

# Neural Mechanisms of Learning and Control

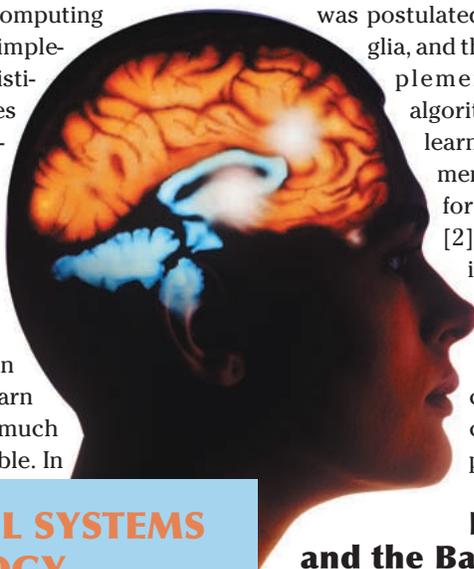
By Kenji Doya, Hidenori Kimura, and Mitsuo Kawato

The recent explosion in computing power has enabled the implementation of highly sophisticated control architectures and algorithms. Yet no artificial control system has been designed that works as flexibly and robustly as a biological control system. This motivates us to study the mechanisms of biological motor control.

Although “learning from nature” is an old concept, what we can actually learn from a biological system depends very much on the theories and technologies available. In the early days of cybernetics, the stability theory of feedback systems played an essential

role in understanding the function and dysfunction of the spinal reflex system. Later, the notion of precalculated, feedforward control led to a better understanding of the role of the cerebellum in the control of rapid movement in the face of feedback delays. More recently, the architecture of reinforcement learning, an online variant of dynamic programming, has provided critical insight about the function of the basal ganglia. These examples suggest that the development of novel system theories and gaining an understanding of biological systems are highly complementary endeavors.

It was traditionally believed that the functions of the cerebellum and the basal ganglia are limited to motor control. Growing evidence suggests, however, that they are involved in nonmotor, cognitive functions, too. Thus, a new theory



## CONTROL SYSTEMS IN BIOLOGY

was postulated that the cerebellum, the basal ganglia, and the cerebral cortex have evolved to implement different kinds of learning algorithms: the cerebellum for supervised learning, the basal ganglia for reinforcement learning, and the cerebral cortex for unsupervised learning (Fig. 1) [1], [2]. Here we introduce recent advances in motor control and learning, namely, the role of the basal ganglia in acquisition of goal-directed behaviors, learning of internal models by the cerebellum, and decomposition of complex tasks by the competition of predictive models.

## Reward Prediction and the Basal Ganglia

The behavior of an animal is generally goal directed; that is, to seek some kind of reward, such as food or water, and to avoid punishment, such as pain or death. Prediction of future reward is essential for learning motor behavior or even making any voluntary movement at all. The theory of *reinforcement learning* [3] provides a computational framework for learning goal-directed behavior through the interaction of an *agent* (e.g., an animal or a robot) with the *environment*. Here we introduce some basic concepts of reinforcement learning and show how they help in understanding the function of the basal ganglia in voluntary movement control.

## Temporal Difference Learning

Reinforcement learning is a paradigm in which an agent learns a policy that maps the sensory state  $x$  of the environment to a motor action  $u$  so that a certain reward  $r$  is maxi-

Doya (doya@isd.ATR.co.jp) and Kawato are with the Information Sciences Division, ATR International, 2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan. Kimura is with the Graduate School of Frontier Science, University of Tokyo, Japan.

mized (Fig. 2). When there are no dynamics in the environment, this is just a simple task of identifying the expected reward  $E[r|x,u]$  and selecting an action  $u(t)$  that maximizes the expectation for each given state  $x(t)$ . When the reward is given depending on the dynamics of the environment, consideration of immediate reward  $r(t)$  alone is not a good policy.

The so-called temporal difference (TD) learning is a framework for dealing with a reinforcement learning problem with delayed reward. Here we assume a simple Markov environment where the state  $x(t)$  evolves with the choice of an action  $u(t)$  and the transition probability  $P(x(t+1)|x(t),u(t))$ . The reward  $r(t)$  is given by either a deterministic function  $r(x,u)$  or a stochastic rule  $P(r(t)|x(t),u(t))$ . The goal is to find an action policy, either a deterministic one  $u(t) = g(x(t))$  or a stochastic one  $P(u(t)|x(t))$ , that maximizes the cumulative future reward

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

where  $0 \leq \gamma \leq 1$  is a *discount factor* for future rewards. The basic strategy is to first learn the above expectation for each starting state  $x$  as the *state value function*  $V(x)$  and then to improve the policy in reference to the value function.

For a given policy, the value functions of temporally adjacent states should satisfy

$$V(x(t)) = E[r(t) + \gamma V(x(t+1))],$$

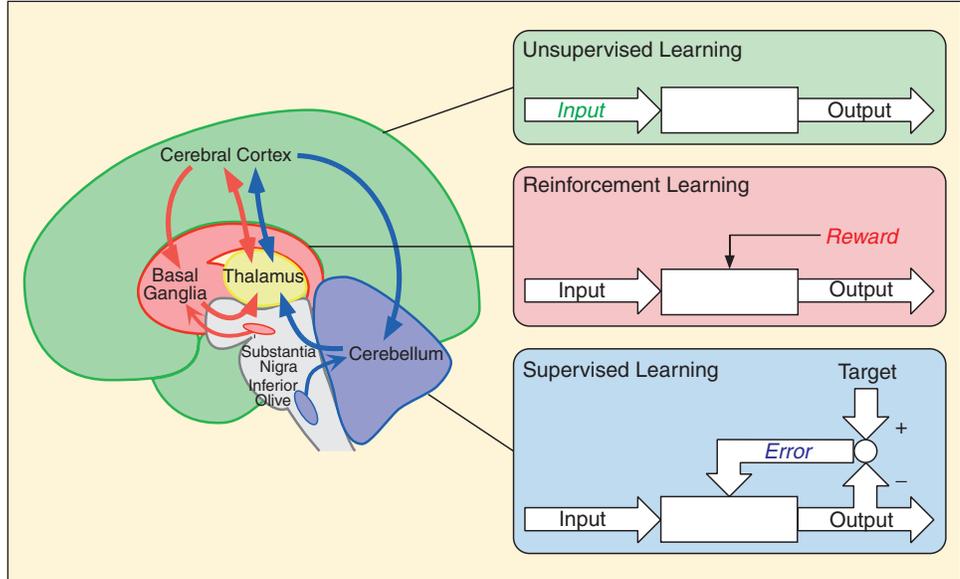
where  $r(t)$  and  $x(t+1)$  are dependent on the action  $u(t)$  and the environmental dynamics. Based on this constraint, the criterion for learning the state value function is to bring the deviation

