure 2). This allowed us to reliably identify song syllables, the basic acoustic units of zebra finch song, across song renditions (Figure 2A). The variability score *(V)*—a measure reflecting the acoustic variability of a syllable across renditions (see Materials and Methods)—was calculated for all identified syllables before and after TTX injection. Without exception, the syllables showed a highly significant reduction in variability as a consequence of LMAN inactivation (Figure 2B; $n = 25$ syllables; $\langle V \rangle_{\text{before}} = 0.46$, $\langle \Delta V \rangle = 0.2$; $p_{\text{ave}} < 0.0001$, t-test). In fact, the juvenile song after inactivation was significantly less variable than songs of adult zebra finches singing undirected song (i.e., songs not directed to a female; Figure 2D; $p < 0.001$, t-test). LMAN inactivation also eliminated 75% of the difference in mean variability between juvenile song and adult directed song—the most highly stereotyped form of song [19].

To verify that the loss of variability resulted from silencing LMAN neurons, and not from inactivating fibers of passage near LMAN, a $GABA_A$ receptor agonist (muscimol, 30 nl, 25 mM) was injected bilaterally into LMAN ($n = 2$ birds; 66 and 70 dph). Again, all syllables showed a dramatic reduction in variability after injection ($n = 8$ syllables; $\langle V \rangle_{\text{before}} = 0.43$, $\langle \Delta V \rangle = 0.16$; $p_{\text{ave}} < 0.0001$, t-test). While the reduction in acoustic variability was similar to that resulting from TTX injections (Figure 2B), the duration of the effect of muscimol was substantially shorter than observed for TTX (Figure 2C). This difference in temporal profile was in good agreement with the known in vivo pharmacology of TTX and muscimol [20,21], suggesting a direct link between suppression of spiking activity in LMAN and loss of song variability.

An additional effect of LMAN inactivation was a significant reduction in sequence variability, a measure of the variability in syllable ordering (Figure 2E; $p < 0.005$, paired t-test; see Materials and Methods). In fact, the sequential ordering of syllables after TTX injection was comparable in stereotypy to that of adult song. Thus, LMAN activity may influence sequence generation, possibly through an indirect feedback pathway going from RA to HVC, the putative sequence generator [6,8,22].

We confirmed that the loss of song variability following injections into LMAN did not result from diffusion of the drugs into the medial magnocellular nucleus of the nidopallium (MMAN), a nucleus approximately 1.25 mm medial from LMAN with projections to HVC. Bilateral injections of TTX into MMAN, done in the same birds in which LMAN injections were previously made, had no significant effect on acoustic variability (Figure 2B).

We next considered the neural mechanisms by which LMAN affects variability in the motor pathway. One intriguing possibility is that song variability is driven by fast synaptic input from LMAN. If true, then acoustic variability should be accompanied by variability in the firing patterns of RA-

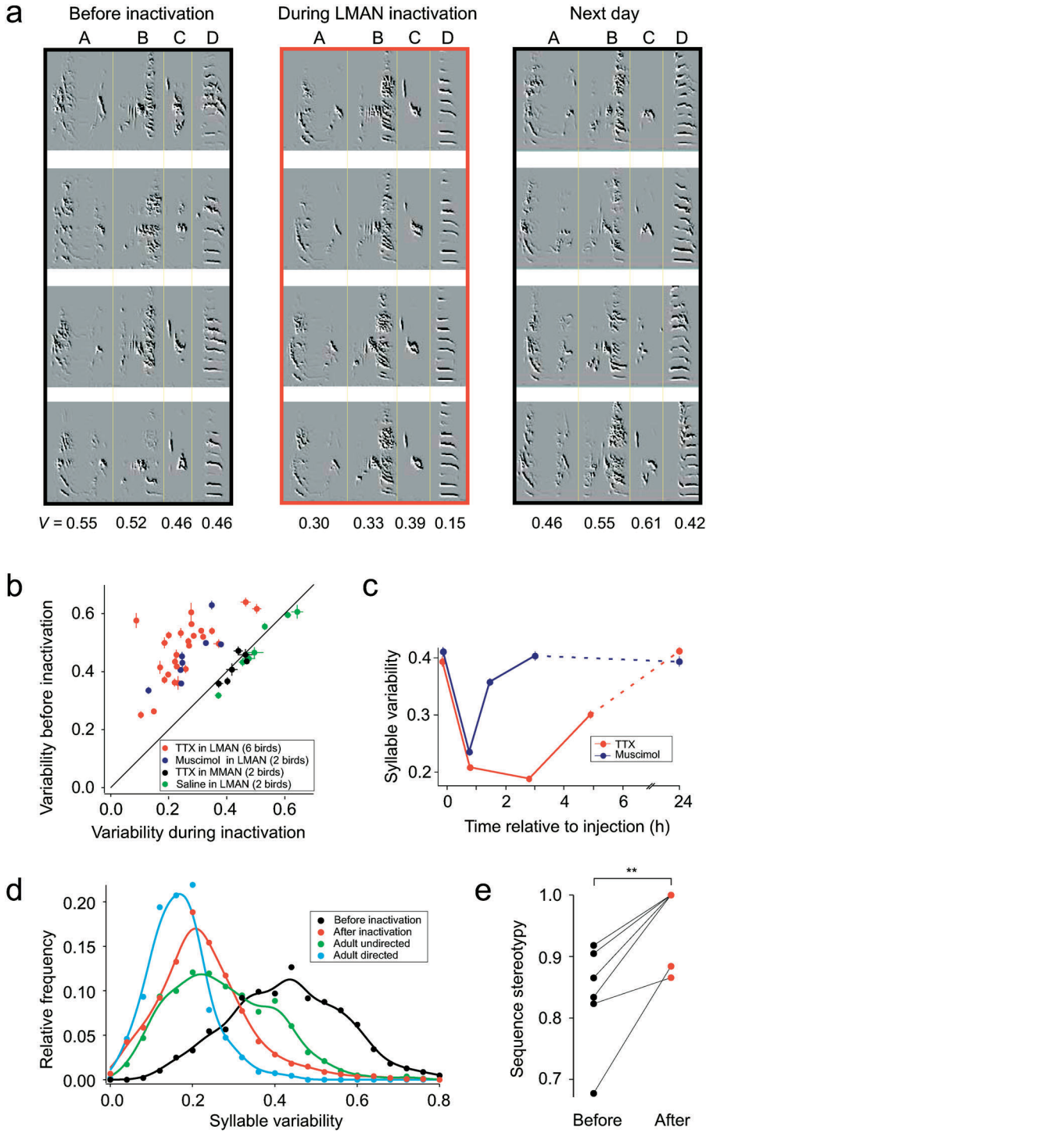


Figure 2. Analysis of the Effect of Bilateral LMAN Inactivation on Song Variability

(A) Consecutive renditions of a repeating song motif of 0.5 s duration in a juvenile bird (59 dph) arranged vertically. Note the large variations in acoustic structure within individual syllables before LMAN inactivation (left). Following TTX injection into LMAN, the acoustic variability is dramatically reduced (middle), only to return to the original level by the following day (right). Numbers below each column indicate the variability index (See Material and Methods section) calculated for the four renditions of the syllables shown.

(B) Scatter plot of variability scores before and during LMAN inactivation with TTX (red) and muscimol (blue). Also shown are results for bilateral TTX injection into MMAN (black; see text), and saline injection into LMAN (green).

(C) Time course of variability reduction following TTX (red) and muscimol (blue) injections show a time dependence that reflects the known in vivo pharmacology of the respective agents. Data were averaged over four identified syllables and taken from the same bird over consecutive days (dph = 70 and 71; muscimol inactivation followed by TTX inactivation).

(D) Distribution of variability scores for all syllables analyzed in the TTX and muscimol experiments (25 unique syllables, six birds) before (black) and during (red) LMAN inactivation in juvenile birds. Shown for comparison are the variability scores for adult zebra finch syllables (18 syllables, 4 birds; undirected song, green; directed song, light blue). Dots represent raw data, while the lines are smoothed running averages.

(E) TTX inactivation of LMAN significantly increased syllable sequence stereotypy. Sequence stereotypy scores (see Materials and Methods) for six birds before (black) and after (red) TTX injections into LMAN. For comparison, the average stereotypy score for adult birds singing directed song was 0.95 (n = 4 birds).

DOI: 10.1371/journal.pbio.0030153.g002

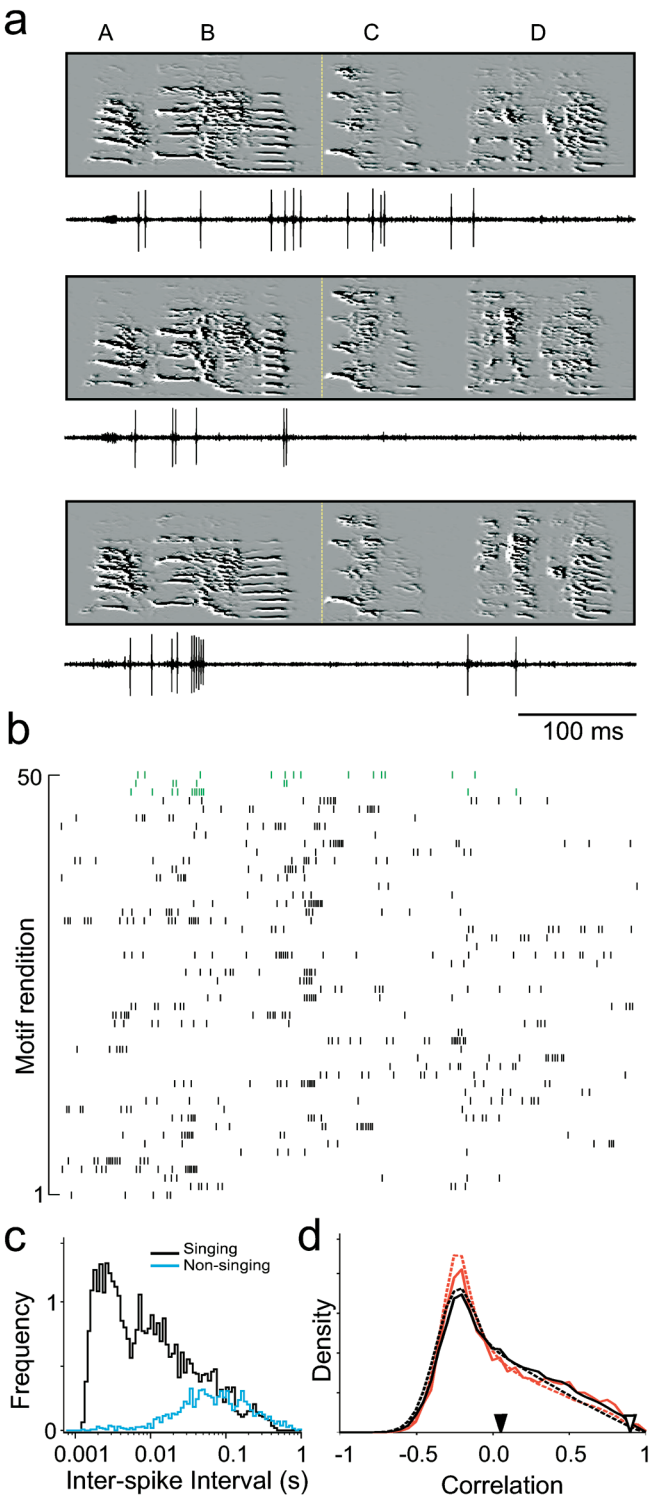


Figure 3. Song-Aligned Firing Patterns of RA-Projecting LMAN Neurons in Singing Juvenile Zebra Finches Are Highly Variable

(A) Three successive renditions of a 67-d-old bird's song motif. Displayed under each spectrogram is the simultaneously recorded voltage waveform of an antidromically identified RA-projecting LMAN neuron (verified by collision testing). Average syllable variability for the three motifs is 0.31. Motif alignment was done at the onset (yellow lines) of syllable C.

(B) Raster plot showing the spike patterns for 50 consecutive motif renditions for the same cell as in (A). The motifs from (A) are indicated in green.

(C) Relative frequency of inter-spike intervals during singing (black) and non-singing (blue) for all the 17 identified projection neurons (units are intervals per second; bin size is 0.04 log units).

(D) Distribution of spike-train correlations across all pairs of motifs for the cell in (B) (solid red line). Correlations calculated with random time shifts added to the spike trains have a similar distribution (dashed red line; see Materials and Methods). Also shown is the correlation distribution for the population of identified projection neurons (solid black line; mean correlation indicated by solid arrowhead), and for the population with random time shifts added (dashed black line). In comparison, spike trains of neurons in premotor nucleus RA of the adult bird are highly stereotyped (from [23]; mean correlation indicated by open arrowhead).

DOI: 10.1371/journal.pbio.0030153.g003

projecting LMAN neurons. To test this idea explicitly, we recorded single-unit signals from 29 LMAN neurons in singing juvenile birds (n = 3 birds; age range, 62–79 dph) (Figure 3). In all, 17 of these were antidromically identified as RA-projecting LMAN neurons (see Materials and Methods). These neurons exhibited song-related changes in firing rate (spontaneous activity, 12 ± 4 Hz; during singing, 39 ± 6 Hz [mean ± standard deviation]), and generated significantly more bursts during singing (Figure 3C). Raster plots of the spike trains aligned to the song motif showed that the patterns of spikes and bursts generated by individual neurons were different each time the bird sang (Figure 3A and 3B).

Correlations in the spike trains across different renditions of the motif were small (0.054 ± 0.34 [mean ± standard deviation]) compared to those observed in premotor neurons of adult birds (0.90 ± 0.1) [7]. We also compared the correlation distributions to those calculated after random time shifts were added to the spike trains (see Materials and Methods). In general, the correlation distributions of the randomized spike trains were very similar to those calculated for the motif-aligned spike trains (Figure 3D), confirming that the firing patterns of LMAN neurons are highly variable. Nevertheless, in 13 out of the 17 identified RA-projecting neurons the correlation distributions were still significantly different from those of the randomly shuffled spike trains ($p < 0.01$, Kolmogorov-Smirnov test), suggesting that while LMAN activity is highly variable, it is not completely random with respect to the song.

Guided by the neural data, we next tested the hypothesis that LMAN drives song variability by providing excitatory glutamatergic input to RA—which in the zebra finch is mediated almost exclusively by N-methyl-D-aspartate (NMDA)–type receptors [24]. In contrast, glutamatergic inputs to RA from HVC are mediated by a mixture of NMDA and α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)–type receptors (Figure 4A) [25]. Thus, if LMAN drives song variability through glutamatergic input to RA, then blocking NMDA receptors should reduce this variability, while sparing the AMPA-mediated drive from HVC. In line with our hypothesis, bilateral injections of the NMDA receptor antagonist 2-amino-5-phosphonovalerate (AP5, 50 nl, 30 mM) into RA significantly reduced acoustic variability in all song syllables examined (Figure 4B and 4C; n = 4 birds; age range, 57–73 dph; 11 syllables; $\langle V \rangle_{\text{before}} = 0.47$, $\langle \Delta V \rangle = 0.16$; $p_{\text{ave}} < 0.0001$, t-test). The time course of the variability reduction (Figure 4D) was consistent with the temporal profile of AP5 effects seen in other in vivo studies [26].