$$\delta(t) = r(t) + \gamma V(x(t+1)) - V(x(t)) \quad (2)$$

to zero on average. This is called the TD error because it is derived from the temporal difference of the predicted future rewards. With this error signal, the value function is updated as

$$V(x(t)) := V(x(t)) + \alpha \delta(t), \quad (3)$$

where  $\alpha$  is a learning rate.

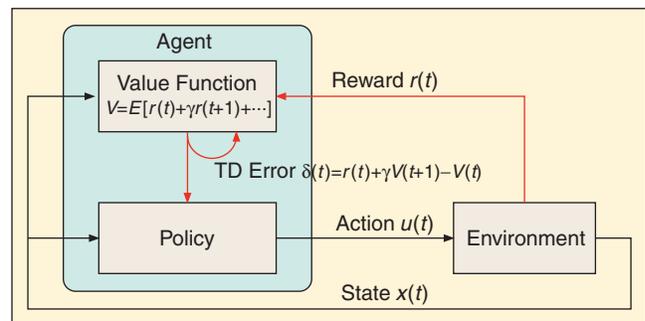


**Figure 1.** Learning-oriented specialization of the cerebellum, the basal ganglia, and the cerebral cortex [1], [2]. The cerebellum is specialized for supervised learning based on the error signal encoded in the climbing fibers from the inferior olive. The basal ganglia are specialized for reinforcement learning based on the reward signal encoded in the dopaminergic fibers from the substantia nigra. The cerebral cortex is specialized for unsupervised learning based on the statistical properties of the input signal.

Several ways of improving the policy based on the predicted future reward have been formulated, but the most popular one is to use the *action value function* defined as

$$Q(x(t),u) = E[r(t) + \gamma V(x(t+1))|x(t),u].$$

This gives an estimate of how much cumulative future reward one would get by trying an action  $u$  for now and then



**Figure 2.** The standard setup of reinforcement learning [3]. The agent observes the state  $x$  of the environment, takes an action  $u$  according to a policy  $u = g(x)$  (or  $P(u|x)$  in the stochastic case), and receives reward  $r$ . The goal of learning is to find a policy that maximizes the amount of reward acquired in the long run. An action has to be chosen by considering not only the reward immediately following the action, but also the reward delivered in the future, depending on the dynamics of the environment. For this reason, the agent learns a value function  $V(x)$ , which is a prediction of the cumulative future rewards. The difference between the predicted reward and the actual reward, or the TD error  $\delta$ , is used both for learning the value function  $V(x)$  and for improving the policy  $g(x)$ .

following the current policy. The action value function is updated by

$$Q(x(t), u(t)) := Q(x(t), u(t)) + \alpha \delta(t). \quad (4)$$

The policy is then updated as a *greedy* action selection

$$u(t) = \arg \max_u Q(x(t), u) \quad (5)$$

or its stochastic variant

$$\text{Prob}[u(t) = u_i] = \frac{\exp(\beta Q(x(t), u_i))}{\sum_k \exp(\beta Q(x(t), u_k))}, \quad (6)$$

where  $u_k$  denote the candidates of actions at  $x(t)$  and  $\beta > 0$  is a parameter that controls the randomness of action selection for exploration.

Thus, in the TD learning framework, the TD error  $\delta(t)$  plays the dual role of the teaching signal for reward prediction ( $V$ ) and action selection ( $Q$ ). TD learning has been successfully applied to a variety of control and optimization problems such as robot navigation and game playing programs [4].

In the early 1990s, it became clear that the heuristic algorithms of reinforcement learning had good correspondence with the framework of dynamic programming, which helped theoretical analyses of the learning algorithms [5]. It has been shown that the value function updated by (3) converges, under certain conditions, to the optimal value function that satisfies the Bellman equation

$$V(x(t)) = \max_u E[r(t) + \gamma V(x(t+1)) | x(t), u],$$

which is the necessary and sufficient condition for the value function of an optimal policy [3], [5]. Reinforcement learning was first developed as a heuristic optimization strategy, but it is now widely recognized as an online, model-free variant of dynamic programming.

The theory of reinforcement learning was first developed for discrete-time, discrete-state systems. A critical issue in applying it to motor control was how to discretize state, action, and time. The characterization of reinforcement learning as an online, model-free variant of dynamic programming also enabled us to derive new learning algorithms for continuous-time, continuous-state systems [6].