Given that AP5 has effects beyond blocking LMAN input to RA, it may influence the song in ways other than reducing variability. To examine whether AP5 injections affected the

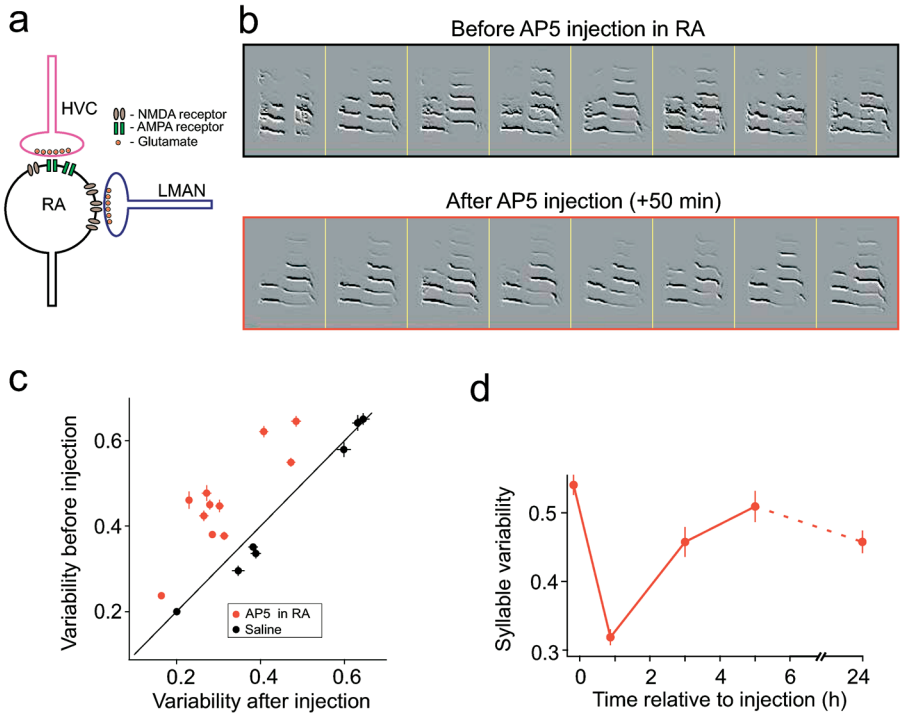


Figure 4. Bilateral Injections of the NMDA Receptor Antagonist AP5 into RA Significantly Reduced Song Variability

(A) Excitatory synaptic inputs to RA from LMAN and HVC are mediated by a different mix of glutamate receptor types (see text). Using AP5 we could block LMAN input while only partially inactivating HVC input.

(B) Eight sequential renditions of one song syllable in a juvenile zebra finch (63 dph) before and after AP5 injection into RA. Note the rapid fluctuations in pitch, the appearance of noisy acoustic structure, and variations in syllable duration before injection. The average variability scores *(V)* before and after injections for the eight shown syllable renditions were 0.50 and 0.25, respectively.

(C) Following injection of AP5 into RA, fluctuations in acoustic structure were substantially reduced. Variability scores of 11 syllables in four birds before and after injection of AP5 into RA.

(D) Time course of acoustic variability following drug injection averaged over all identifiable syllables for the bird in (B).

DOI: 10.1371/journal.pbio.0030153.g004

acoustic structure of syllables, we compared the acoustic features of syllables after AP5 injection to the same syllables before injection (average similarity score 78.0, 11 syllables; see Materials and Methods). In comparison, the average similarity score across renditions of the same syllables prior to injection was 77.7, suggesting that the effect of AP5 injection was largely limited to song variability.

Discussion

Previous studies have shown that permanent LMAN lesions in the juvenile bird disrupt song learning and result in an impoverished and prematurely stereotyped song [12,13]. Such lesions are known to produce synaptic maturation in RA within a few days [27], perhaps because of a loss of neurotrophic input from LMAN [12,13]. Because of the long delay from lesioning to singing (often several days), these studies could not address whether increased stereotypy was caused by synaptic reorganization in RA, or by a more immediate mechanism such as the loss of fast synaptic input from LMAN. In our experiments, we observe singing within an hour after injection, and find that LMAN inactivation reduces song variability reversibly and on a short timescale. This observation implies that, in addition to slow neurotrophic effects, LMAN acts on RA rapidly to drive or control song variability, a necessary ingredient of reinforcement learning. Thus, our results suggest that the loss of vocal plasticity following permanent lesions of LMAN may, in part at least, be due to the immediate loss of exploratory behavior.

What is the mechanism by which neural activity in LMAN controls motif-to-motif variability in the song? Our experiments tested the hypothesis that fluctuations in the song are driven directly by synaptic input from LMAN [25]. In this view, the premotor circuit generates a stereotyped song sequence upon which the AFP acts to drive variations. This hypothesis requires that neural activity in LMAN be highly variable across different song motifs, a prediction that was borne out by our recordings in LMAN (see Figure 3). In comparison, premotor neurons in adult birds (singing song of comparable stereotypy to our LMAN-inactivated juvenile birds) generate extremely stereotyped, song-locked spike patterns [6,7,8]. In itself, the result that LMAN neurons are only weakly time-locked to the song may not be surprising. The significance of this observation becomes apparent when considering that these neurons send excitatory projections to the motor pathway, and that they are necessary for the expression of song variability as demonstrated by our inactivation results. Together with the finding that electrical stimulation of LMAN in adult birds can drive transient changes in the song [19], these observations make LMAN a likely source for the variability in the premotor pathway.

Because LMAN input to RA neurons is mediated almost exclusively by NMDA receptors, another strong prediction of

our hypothesis was that blockade of NMDA receptors in RA should reduce song variability. Our results from the injection of AP5 into RA confirmed this. However, given the presence of NMDA receptors in the projection from HVC to RA [24], and perhaps in recurrent connections within RA, blockade of NMDA receptors is likely to have effects on RA circuitry beyond the loss of direct synaptic input from LMAN. Thus, this experiment cannot preclude other hypotheses—for example, that LMAN acts to regulate stochastic processes intrinsic to the premotor circuit, through some yet unknown mechanism.

Further support for the idea that LMAN can drive song variations comes from studies in the adult zebra finch. Song-related neural activity in LMAN is variable also in the adult bird, and this variability has been shown to be larger during undirected as compared to directed singing [27,28]. A recent study [19] linked the increased neural variability in LMAN during undirected singing to an increase in motif-to-motif variability in song features (see also Figure 2D).

How does the role and function of LMAN change as song variability is reduced during learning and finally during song crystallization? To the extent that the variability of LMAN firing patterns in the adult bird during undirected song [28] is similar to that in the juvenile bird, an essential part of song development may be a reduction of the gain by which LMAN drives RA. This could occur as a result of synaptic changes within RA that weaken input from LMAN and/or strengthen the projections from HVC. While there is evidence that this may indeed occur [26,29], more experiments are needed to establish how the developmental reduction in song variability is related to changes in song circuitry.

Reinforcement learning requires that variability in the motor output be accompanied by a mechanism that evaluates the resulting performance. In the songbird, such an evaluation signal could be sent directly to the motor system (e.g., to RA), perhaps via a neuromodulator [30,31], to reinforce the states of the motor pathway that lead to a better-than-expected match to the memorized template. A reinforcement signal could also be sent to the AFP to shape or regulate the fluctuations introduced into the motor pathway via LMAN. This would make LMAN more than a simple "noise generator," allowing it to bias vocal fluctuations in the direction of the desired song. Such bias is suggested by the presence of small but significant correlations in the motif-aligned firing pattern of LMAN neurons (see Figure 3). This bias could permit a more efficient exploration of motor space, and even allow LMAN activity to drive plastic changes in the motor circuitry.