In the continuous-time case, the value function is defined as the weighted integral

$$V(x) = E \left[ \int_t^\infty e^{-(t-s)/\tau} r(s) ds \right], \quad (7)$$

where  $\tau$  is the time constant of reward prediction. The TD error is then defined as

$$\delta(t) = r(t) + \frac{dV(x(t))}{dt} - \frac{1}{\tau} V(x(t))$$

and is used for updating the value function and the policy in a similar way as in the discrete-time case. One merit of the continuous formulation is that we can utilize the gradient of the value function for efficient design of a policy. Suppose that the dynamics of the environment are given by

$$\frac{dx(t)}{dt} = f(x(t), u(t))$$

and the reward is given by

$$r(t) = r(x(t), u(t)).$$

Then the condition for the optimal value function is given by the Hamilton-Jacobi-Bellman (HJB) equation [6]

$$\frac{1}{\tau} V(x) = \max_u E \left[ r(x, u) + \frac{\partial V(x)}{\partial x} f(x, u) \right].$$

If the dynamics  $f(x, u)$  are linear with respect to the input  $u$  and the reward is convex with respect to each component  $u_j$  of the action variable, i.e.,

$$r(t) = R(x(t)) + \sum_j S_j(u_j(t)),$$

then the action that maximizes the right-hand side of the HJB equation is given by

$$u_j(t) = (S'_j)^{-1} \left( \left( \frac{\partial f(x(t), u)}{\partial u_j} \right)^T \left( \frac{\partial V(x(t))}{\partial x} \right)^T \right), \quad (8)$$

where ' denotes the derivative and  $T$  denotes the transpose. The gradient  $\partial V(x(t)) / \partial x$  of the value function represents the desired direction of movement in the state space. The action that realizes the movement under the system dynamics and the action cost constraint is calculated by the input gain matrix  $\partial f(x, u) / \partial u$  and a sigmoid function  $(S')^{-1}$  derived from the cost for the action [6].

If the value function  $V$  satisfies the HJB equation, (8) gives an optimal feedback control policy, which generalizes the popular linear quadratic control to the case of nonlinear dynamics and nonquadratic reward. Even if the optimal value function is unknown, we can use the above greedy policy (8) with respect to the current estimate of the value function. Fig. 3 shows an example of applying the above continuous reinforcement learning method to swing-up control of a cart-pole system [6]. With the use of the policy (8) while the value function  $V$  and the dynamic model  $f(x, u)$  are learned, the task was learned in significantly fewer trials than with a more conventional approach that does not use the gradient of the value function and the dynamic model.

## Reward-Predicting Response of Dopamine Neurons

The neurotransmitter dopamine is known to be involved in both the processing of reward and the control of movement. Most addictive drugs are known to increase the activity of dopamine in the brain. The major cause of Parkinson's disease is the loss of dopamine-delivering neurons.

Schultz and colleagues performed systematic experiments on the activities of dopamine neurons in the midbrain of monkeys while they performed visual response tasks [7]. Dopamine neurons initially responded to the delivery of reward, such as food and water (Fig. 4(a)). As the monkey became proficient in the lever-pressing task, however, the response to the reward disappeared. Instead, dopamine neurons responded to the visual stimulus that evoked the lever-pressing response, which in turn caused the delivery of reward (Fig. 4(b)). When the reward that usually followed a successful response was withheld, the activity of dopamine neurons was depressed (Fig. 4(c)). These results showed that the response of dopamine neurons does not just encode the reward itself, but the increase or decrease of the reward compared to what has been predicted.

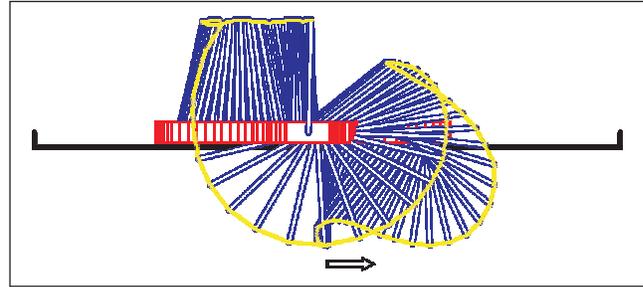
## TD Learning Model of the Basal Ganglia

The findings regarding the response of dopamine neurons in the course of learning was a big surprise to theoretical neuroscientists who were familiar with TD learning. The response to the reward itself before learning and the response to the reward predicting sensory state after learning are exactly how the TD error (2) should behave in the course of learning (Fig. 4).

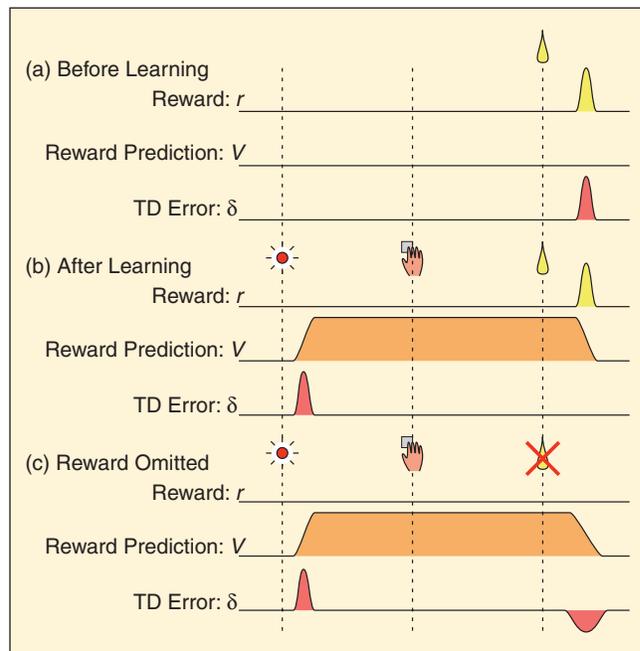
A major target of dopamine neurons is the basal ganglia, which are located between the brain stem and the cerebral cortex. They are known to be involved in motor control because damage to them results in severe motor deficits such as Parkinson's disease and Huntington's disease. However, their exact role under normal conditions has been quite an enigma.

Inspired by Schultz's findings, a number of models of the basal ganglia based on the TD learning paradigm have been proposed (Fig. 5) [8], [9]. The input part of the basal ganglia, called the *striatum*, receives strong dopaminergic input from the *compact part of the substantia nigra* (SNc). The striatum comprises two compartments: the *striosome* and the *matrix*. The striosome projects to the dopaminergic neurons in SNc. The matrix projects to the *reticular part of the substantia nigra* (SNr) and the *globus pallidus* (GP), whose outputs are sent to motor nuclei in the brain stem and through the thalamus to the cerebral cortex. This two-part architecture is reminiscent of the architecture of TD learning (Fig. 2): one for reward prediction and another for action selection [8].