The exploratory motor behavior exhibited by juvenile songbirds may also provide general insights into how the brain generates fluctuations required for learning. Such fluctuations could be generated within the motor pathway or by brain regions projecting to it, and could result from stochastic processes, such as randomness in synaptic release [32], noise propagated by summation of irregular patterns of inhibitory postsynaptic potentials and excitatory postsynaptic potentials [33], or complex collective dynamics of the neuronal network [34]. Our results strongly suggest that, whatever the detailed biophysical mechanisms, the neural circuits generating these fluctuations are located outside the motor pathway in a specialized pathway involving the basal ganglia. The output of this circuit acts on the motor pathway, allowing the song system to explore the vocal space in a purposeful manner. Whether inducing exploratory motor behavior is a general feature of basal ganglia circuits is an intriguing idea that remains to be explored.

Materials and Methods

Subjects. Subjects were juvenile male zebra finches (54–79 dph). Birds were obtained from the Massachusetts Institute of Technology zebra finch breeding facility (Cambridge, Massachusetts), and from the aviary at the Rockefeller Field Research Station (Millbrook, New York). The care and experimental manipulation of the animals were carried out in accordance with guidelines of the National Institutes of Health and were reviewed and approved by the Massachusetts Institute of Technology Institutional Animal Care and Use Committee.

Reversible inactivation. Birds underwent a brief surgery to attach to the skull a means of restraining the head during drug injections. The animals were anesthetized with isoflurane (2%) and placed in a stereotaxic apparatus (MyNeuroLab.com, St. Louis, Missouri, United States). Two stainless-steel screws (#0–80 6 mm long) were secured to the skull with dental acrylic. Small holes (approximately 300 μm in diameter) were drilled through the cranium bilaterally over LMAN or MMAN, or RA using stereotaxic coordinates. The holes were covered with a thin layer of Kwik-Kast (World Precision Instruments, Sarasota, Florida, United States). The animals were then placed in a custom sound-isolation chamber where they began to sing prolifically after a few days—typically 200–1,000 song motifs per hour.

Inactivation of song control nuclei in the singing bird was carried out by placing the bird, unanesthetized, in a small foam restraint and attaching the head-mounted screws to a metal plate bolted to the stereotaxic apparatus. The Kwik-Kast over the cranial holes was removed, and 30 nl of TTX (50 μM, #T5651, Sigma, St. Louis, Missouri, United States) or muscimol (25 mM, #M1523, Sigma) was injected bilaterally into the brain region of interest using a Nanoject II injector (Drummond Scientific, Broomall, Pennsylvania, United States). The procedure of injecting the birds took approximately 10 min. Experimental confirmation of the physiological effects of TTX injections showed that LMAN was likely completely inactivated after our injections (see Figure S2). Regions immediately surrounding LMAN were also affected, and we cannot rule out an indirect contribution from the partial inactivation of these regions. For inactivation of NMDA-mediated synapses in RA, AP5 (#A5282, Sigma) was injected bilaterally into RA (50 nl, 30 mM). The injection site was guided by electrophysiological recordings of spontaneous activity in RA.

Injected solutions also contained dye-conjugated dextrans (#D22912, Molecular Probes, Eugene, Oregon, United States). All injection sites were verified by histological examination and were found to be within the target nucleus (see Figure S1), except for TTX injections in LMAN in two birds: one in which the LMAN injection site in one hemisphere was found to be approximately 100 μm anterior to the edge of LMAN, the other in which the injections were approximately 200 μm posterior to LMAN, but right in the middle of the fiber tract leading from LMAN to RA. The results from these birds were similar to those from other birds, and were included in the analysis.

Chronic neural recordings in LMAN. Experiments were timed such that the birds were at an age at which they produced readily identifiable syllable sequences, yet showed variable acoustic syllable structure across song renditions. Recordings were carried out using a motorized microdrive described previously [35]. Cells were isolated by searching for spontaneous or antidromically evoked spiking activity; units typically had signal-to-noise ratios greater than 10:1. Antidromic identification of RA-projecting LMAN neurons was carried out with a bipolar stimulating electrode implanted in RA using techniques described previously for antidromic identification of RA-projecting HVC neurons [8]. Neurons exhibiting a short-latency antidromic spike ($<$5 ms) with a root-mean-squared latency jitter of less than 100 μs (at a stimulation current of approximately 10% above threshold) were counted as identified RA-projecting neurons. Of the 17 antidromically identified neurons in our dataset, ten were further validated with collision tests [8]. An additional ten putative projection neurons did not respond to RA stimulation with a short-latency spike, but exhibited spike patterns and correlations similar to the identified projection neurons. For the cells in our dataset, we recorded signals for many song motifs (range, 5–133 motifs; mean, 56).

Data analysis. To assess the effects of drug injections on acoustic variability and average acoustic structure, analysis was done on reliably identifiable song syllables (range, 2–5 per bird; see Figure 2A

for an example). Each data point was derived from 45 pairwise comparisons made across ten consecutive renditions of a given syllable, recorded immediately before and after injection. Acoustic variability was quantified using the Sound Analysis Pro 1.04 software [36], and pairwise comparisons of the acoustic features of identified syllables were made using the local similarity measure ("accuracy"). This measure is based on pitch, frequency modulation, amplitude modulation, Wiener entropy, and goodness of pitch, and is calculated in 9-ms intervals and averaged over the duration of the syllable; syllables were aligned in time so as to maximize the similarity, allowing for 5% time warping. For the variability measurements, the resulting similarity score (S, ranging from zero to 100) was converted, through a linear remapping, to a variability score *(V)* by the following formula:

$$V = \frac{S_{\max} - \langle S \rangle}{S_{\max} - \langle S_{\min} \rangle}. \quad (1)$$

$\langle S_{\min} \rangle$ is the average similarity score of randomly chosen pairs of syllables from unrelated birds, which in our finch colony was measured to be 50 ± 12 (mean ± standard deviation, $n = 200$ pairwise comparisons; comparisons were made across syllables of birds from different fathers). The similarity of identical syllables, $S_{\max}$, is 100 by definition of the similarity measure. Thus, a variability score of one means that syllables are as different as two unrelated syllables would be on average, while variability score of zero means that the syllables are identical. Error bars for V in the figures all denote standard error of the mean. $\langle V \rangle$ denotes the average variability score across birds and syllables for a given condition.

The variability of syllable ordering in a song was quantified using the stereotypy score of Scharff and Nottebohm [13], excluding the variability in the number of introductory notes and in the end syllable of a song bout. The score is a combination of "sequence linearity," which addresses the way in which notes are ordered, and "sequence consistency," a measure of the frequency with which the main motif sequence appears. Complete stereotypy yields a score of one, while a completely random sequencing will have a score close to zero. Stereotypy scores were calculated over ten consecutive song bouts, before and after LMAN injections.

For the analysis of the neural recordings in LMAN, we determined the sequence of song syllables most frequently produced by each bird. Motifs that matched this sequence were identified and time-aligned using the onset of one of the syllables. The alignment syllable was chosen for a sharp onset in acoustic power. The relative jitter in the timing of other syllables in the motif was found to be less than 9 ms (root mean squared). Spike times were extracted, and the instantaneous firing rate during each motif rendition was estimated by smoothing the spike train with a Gaussian of half-width 20 ms (to the 1/e points). Correlations were calculated between the firing rate functions for all pairs of smoothed spike trains. Correlations were also calculated for all pairs of spike trains after a random time shift. The shift was circular, such that spikes wrapped around to the beginning of the motif; time shifts were chosen randomly from a uniform distribution with the width of the motif. For each cell the correlation distribution of the time-shifted firing rates was calculated with 100 different ensembles of random shifts. This random shift ensured zero mean correlation while preserving spike statistics. Thus, the distribution of time-shifted correlations provides a zero-correlation baseline with which to compare our results.

Supporting Information

Figure S1. Histology Confirming the Injection Sites for the LMAN Inactivation Experiments in Figures 1 and 2.