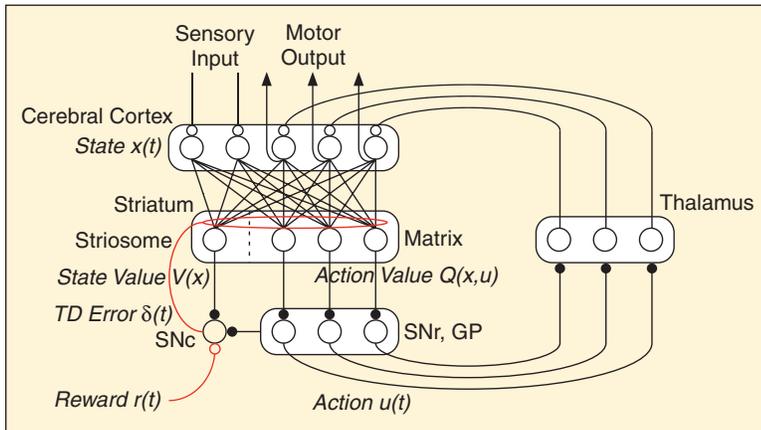
The circuit of the basal ganglia could implement TD learning as follows (Fig. 5). The state of the environment and the context of the task are represented in the cerebral cortex, denoted by  $x$ . The state value function  $V(x)$  is learned in the



**Figure 3.** An example application of continuous reinforcement learning for the task of swinging up a cart-pole system [6]. The state variables were the angle and the angular velocity of the pole (1 m, 0.1 kg) and the position and the velocity of the cart (1.0 kg) on the track (4.8 m). The action was the driving force to the cart (bounded by  $\pm 10$  N). The reward was given by the height of the tip of the pole. The value function and the policy were learned on the four-dimensional state space using radial basis function networks. Appropriate swing-up behaviors were learned after about 2,700 trials using a model-free learning algorithm and after 800 trials using a model-based learning algorithm.



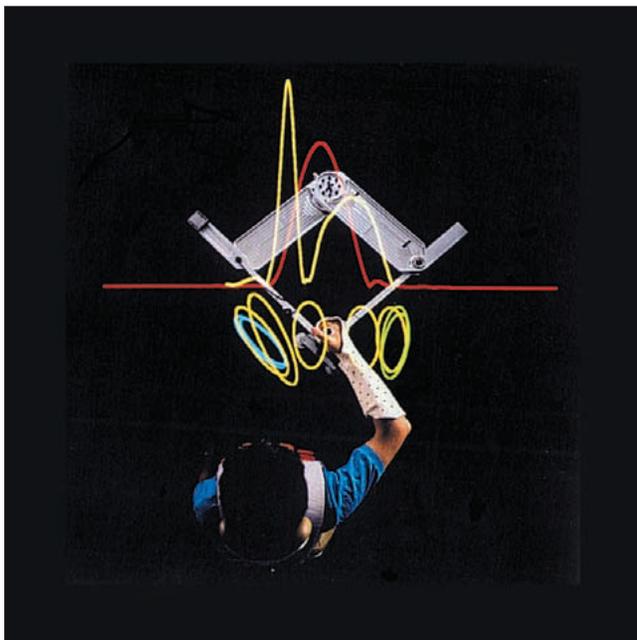
**Figure 4.** The response of midbrain dopamine neurons during a visual response task [7] and its interpretation by a TD learning model. (a) Initially, the dopamine neurons respond to the reward itself. Since the reward is not predicted (i.e.,  $V(t) \equiv 0$ ), the TD error (2) is the same as the reward, i.e.,  $\delta(t) = r(t)$ . (b) After learning, the visual stimulus elicits the response of dopamine neurons. Since the correct response is well established, the presentation of the stimulus lets the animal predict the delivery of reward. Thus, the increase in the value function  $V$  makes a positive TD error  $\delta(t) = V(t+1) - V(t)$ . (c) If the promised reward is withheld, the downturn in the predicted reward is observed as a dip in the TD error, which is usually cancelled with the actual delivery of reward.



**Figure 5.** A schematic diagram of the circuit of the basal ganglia and their loop connection with the cerebral cortex. The labels in italics show the hypothetical roles of the anatomical areas in the reinforcement learning model.

striosome, while the action value function  $Q(x,u)$  is learned in the matrix. One of the candidate actions  $u$  is selected in the output pathway of the basal ganglia by the competition of the action values, as in (5) or (6). Based on the resulting reward signal  $r$  and the value function of the new state, the TD error  $\delta$  is represented as the firing of dopamine neurons in the substantia nigra. Their output is fed back to the striatum and used as the learning signal for both the state value function  $V(x)$  and the action value function  $Q(x,u)$  in the striatum.

Several lines of evidence support the TD learning models of the basal ganglia. Projection neurons in the striatum have



**Figure 6.** A top view of the manipulandum (PFM: parallel link direct-drive air and magnet floating manipulandum) for measuring the mechanical impedance of the human arm [14]. The ellipses represent the stiffness of the arm in different directions.

many small spines in the dendrites. Each spine receives synaptic input from both cortical neurons and dopamine neurons, which suggests a tight interaction between the two inputs. Indeed, it has been shown that the change in the efficacy of the synapses from the cerebral cortex to the striatal neurons is modulated by the dopaminergic input [10]. This is consistent with the TD learning model where learning is based on the TD error  $\delta$ , as in (3) and (4). It has also been demonstrated in simulation that the TD model of the basal ganglia can replicate reward-based learning behaviors [11], [12].

Thus, by combining the experimental data from the basal ganglia and the theory of reinforcement learning, the role of the basal ganglia has become much clearer in the last several years. Further, the reinforcement learning model of the basal ganglia ignited a flood of new experiments on the activity of the basal ganglia and related brain areas in learning and decision making based

on the prediction of reward. It should be noted, however, that there are unresolved issues in the TD learning model of the basal ganglia. One is how the temporal difference of the value function  $V$  is actually calculated in the circuit linking the striatum and the dopamine neurons in SNc. Some other models have been proposed that explain the reward predictive response of the dopamine neurons [13].

## Internal Models and the Cerebellum

When a subject is asked to reach toward a target with an arm, the trajectory of the hand is roughly straight and its velocity has a smooth, bell-shaped profile. The mechanism underlying such a smooth arm trajectory has been a subject of much debate. One idea is that the viscoelastic property of the muscle plays a major role. In the *virtual trajectory* hypothesis, even if the brain sends a simple motor command in a step or ramp profile, a smooth trajectory is realized due to the physical properties of the muscle and the spinal feedback loop. For this mechanism to work, however, the arm should have rather high stiffness and damping.

Recently, Gomi and Kawato [14] measured the mechanical impedance of the human arm using a high-performance *manipulandum*, a robot manipulator that perturbs and monitors the movement of the subject's arm (Fig. 6). The results indicated that the arm has low stiffness and viscosity during movement. From a simulation with the measured mechanical impedance of the arm, it was concluded that the trajectory of the equilibrium point should have quite a complex shape to reproduce the smooth movement trajectory with a bell-shaped velocity profile. This result suggests that a control strategy which takes into account the dynamic property of the arm should be used, even for a simple arm-reaching movement. This motivated studies on how the *internal models* of the body and the environment are acquired by learning and used for control [15], [16].

## Internal Models in the Cerebellum

An obvious question is where in the brain are internal models of the body and the environment stored. Several lines of evidence suggest the cerebellum as a good candidate [15]. First, damage to the cerebellum often leads to a motor deficit, especially in quick, ballistic movements and coordination of multiple-joint movements. In such cases, subjects must rely on preprogrammed motor command rather than feedback control. Second, the anatomy and physiology of the cerebellum is well suited to the learning of internal models.

There are two major inputs to the cerebellar cortex: the *mossy fiber* inputs and the *climbing fiber* inputs (Fig. 7). The mossy fiber inputs are relayed by an enormous number of *granule cells*, larger than the number of all the neurons in the cerebral cortex, and their output, the parallel fibers, converge with the climbing fiber input at the Purkinje cells, the output neurons of the cerebellar cortex. A marked feature is that a single Purkinje cell receives inputs from about 200,000 parallel fibers, whereas it receives input from only a single climbing fiber.

This peculiar structure inspired Marr [17] and Albus [18] to propose a hypothesis that the cerebellum works as a pattern-classifier *perceptron* where the massive number of granule neurons work as a signal multiplexer and the climbing fiber input provides the teaching signal for Purkinje cells. As predicted by this hypothesis, Ito showed that the synaptic strength from the parallel fiber to the Purkinje cell is modified when the parallel fiber input coincides with the climbing fiber input [19].

Whereas the dopaminergic fibers to the striatum carry global, scalar information, climbing fibers to the cerebellum carry more specific information. Kobayashi and colleagues analyzed the *complex spike* response of Purkinje cells, which follows each climbing fiber spike input, during an eye movement task [20]. The result showed that each climbing fiber represents the direction and amplitude of error in eye movement. It was also shown in an arm-reaching task that climbing fiber input contains information about the direction in which the hand deviated from the target at the end of the movement [21].

These anatomical and physiological data suggest the following learning mechanism for the cerebellum. The object to be modeled, such as the dynamics or kinematics of the arm, determines the target mapping  $x \rightarrow y$ . For example,  $x$  is the motor command and  $y$  is the sensory outcome. The input signal  $x$ , provided as the mossy

fiber input, is transformed into an intermediate representation  $b_j(x(t))$  of a massive number of granule cells. The output of the Purkinje cell is given by the summation of the granule cell outputs  $b_j(x(t))$  multiplied by the synaptic weights  $w_j$ , namely,

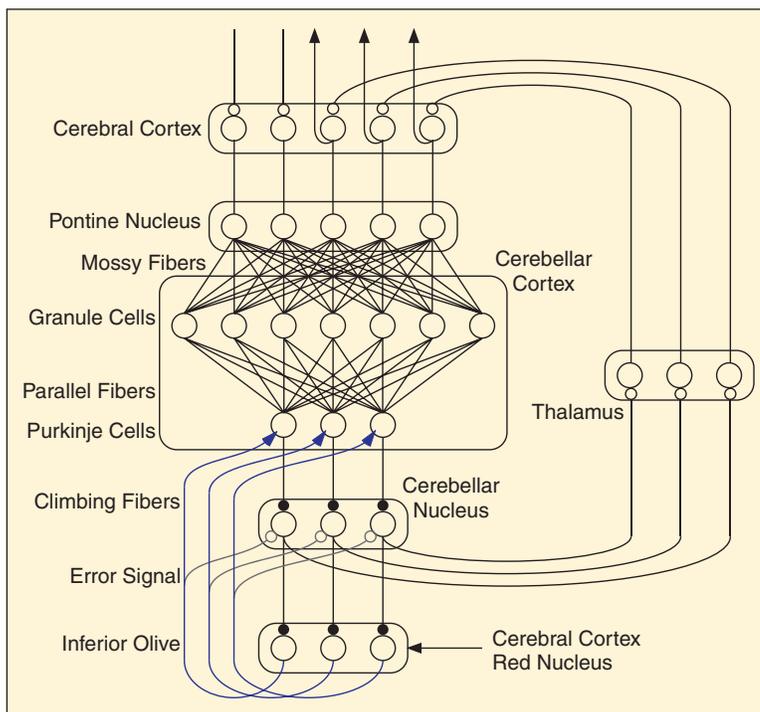
$$\hat{y}(t) = \sum_j w_j b_j(x(t)).$$

**Our brain implements the most efficient and robust control system available to date. How it really works cannot be understood just by watching its activity or by breaking it down piece by piece.**

The basic learning algorithm is to update the weights by the product of the output error and the input, i.e.,

$$w_j := w_j - \alpha(\hat{y}(t) - y(t))b_j(x(t)),$$

where  $\alpha$  is the learning rate. This learning can be implemented by the synaptic plasticity of the Purkinje cell if the appropriate error signal  $\hat{y}(t) - y(t)$  is provided as the climbing fiber input.