(A) A parasaggital Nissl-stained section of a zebra finch brain showing the location of LMAN.
(B) Inverted darkfield image showing LMAN in one of the juveniles injected (red markers in [D] and [E]).
(C) Combined darkfield and fluorescence image showing the spread of the dye that was co-injected with the drug.
(D and E) Estimated injection sites relative to the boundaries of LMAN for all birds in Figures 1 and 2 in the saggital (D) and coronal (E) planes, respectively (individual birds are color coded).
(F) Estimated maximum diameter of LMAN in the saggital plane.
(G) Estimated lateral extent of LMAN in the coronal plane.
The estimates in (F) and (G) are based on the contrast borders seen in the darkfield images (see [B]). Note that fibers from LMAN to RA leave the posterior edge of LMAN.

Found at DOI: 10.1371/journal.pbio.0030153.sg001 (369 KB PDF).

Figure S2. Dose- and Distance-Dependent Effects of TTX Injections in and around LMAN

(A) Decrease in acoustic variability (ΔV) approximately 1 h after injection, as a function of location and concentration of TTX injections. Red bars indicate dose response for TTX injections in LMAN ($n = 2$ birds; 8 syllables; injection sites for the two birds correspond to the blue and grey markers in Figure S1). Blue bars indicate 30-nl saline injections in LMAN ($n = 2$ birds; 7 syllables). Green bars indicate 30-nl (50 μM) TTX injections 1.25 mm medial (MMAN, $n = 2$ birds; 6 syllables) and dorsal ("above," $n = 2$; 8 syllables) from the center of LMAN.
(B and C) Summary of experiments done to verify the physiological spread of TTX. Experiments were done in anesthetized birds (2% isoflurane). A bipolar stimulating electrode was placed in RA, and a recording electrode in LMAN, producing antidromically evoked activity in LMAN (stimulus pulses, 175 μA, 0.2 ms, 0.5 Hz). TTX (30 nl, 50 μM) was injected at different distances away from the recording electrode. (B) Examples of recorded signals for TTX injections 400 μm (top) and 1,250 μm (bottom) away from the recording electrode (averaged over 30 stimulus pulses). The baseline stimulus artifact recorded 1 mm above LMAN is shown in the green boxes (left). Signal recorded in LMAN immediately before injection is shown in the black boxes (middle). Signal recorded 1 h after injection is shown in the red boxes (right). (C) Summary of evoked activity 1 h after TTX injections made at different distances away from the recording site. Evoked activity was measured as the root-mean-squared deviation of the signal from the baseline in the interval 1.5–4.5 ms after the stimulation pulse (six birds, two at 400 μm, two at 600 μm, and one each at 800 μm and 1,250 μm).

Found at DOI: 10.1371/journal.pbio.0030153.sg002 (1.1 MB PDF).

Figure S3. Example of a Juvenile Zebra Finch Song (54 dph) Showing a Loss of Sequence and Acoustic Variability following LMAN Inactivation by TTX Injection

The song snippets shown are from three consecutive song bouts, immediately before and 1 h after TTX injection. Tutor song is shown for comparison.

Found at DOI: 10.1371/journal.pbio.0030153.sg003 (1.8 MB PDF).

Audio S1. Example of a Song from the Bird in Figure 1 prior to TTX Inactivation of LMAN (Bout 1)

Found at DOI: 10.1371/journal.pbio.0030153.sa001 (545 KB WAV).

Audio S2. Example of a Song from the Bird in Figure 1 prior to TTX Inactivation of LMAN (Bout 2)

Found at DOI: 10.1371/journal.pbio.0030153.sa002 (455 KB WAV).

Audio S3. Example of a Song from the Bird in Figure 1 during TTX Inactivation of LMAN (Bout 1)

Found at DOI: 10.1371/journal.pbio.0030153.sa003 (430 KB WAV).

Audio S4. Example of a Song from the Bird in Figure 1 during TTX Inactivation of LMAN (Bout 2)

Found at DOI: 10.1371/journal.pbio.0030153.sa004 (360 KB WAV).

Acknowledgments

We thank Edward Soucy, Stephen Baccus, Isabella Nebel, Carlos Lois, and members of the Fee lab for comments on the manuscript. We also acknowledge Thomas Ramée for assistance with histology and animal care.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. BPÖ, ASA, and MSF conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, and wrote the paper. ■

References

1. Sutton RS, Barto AG (1998) Reinforcement learning: An introduction. Cambridge: MIT Press. 322 p.
2. Immelmann K (1969) Song development in the zebra finch and other estrilid finches. In: Hinde RA, editor. Bird vocalizations. New York: Cambridge Univ. Press. pp. 61–74.

3. Konishi M (1965) The role of auditory feedback in the control of vocalizations in the white-crowned sparrow. Z Tierpsychol 22: 770–783.
4. Vu ET, Mazurek ME, Kuo Y (1994) Identification of a forebrain motor programming network for the learned song of zebra finches. J Neurosci 14: 6924–6934.
5. Nottebohm F, Kelley DB, Paton JA (1982) Connections of vocal control nuclei in the canary telencephalon. J Comp Neurol 207: 344–357.
6. Yu AC, Margoliash D (1996) Temporal hierarchical control of singing in birds. Science 273: 1871–1875.
7. Leonardo A, Fee MS (2005) Ensemble coding of vocal control in birdsong. J Neurosci 25: 652–661.
8. Hahnloser RHR, Kozhevnikov AA, Fee MS (2002) An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature 419: 65–70.
9. Luo M, Perkel DJ (2001) An avian basal ganglia pathway essential for vocal learning forms a closed topographic loop. J Neurosci 21: 6836–6845.
10. Farries MA, Perkel DJ (2002) A telencephalic nucleus essential for song learning contains neurons with physiological characteristics of both striatum and globus pallidus. J Neurosci 22: 3776–3787.
11. Canales JJ, Graybiel AM (2000) A measure of striatal function predicts motor stereotypy. Nat Neurosci 3: 377–383.
12. Bottjer SW, Miesner EA, Arnold AP (1984) Forebrain lesions disrupt development but not maintenance of song in passerine birds. Science 224: 901–903.
13. Scharff C, Nottebohm F (1991) A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: Implications for vocal learning. J Neurosci 11: 2896–2913.
14. Margoliash D (2002) Evaluating theories of bird song learning: Implications for future directions. J Comp Physiol A 188: 851–866.
15. Troyer TW, Bottjer SW (2001) Birdsong: Models and mechanisms. Curr Opin Neurobiol 11: 721–726.
16. Hessler N, Doupe A (1999) Singing-related neural activity in a dorsal forebrain—Basal ganglia circuit of adult zebra finches. J Neurosci 19: 10461–10481.
17. Leonardo A (2004) Experimental test of the birdsong error-correction model. Proc Nat Acad Sci U S A 101: 16935–16940.
18. Doya K, Sejnowski TJ (1995) A novel reinforcement model of birdsong vocalization learning. In: Tesauro G, Touretzky DS, Leen TK, editors. Advances in neural information processing systems, Volume 7. Cambridge: MIT Press. pp. 101–108.
19. Kao MH, Doupe AJ, Brainard MS (2005) Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. Nature 433: 638–643.
20. Boehnke SE, Rasmusson DD (2001) Time course and effective spread of lidocaine and tetrodotoxin delivered via microdialys: An electrophysiological study in cerebral cortex. J Neurosci Meth 105: 133–141.
21. Martin JH, Ghez C (1999) Pharmacological inactivation in the analysis of the central control of movement. J Neurosci Meth 86: 145–159.
22. Schmidt M, Ashmore RC, Vu ET (2004) Bilateral control and interhemispheric coordination in the avian song motor system. In: Zeigler HP, Marler P, editors. Behavioral neurobiology of birdsong. New York: New York Academy of Science. pp. 171–186.
23. Mooney R, Konishi M (1991) Two distinct inputs to an avian song nucleus activate different glutamate receptor subtypes on individual neurons. Proc Natl Acad Sci U S A 88: 4075–4079.
24. Stark LL, Perkel DJ (1999) Two-stage, input-specific synaptic maturation in a nucleus essential for vocal production in the zebra finch. J Neurosci 19: 9107–9116.
25. Steele RJ, Morris RGM (1999) Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. Hippocampus 9: 118–138.
26. Kittelberger J, Mooney R (1999) Lesions of an avian forebrain nucleus that disrupt song development alter synaptic connectivity and transmission in the vocal premotor pathway. J Neurosci 19: 9385–9398.
27. Hessler N, Doupe A (1999) Social context modulates singing-related neural activity in the songbird forebrain. Nat Neurosci 2: 209–211.
28. Leonardo A (2002) Neural dynamics underlying complex behavior in a songbird [dissertation]. Pasadena: California Institute of Technology. 97 p. Available: http://etd.caltech.edu/etd/available/etd-05092002–165316/. Accessed 7 March 2005.
29. Herrmann K, Arnold AP (1991) The Development of Afferent Projections to the Robust Archistriatal Nucleus in Male Zebra Finches: A Quantitative Electron Microscopic Study. J Neurosci 11: 2063–2074.
30. Appeltants D, Ball GF, Balthazart J (2002) The origin of catecholaminergic inputs to the song control nucleus RA in canaries. Neuroreport 13: 649–653.
31. Schultz W (2002) Getting formal with dopamine and reward. Neuron 36: 241–263.
32. Seung H (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron 40: 1063–1073.
33. Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. J Neurosci 18: 3870–3896.
34. Kenet T, Bibitchkov KT, Tsodyks M, Grinvald A, Arieli A (2003) Spontaneously emerging cortical representations of visual attributes. Nature 425: 954–956.
35. Fee MS, Leonardo A (2001) Miniature motorized microdrive and commutator system for chronic neural recordings in small animals. J Neurosci Meth 112: 83–94.
36. Tchernichovski O, Nottebohm F, Ho CE, Pesaran B, Mitra PP (2000) A procedure for an automated measurement of song similarity. Anim Behav 59: 1167–1176.