**Figure 7.** A schematic diagram of the circuit of the cerebellum and its loop connection with the cerebral cortex.

## Feedback Error Learning of Inverse Models

The above learning mechanism of the cerebellum potentially can be used for building an adaptive controller. However, a basic problem in building a controller is how to derive an appropriate error signal for the controller. In the case of pointing or tracking control, the role of the controller is to provide an *inverse model* of the controlled system. If the mapping from the motor command  $u$  to the sensory outcome  $y$  is given by  $F:u \rightarrow y$ , for a given sensory target  $y^d$ , the ideal motor command is given by the inverse  $u(t) = F^{-1}(y^d(t))$ .

## Findings suggest that the internal model in the cerebellum is involved in cancellation of ticklishness when a subject tickles him- or herself.

A naive method for realizing an inverse model is to train a network with the sensory output  $y(t)$  as the input and the motor input  $u(t)$  as the target output. Although this method could work in simple problems where the inverse is uniquely determined, it is not applicable to systems with redundancy and strong nonlinearity, which is usually the case with biological control systems.

One way of circumventing this problem was found by considering the circuit of the cerebellum. The cerebellum often serves as a side path to another control system, such as the brain stem circuit in the case of eye movement and the spinal reflex loop in the case of limb movement. The inferior olive, which sends climbing fibers to the cerebellum, receives inputs from these lower-level feedback loops.

Based on this notion, Kawato proposed the *feedback error learning* scheme, as shown in Fig. 8 [22]. This architecture comprises a fixed feedback controller that ensures the stability of the system and an adaptive feedforward controller that improves control performance. The output of the feedback controller is given by

$$u^{fb}(t) = K(y^d(t) - y(t)),$$

where  $y(t)$  is the actual output of the system. The key idea in feedback error learning is to use the output of the feedback controller  $u^{fb}$  as the error signal for the adaptive feedforward controller

$$u^{ff}(t) = G(y^d(t)).$$

The control output is given by the sum of both controllers

$$u(t) = u^{ff}(t) + u^{fb}(t) = G(y^d(t)) + K(y^d(t) - y(t)).$$

If the learning is complete, i.e.,  $y(t) = y^d(t)$ , then the feedforward controller  $G$  should be serving as the inverse  $F^{-1}$  of the system  $y = F(u)$ . One may wonder why the output of a simple, linear feedback controller can serve as the teaching signal for a more complex, nonlinear feedforward controller. This is possible under the assumption that the feedback gain matrix  $K$  provides an approximate Jacobian of the nonlinear system [23].

The feedback error learning architecture has been successfully applied to a variety of control tasks, such as the control of a robot arm with pneumatic actuators, for which standard linear feedback control does not work due to large time delays and nonlinearities. This architecture has also successfully replicated the experimental data of eye movement adaptation under realistic assumptions.

These successes motivated a rigorous analysis of the properties of the feedback error learning scheme from a control theoretic viewpoint. In [24] and [25], the feedback error learning method is treated as a new type of two-degree-of-freedom control scheme. A rigorous proof of its convergence is given in the case of linear invertible plants. Extensions to systems with delay are also discussed in [25].

## Model-Based Reinforcement Learning

Reinforcement learning can be considered an online, *model-free* version of dynamic programming, which is an offline, *model-based* optimal control method. Although a model-free strategy has the merits of simplicity and directness, learning tends to require a large number of trials. Incorporation of environmental models, either given in advance or learned online, has been considered to make reinforcement learning control more practical. We showed in the continuous-time case, using the example of cart-pole swing-up (Fig. 3), that an environmental model can considerably accelerate acquisition of an appropriate policy [6].

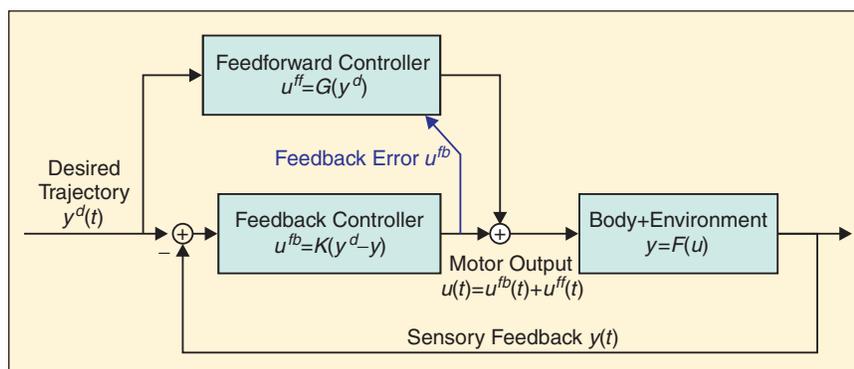


Figure 8. The feedback error learning architecture.

In the discrete-time, deterministic case, if a *forward model* of the system dynamics

$$x(t+1) = F(x(t), u(t))$$

and a model of reward condition  $r(x, u)$  are available, a greedy action can be found by

$$u(t) = \arg \max_u [r(x(t), u) + \gamma V(F(x(t), u))]. \quad (9)$$

This enables a focused search in relevant actions and accordingly facilitates learning of the value function. This action selection scheme requires predicting the next state  $x(t+1) = F(x(t), u)$  if an action  $u$  is taken at state  $x(t)$  and then evaluating the predicted state by the value function  $V(x)$ . If the internal model of the system dynamics  $F(x, u)$  is acquired in the cerebellum and the value function  $V(x)$  is learned in the basal ganglia, the above action selection mechanism can be realized by the collaboration of the cerebellum and the basal ganglia. Although there is no direct anatomical connection between the cerebellum and the basal ganglia, they both have loop connections with the cerebral cortex (Figs. 5 and 7). Thus interaction between the cerebellum and the basal ganglia can be realized by way of the cerebral cortex [1].