1.3 Elements of Reinforcement Learning

Beyond the agent and the environment, one can identify four main subelements of a reinforcement learning system: a *policy*, a *reward function*, a *value function*, and, optionally, a *model* of the environment.

A *policy* defines the learning agent's way of behaving at a given time. Roughly speaking, a policy is a mapping from perceived states of the environment to actions to be taken when in those states. It corresponds to what in psychology would be called a set of stimulus–response rules or associations. In some cases the policy may be a simple function or lookup table, whereas in others it may involve extensive computation such as a search process. The policy is the core of a reinforcement learning agent in the sense that it alone is sufficient to determine behavior. In general, policies may be stochastic.

A *reward function* defines the goal in a reinforcement learning problem. Roughly speaking, it maps each perceived state (or state–action pair) of the environment to a single number, a *reward*, indicating the intrinsic desirability of that state. A reinforcement learning agent's sole objective is to maximize the total reward it receives in the long run. The reward function defines what are the good and bad events for the agent. In a biological system, it would not be inappropriate to identify rewards with pleasure and pain. They are the immediate and defining features of the problem

faced by the agent. As such, the reward function must necessarily be unalterable by the agent. It may, however, serve as a basis for altering the policy. For example, if an action selected by the policy is followed by low reward, then the policy may be changed to select some other action in that situation in the future. In general, reward functions may be stochastic.

Whereas a reward function indicates what is good in an immediate sense, a *value function* specifies what is good in the long run. Roughly speaking, the *value* of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state. Whereas rewards determine the immediate, intrinsic desirability of environmental states, values indicate the *long-term* desirability of states after taking into account the states that are likely to follow, and the rewards available in those states. For example, a state might always yield a low immediate reward but still have a high value because it is regularly followed by other states that yield high rewards. Or the reverse could be true. To make a human analogy, rewards are like pleasure (if high) and pain (if low), whereas values correspond to a more refined and farsighted judgment of how pleased or displeased we are that our environment is in a particular state. Expressed this way, we hope it is clear that value functions formalize a basic and familiar idea.

Rewards are in a sense primary, whereas values, as predictions of rewards, are secondary. Without rewards there could be no values, and the only purpose of estimating values is to achieve more reward. Nevertheless, it is values with which we are most concerned when making and evaluating decisions. Action choices are made based on value judgments. We seek actions that bring about states of highest value, not highest reward, because these actions obtain the greatest amount of reward for us over the long run. In decision-making and planning, the derived quantity called value is the one with which we are most concerned. Unfortunately, it is much harder to determine values than it is to determine rewards. Rewards are basically given directly by the environment, but values must be estimated and reestimated from the sequences of observations an agent makes over its entire lifetime. In fact, the most important component of almost all reinforcement learning algorithms is a method for efficiently estimating values. The central role of value estimation is arguably the most important thing we have learned about reinforcement learning over the last few decades.

Although all the reinforcement learning methods we consider in this book are structured around estimating value functions, it is not strictly necessary to do this to solve reinforcement learning problems. For example, search methods such as genetic algorithms, genetic programming, simulated annealing, and other function optimization methods have been used to solve reinforcement learning problems. These

methods search directly in the space of policies without ever appealing to value functions. We call these *evolutionary* methods because their operation is analogous to the way biological evolution produces organisms with skilled behavior even when they do not learn during their individual lifetimes. If the space of policies is sufficiently small, or can be structured so that good policies are common or easy to find, then evolutionary methods can be effective. In addition, evolutionary methods have advantages on problems in which the learning agent cannot accurately sense the state of its environment.

Nevertheless, what we mean by reinforcement learning involves learning while interacting with the environment, which evolutionary methods do not do. It is our belief that methods able to take advantage of the details of individual behavioral interactions can be much more efficient than evolutionary methods in many cases. Evolutionary methods ignore much of the useful structure of the reinforcement learning problem: they do not use the fact that the policy they are searching for is a function from states to actions; they do not notice which states an individual passes through during its lifetime, or which actions it selects. In some cases this information can be misleading (e.g., when states are misperceived), but more often it should enable more efficient search. Although evolution and learning share many features and can naturally work together, as they do in nature, we do not consider evolutionary methods by themselves to be especially well suited to reinforcement learning problems. For simplicity, in this book when we use the term "reinforcement learning" we do not include evolutionary methods.

The fourth and final element of some reinforcement learning systems is a *model* of the environment. This is something that mimics the behavior of the environment. For example, given a state and action, the model might predict the resultant next state and next reward. Models are used for *planning*, by which we mean any way of deciding on a course of action by considering possible future situations before they are actually experienced. The incorporation of models and planning into reinforcement learning systems is a relatively new development. Early reinforcement learning systems were explicitly trial-and-error learners; what they did was viewed as almost the *opposite* of planning. Nevertheless, it gradually became clear that reinforcement learning methods are closely related to dynamic programming methods, which do use models, and that they in turn are closely related to state-space planning methods. In Chapter 9 we explore reinforcement learning systems that simultaneously learn by trial and error, learn a model of the environment, and use the model for planning. Modern reinforcement learning spans the spectrum from low-level, trial-and-error learning to high-level, deliberative planning.

2.1 An n-Armed Bandit Problem

Consider the following learning problem. You are faced repeatedly with a choice among n different options, or actions. After each choice you receive a numerical reward chosen from a stationary probability distribution that depends on the action you selected. Your objective is to maximize the expected total reward over some time period, for example, over 1000 action selections. Each action selection is called a *play*.

This is the original form of the *n-armed bandit problem*, so named by analogy to a slot machine, or "one-armed bandit," except that it has n levers instead of one. Each action selection is like a play of one of the slot machine's levers, and the rewards are the payoffs for hitting the jackpot. Through repeated plays you are to maximize your winnings by concentrating your plays on the best levers. Another analogy is that of a doctor choosing between experimental treatments for a series of seriously ill patients. Each play is a treatment selection, and each reward is the survival or well-being of the patient. Today the term "n-armed bandit problem" is often used for a generalization of the problem described above, but in this book we use it to refer just to this simple case.