Brain imaging studies have shown that some parts of the cerebellum and the cortical areas that receive inputs from the cerebellum are activated during imagined movement [26]. Furthermore, the frontal part of the basal ganglia, which receives the inputs from the frontal cortex, is involved early in the acquisition of novel movements [27]. These findings are in agreement with the serial, model-based action selection scheme, which could be particularly useful before a stereotypical sensory-motor mapping is established.

### ***Internal Models for Perception, Simulation, and Encapsulation***

The phylogenetically old, medial part of the cerebellum receives major inputs from and sends outputs to the spinal cord (spinocerebellum); the phylogenetically newer, lateral part of the cerebellum receives major inputs from the cerebral cortex and sends the outputs back to the cerebral cortex (cerebrocerebellum). Although inverse models, whose outputs are motor commands, are likely to be located mainly in the medial cerebellum, forward models may reside in the lateral cerebellum for more versatile use by way of the cerebral cortex. The above model-based action selection (9) is one example, but there are many other ways of using forward models for learning and control [15], [16].

Forward models are also important in sensory perception. Sensory inputs are often affected by the subject's own motor outputs. Information about the external world could be pre-

cisely extracted by subtracting the sensory signal that is predicted from the motor output. A brain imaging study showed that the cerebellum is activated in a tactile object discrimination task using finger movement [28]. Findings in another study suggested that the internal model in the cerebellum is involved in cancellation of ticklishness when a subject tickles him- or herself [29].

Sensing of the state of the outside world is often delayed, corrupted by noise, or cannot be done directly. In such cases, it is useful to predict the current state using the internal model of the system dynamics, such as in the form of a Smith predictor or a Kalman filter. In the model-based action selection scheme (9), we considered only one-step prediction of the sensory outcome of an imaginary action. However, if there is enough working memory capacity, it is possible to perform a multistep simulation of a sensory-motor loop. Activation of the cerebellum, as well as the premotor and parietal cortex, have been reported in brain imaging studies that involve mental simulation of movement.

## **Reinforcement learning can be considered an online, model-free version of dynamic programming, which is an offline, model-based optimal control method.**

The target of learning by the cerebellum may not be limited to the body parts and the external world. It can be useful to learn a model of some other parts of the brain. For example, learning a new task by trial and error using the model-based action selection (9) involves global communication among different brain areas. However, once a task is learned, what is required is just to reproduce the appropriate sensory-motor mapping. A complex input-output mapping that was learned elsewhere in the brain could be used as the teaching signal for the cerebellum. This would allow a set of sensory-motor mapping functions, or a procedure, to be memorized in the cerebellum for rapid execution without widespread brain activation [30].

Although the circuit of the cerebellum is highly suitable for learning of internal models, this does not exclude the possibility that other parts of the nervous system also provide internal models of the body and the environment. For example, the storage and use of internal models in the circuit of the spinal cord has been suggested [31].

### **Modular Representation and the Cerebral Cortex**

In the above discussion, we considered the role of the cerebral cortex simply as a buffer or a patch board for the cerebel-

lum and the basal ganglia to work together. However, the cerebral cortex presumably performs a more sophisticated function. Essential roles of the cerebral cortex may be to extract important components from high-dimensional sensory inputs, to combine inputs from different sensory modalities, and to supplement missing information based on the context.

### Multimodal Representations for Sequence Learning

In studies of arm-reaching movement, a much debated issue has been the space in which movement trajectories are planned. Trajectory planning in the extrinsic, Cartesian coordinate system has the virtue of simplicity. However, the

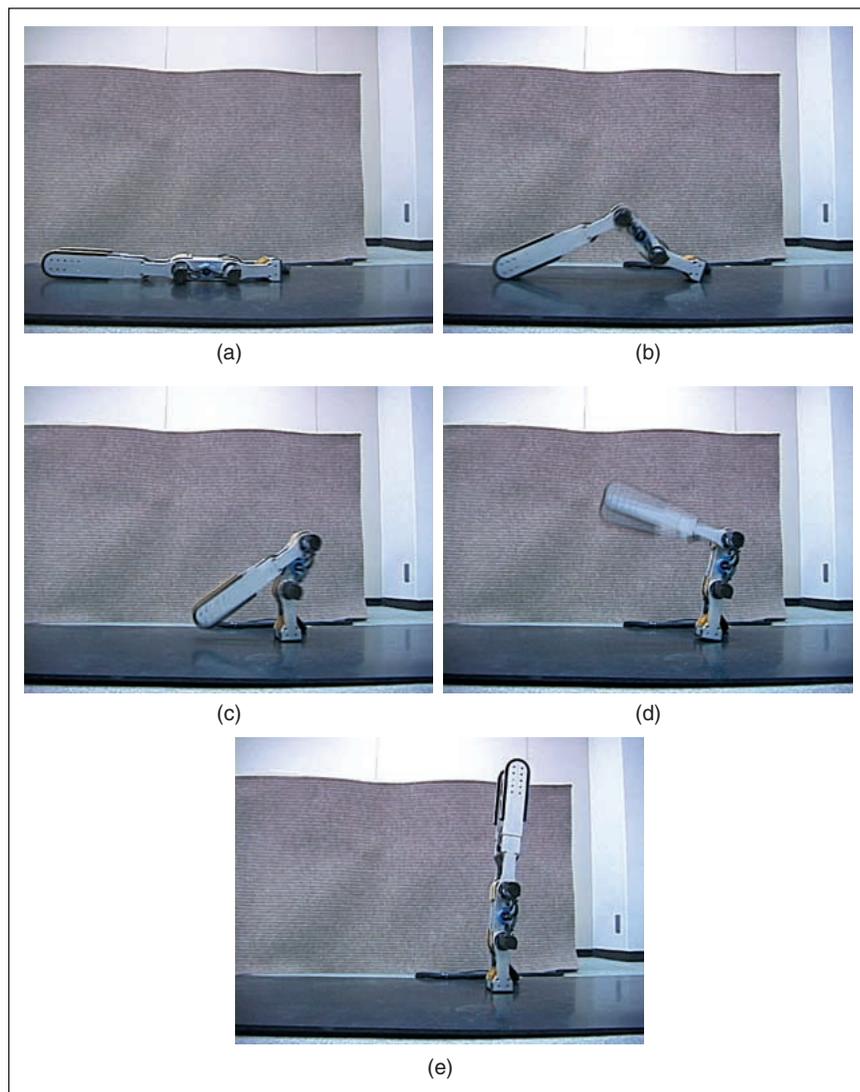
experimental data of arm movement are better explained by taking into account the nonlinear dynamics of the arm [32]. In real life, we may be employing both: extrinsic visual coordinates for easy planning and intrinsic motor coordinates for efficient execution.