In our n-armed bandit problem, each action has an expected or mean reward given that that action is selected; let us call this the *value* of that action. If you knew the value of each action, then it would be trivial to solve the n-armed bandit problem: you would always select the action with highest value. We assume that you do not know the action values with certainty, although you may have estimates.

If you maintain estimates of the action values, then at any time there is at least one action whose estimated value is greatest. We call this a *greedy* action. If you select a greedy action, we say that you are *exploiting* your current knowledge of the values of the actions. If instead you select one of the nongreedy actions, then we say you are *exploring* because this enables you to improve your estimate of the nongreedy action's value. Exploitation is the right thing to do to maximize the expected reward on the one play, but exploration may produce the greater total reward in the long run. For example, suppose the greedy action's value is known with certainty, while several other actions are estimated to be nearly as good but with substantial uncertainty. The uncertainty is such that at least one of these other actions probably is actually better than the greedy action, but you don't know which one. If you have many plays yet to make, then it may be better to explore the nongreedy actions and discover which of them are better than the greedy action. Reward is lower in the short run, during exploration, but higher in the long run because after you have discovered the

better actions, you can exploit *them*. Because it is not possible both to explore and to exploit with any single action selection, one often refers to the "conflict" between exploration and exploitation.

In any specific case, whether it is better to explore or exploit depends in a complex way on the precise values of the estimates, uncertainties, and the number of remaining plays. There are many sophisticated methods for balancing exploration and exploitation for particular mathematical formulations of the n-armed bandit and related problems. However, most of these methods make strong assumptions about stationarity and prior knowledge that are either violated or impossible to verify in applications and in the full reinforcement learning problem that we consider in subsequent chapters. The guarantees of optimality or bounded loss for these methods are of little comfort when the assumptions of their theory do not apply.

In this book we do not worry about balancing exploration and exploitation in a sophisticated way; we worry only about balancing them at all. In this chapter we present several simple balancing methods for the n-armed bandit problem and show that they work much better than methods that always exploit. In addition, we point out that supervised learning methods (or rather the methods closest to supervised learning methods when adapted to this problem) perform poorly on this problem because they do not balance exploration and exploitation at all. The need to balance exploration and exploitation is a distinctive challenge that arises in reinforcement learning; the simplicity of the n-armed bandit problem enables us to show this in a particularly clear form.

2.2 Action-Value Methods

We begin by looking more closely at some simple methods for estimating the values of actions and for using the estimates to make action selection decisions. In this chapter, we denote the true (actual) value of action a as $Q^*(a)$, and the estimated value at the tth play as $Q_t(a)$. Recall that the true value of an action is the mean reward received when that action is selected. One natural way to estimate this is by averaging the rewards actually received when the action was selected. In other words, if at the tth play action a has been chosen k_a times prior to t, yielding rewards $r_1, r_2, \ldots, r_{k_a}$, then its value is estimated to be

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}. \tag{2.1}$$

2.5 Incremental Implementation

The action-value methods we have discussed so far all estimate action values as sample averages of observed rewards. The obvious implementation is to maintain, for each action a, a record of all the rewards that have followed the selection of that action. Then, when the estimate of the value of action a is needed at time t, it can be computed according to (2.1), which we repeat here:

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a},$$

where $r_1, \ldots, r_{k_a}$ are all the rewards received following all selections of action a prior to play t. A problem with this straightforward implementation is that its memory and computational requirements grow over time without bound. That is, each additional reward following a selection of action a requires more memory to store it and results in more computation being required to determine $Q_t(a)$.

As you might suspect, this is not really necessary. It is easy to devise incremental update formulas for computing averages with small, constant computation required to process each new reward. For some action, let Q_k denote the average of its first k rewards (not to be confused with $Q_k(a)$, the average for action a at the kth *play*). Given this average and a $(k+1)$st reward, r_{k+1}, then the average of all $k+1$ rewards can be computed by

$$
\begin{aligned}
Q_{k+1} &= \frac{1}{k+1}\sum_{i=1}^{k+1} r_i \\
&= \frac{1}{k+1}\left(r_{k+1} + \sum_{i=1}^{k} r_i\right) \\
&= \frac{1}{k+1}\left(r_{k+1} + kQ_k + Q_k - Q_k\right) \\
&= \frac{1}{k+1}\left(r_{k+1} + (k+1)Q_k - Q_k\right) \\
&= Q_k + \frac{1}{k+1}\left[r_{k+1} - Q_k\right], \qquad (2.4)
\end{aligned}
$$

which holds even for $k = 0$, obtaining $Q_1 = r_1$ for arbitrary Q_0. This implementation requires memory only for Q_k and k, and only the small computation (2.4) for each new reward.

The update rule (2.4) is of a form that occurs frequently throughout this book. The general form is

$$NewEstimate \leftarrow OldEstimate + StepSize\left[Target - OldEstimate\right]. \qquad (2.5)$$

The expression $\left[Target - OldEstimate\right]$ is an *error* in the estimate. It is reduced by taking a step toward the "Target." The target is presumed to indicate a desirable direction in which to move, though it may be noisy. In the case above, for example, the target is the $(k+1)$st reward.

Note that the step-size parameter ($StepSize$) used in the incremental method described above changes from time step to time step. In processing the kth reward for action a, that method uses a step-size parameter of $\frac{1}{k}$. In this book we denote the step-size parameter by the symbol α or, more generally, by $\alpha_k(a)$. For example, the above incremental implementation of the sample-average method is described by the equation $\alpha_k(a) = \frac{1}{k_a}$. Accordingly, we sometimes use the informal shorthand $\alpha = \frac{1}{k}$ to refer to this case, leaving the action dependence implicit.

Exercise 2.5 Give pseudocode for a complete algorithm for the n-armed bandit problem. Use greedy action selection and incremental computation of action values with $\alpha = \frac{1}{k}$ step-size parameter. Assume a function $bandit(a)$ that takes an action and returns a reward. Use arrays and variables; do not subscript anything by the time index t. Indicate how the action values are initialized and updated after each reward.

3.3 Returns

So far we have been imprecise regarding the objective of learning. We have said that the agent's goal is to maximize the reward it receives in the long run. How might this be formally defined? If the sequence of rewards received after time step t is denoted $r_{t+1}, r_{t+2}, r_{t+3}, \ldots$, then what precise aspect of this sequence do we wish to maximize? In general, we seek to maximize the *expected return*, where the return, R_t, is defined as some specific function of the reward sequence. In the simplest case the return is the sum of the rewards:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T, \tag{3.1}$$

where T is a final time step. This approach makes sense in applications in which there is a natural notion of final time step, that is, when the agent–environment interaction

breaks naturally into subsequences, which we call *episodes*,[5] such as plays of a game, trips through a maze, or any sort of repeated interactions. Each episode ends in a special state called the *terminal state*, followed by a reset to a standard starting state or to a sample from a standard distribution of starting states. Tasks with episodes of this kind are called *episodic tasks*. In episodic tasks we sometimes need to distinguish the set of all nonterminal states, denoted $\mathcal{S}$, from the set of all states plus the terminal state, denoted $\mathcal{S}^+$.

On the other hand, in many cases the agent–environment interaction does not break naturally into identifiable episodes, but goes on continually without limit. For example, this would be the natural way to formulate a continual process-control task, or an application to a robot with a long life span. We call these *continuing tasks*. The return formulation (3.1) is problematic for continuing tasks because the final time step would be $T = \infty$, and the return, which is what we are trying to maximize, could itself easily be infinite. (For example, suppose the agent receives a reward of $+1$ at each time step.) Thus, in this book we usually use a definition of return that is slightly more complex conceptually but much simpler mathematically.

The additional concept that we need is that of *discounting*. According to this approach, the agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses a_t to maximize the expected *discounted return*:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{3.2}$$

where γ is a parameter, $0 \leq \gamma \leq 1$, called the *discount rate*.

The discount rate determines the present value of future rewards: a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately. If $\gamma < 1$, the infinite sum has a finite value as long as the reward sequence $\{r_k\}$ is bounded. If $\gamma = 0$, the agent is "myopic" in being concerned only with maximizing immediate rewards: its objective in this case is to learn how to choose a_t so as to maximize only r_{t+1}. If each of the agent's actions happened to influence only the immediate reward, not future rewards as well, then a myopic agent could maximize (3.2) by separately maximizing each immediate reward. But in general, acting to maximize immediate reward can reduce access to future rewards so that the return may actually be reduced. As γ approaches 1, the objective takes future rewards into account more strongly: the agent becomes more farsighted.

5. Episodes are often called "trials" in the literature.

3.7 Value Functions

Almost all reinforcement learning algorithms are based on estimating *value functions*—functions of states (or of state–action pairs) that estimate *how good* it is for the agent to be in a given state (or how good it is to perform a given action in a given state). The notion of "how good" here is defined in terms of future rewards that can be expected, or, to be precise, in terms of expected return. Of course the

rewards the agent can expect to receive in the future depend on what actions it will take. Accordingly, value functions are defined with respect to particular policies.

Recall that a policy, π, is a mapping from each state, $s \in \mathcal{S}$, and action, $a \in \mathcal{A}(s)$, to the probability $\pi(s,a)$ of taking action a when in state s. Informally, the *value* of a state s under a policy π, denoted $V^\pi(s)$, is the expected return when starting in s and following π thereafter. For MDPs, we can define $V^\pi(s)$ formally as

$$V^\pi(s) = E_\pi\{R_t \mid s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s\right\}, \tag{3.8}$$

where $E_\pi\{\}$ denotes the expected value given that the agent follows policy π. Note that the value of the terminal state, if any, is always zero. We call the function V^π the *state-value function for policy* π.

Similarly, we define the value of taking action a in state s under a policy π, denoted $Q^\pi(s,a)$, as the expected return starting from s, taking the action a, and thereafter following policy π:

$$Q^\pi(s,a) = E_\pi\{R_t \mid s_t = s, a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s, a_t = a\right\}. \tag{3.9}$$

We call Q^π the *action-value function for policy* π.

The value functions V^π and Q^π can be estimated from experience. For example, if an agent follows policy π and maintains an average, for each state encountered, of the actual returns that have followed that state, then the average will converge to the state's value, $V^\pi(s)$, as the number of times that state is encountered approaches infinity. If separate averages are kept for each action taken in a state, then these averages will similarly converge to the action values, $Q^\pi(s,a)$. We call estimation methods of this kind *Monte Carlo methods* because they involve averaging over random samples of actual returns. These kinds of methods are presented in Chapter 5. Of course, if there are very many states, then it may not be practical to keep separate averages for each state individually. Instead, the agent would have to maintain V^π and Q^π as parameterized functions and adjust the parameters to better match the observed returns. This can also produce accurate estimates, although much depends on the nature of the parameterized function approximator (Chapter 8).

$$V(s_t) \leftarrow V(s_t) + \alpha \left[R_t - V(s_t) \right], \tag{6.1}$$

where R_t is the actual return following time t and α is a constant step-size parameter (cf., Equation 2.5). Let us call this method *constant-α MC*. Whereas Monte Carlo methods must wait until the end of the episode to determine the increment to $V(s_t)$ (only then is R_t known), TD methods need wait only until the next time step. At time $t+1$ they immediately form a target and make a useful update using the observed reward r_{t+1} and the estimate $V(s_{t+1})$. The simplest TD method, known as *TD(0)*, is

$$V(s_t) \leftarrow V(s_t) + \alpha \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right]. \tag{6.2}$$

In effect, the target for the Monte Carlo update is R_t, whereas the target for the TD update is $r_{t+1} + \gamma V_t(s_{t+1})$.

Because the TD method bases its update in part on an existing estimate, we say that it is a *bootstrapping* method, like DP. We know from Chapter 3 that

$$\begin{aligned} V^\pi(s) &= E_\pi \{ R_t \mid s_t = s \} \qquad (6.3) \\ &= E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s \right\} \\ &= E_\pi \left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\middle|\, s_t = s \right\} \\ &= E_\pi \{ r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s \}. \qquad (6.4) \end{aligned}$$

Roughly speaking, Monte Carlo methods use an estimate of (6.3) as a target, whereas DP methods use an estimate of (6.4) as a target. The Monte Carlo target is an estimate because the expected value in (6.3) is not known; a sample return is used in place of the real expected return. The DP target is an estimate not because of the expected value, which is assumed to be completely provided by a model of the environment, but because $V^\pi(s_{t+1})$ is not known and the current estimate, $V_t(s_{t+1})$, is used instead. The TD target is an estimate for both reasons: it samples the expected value in (6.4) *and* it uses the current estimate V_t instead of the true V^π. Thus, TD methods combine the sampling of Monte Carlo with the bootstrapping of DP. As we shall see, with care and imagination this can take us a long way toward obtaining the advantages of both Monte Carlo and DP methods.

Figure 6.1 specifies TD(0) completely in procedural form, and Figure 6.2 shows its backup diagram. The value estimate for the state node at the top of the backup diagram is updated on the basis of the one sample transition from it to the immediately

Initialize $V(s)$ arbitrarily, π to the policy to be evaluated
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    $a \leftarrow$ action given by π for s
    Take action a; observe reward, r, and next state, s'
    $V(s) \leftarrow V(s) + \alpha \left[r + \gamma V(s') - V(s) \right]$
    $s \leftarrow s'$
  until s is terminal

Figure 6.1 Tabular TD(0) for estimating V^π.

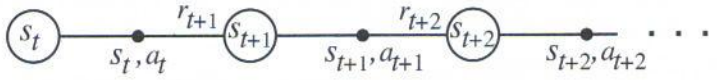
Figure 6.2 The backup diagram for TD(0).

following state. We refer to TD and Monte Carlo updates as *sample backups* because they involve looking ahead to a sample successor state (or state–action pair), using the value of the successor and the reward along the way to compute a backed-up value, and then changing the value of the original state (or state–action pair) accordingly. *Sample* backups differ from the *full* backups of DP methods in that they are based on a single sample successor rather than on a complete distribution of all possible successors.

6.4 Sarsa: On-Policy TD Control

We turn now to the use of TD prediction methods for the control problem. As usual, we follow the pattern of generalized policy iteration (GPI), only this time using TD methods for the evaluation or prediction part. As with Monte Carlo methods, we face the need to trade off exploration and exploitation, and again approaches fall into two main classes: on-policy and off-policy. In this section we present an on-policy TD control method.

The first step is to learn an action-value function rather than a state-value function. In particular, for an on-policy method we must estimate $Q^\pi(s, a)$ for the current behavior policy π and for all states s and actions a. This can be done using essentially the same TD method described above for learning V^π. Recall that an episode consists of an alternating sequence of states and state–action pairs:



In the previous section we considered transitions from state to state and learned the values of states. Now we consider transitions from state–action pair to state–action pair, and learn the value of state–action pairs. Formally these cases are identical: they are both Markov chains with a reward process. The theorems assuring the convergence of state values under TD(0) also apply to the corresponding algorithm for action values:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]. \tag{6.5}$$

This update is done after every transition from a nonterminal state s_t. If s_{t+1} is terminal, then $Q(s_{t+1}, a_{t+1})$ is defined as zero. This rule uses every element of the

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Choose a from s using policy derived from Q (e.g., ϵ-greedy)
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q (e.g., ϵ-greedy)
    $Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma Q(s', a') - Q(s, a) \right]$
    $s \leftarrow s'; a \leftarrow a';$
  until s is terminal

Figure 6.9 Sarsa: An on-policy TD control algorithm.