In a series of sequence learning experiments, Hikosaka and colleagues found that sequence learning has at least two components: short-term learning of the correct order of component movements and long-term learning for fluently executing a series of movements [27]. Interestingly, they also found through pharmacological blockade experiments, as well as functional brain imaging studies, that different parts of the basal ganglia and the associated cortical areas are differentially involved in short-term and long-term

learning of sequences. The anterior (front) part of the basal ganglia and the cortical area it projects to are primarily involved in the acquisition of novel sequences, whereas the posterior (rear) part of the basal ganglia and the recipient cortical area are mainly involved in the execution of well-learned sequences. A question then is why the information about a sequence once acquired in one part of the control basal ganglia loop has to be sent to another part for skilled execution.

A computationally reasonable explanation is that there should be some change of information coding between different brain areas. In visually guided motor sequence learning, there can be at least two ways of defining a sequence. One is to describe the sequence of target positions in the extrinsic visual coordinates. Another is to describe the sequence of movement commands, such as arm postures or muscle commands, in intrinsic, body-specific coordinates. Visual coordinates are useful in learning a new sequence because the candidates of action targets are explicitly given. With motor coordinates, the same target can be reached with multiple postures and in various trajectories, but once this ill-posed problem is solved, quick, efficient movement can be realized. Anatomical and physiological data support the possibility that visual coordinates are used in the anterior basal ganglia loop and motor coordinates are used in the posterior basal ganglia loop.

We have built a network model that simulates the dual architecture with dif-



**Figure 9.** An example of stand-up behavior learned by a hierarchical reinforcement learning scheme [34]. The two-joint, three-link robot (70 cm in body length and 5 kg in mass) learned a nonlinear feedback policy in the six-dimensional state space (the head pitch angle, the two joint angles, and their temporal derivatives). Reward was based on the height of the head, and punishment was given when the robot tumbled. Successful stand-up was achieved after about 750 learning trials in simulation and then an additional 150 trials using real hardware.

ferent state representations [11]. The model replicated many of the experimental findings, including the differential effects of blockade experiments.

The hypothesis that different coordinate frames are used in different stages of sequence learning was tested in a human sequence learning experiment [33]. Once the subjects learned a button press sequence, their generalization performance was tested in two conditions: one in which the targets in the same spatial location were pressed with different finger movements and another in which different spatial targets were pressed with the same finger movements. Analysis of response times showed significantly better performance in the latter case (i.e., sequences in the same motor coordinates) after extended training. This supports our hypothesis that sequence representation in motor coordinates takes time to be developed, but, once developed, allows quick execution.

### Learning to Stand with Hierarchical Representations

The application of reinforcement learning to a high-dimensional dynamical system is quite difficult because of the complexity of learning the value function in a high-dimensional space, known as the *curse of dimensionality*. Thus, we developed a hierarchical reinforcement learning architecture [34]. In the upper level, coarsely discretized states in a reduced-dimensional space are used to make global exploration feasible. In the lower level, local dynamics in a continuous, high-dimensional state space are considered for smooth control performance.

We applied this hierarchical reinforcement learning architecture to the task of learning to vertically balance a three-link robot (Fig. 9). The goal was to find the dynamic movement sequence for standing. The reward was given by the height of the head and a punishment was given when the robot tumbled. Within several hundred trials, a successful pattern of standing up was achieved by the hierarchical reinforcement learning system. The learning was several times quicker than with simple reinforcement learning [34].

The reason for this quick learning was the selection of the upper-level state representation, which included kinematic, task-oriented variables such as the relative center of mass position measured from the foot. Although the hierarchical architecture was developed simply to let the robot learn the task quickly (before the hardware broke down), it is interesting that the multimodal representation resembles the architecture of the multiple cortico-basal ganglia loops [11], [27].

### Emergence of Modular Organization

A remarkable feature of the biological motor control system is its realization of both flexibility and robustness. For example, in a motor adaptation experiment of reaching to a target while wearing prism glasses, if a subject is well adapted to an altered condition and then returns to a normal condition, there is an after-effect (i.e., an error in the opposite direction). However, de-adaptation to a normal condition is usually much faster than adaptation to a novel condition. Further, if a subject is trained alternately in normal and altered conditions, she will adapt to either condition very quickly. Such results suggest that a subject does not simply modify the parameters of a single controller, but retains multiple controllers for different conditions and can switch between them easily. Evidence from arm reaching experiments suggests that the outputs of controllers for similar conditions can be smoothly interpolated for a novel, intermediate condition [15].

The idea of switching among multiple controllers is quite common; however, a difficult problem in designing an adaptive modular control system is how to select an appropriate module for a given situation. To evaluate a set of controllers, we basically have to test the performance of each controller

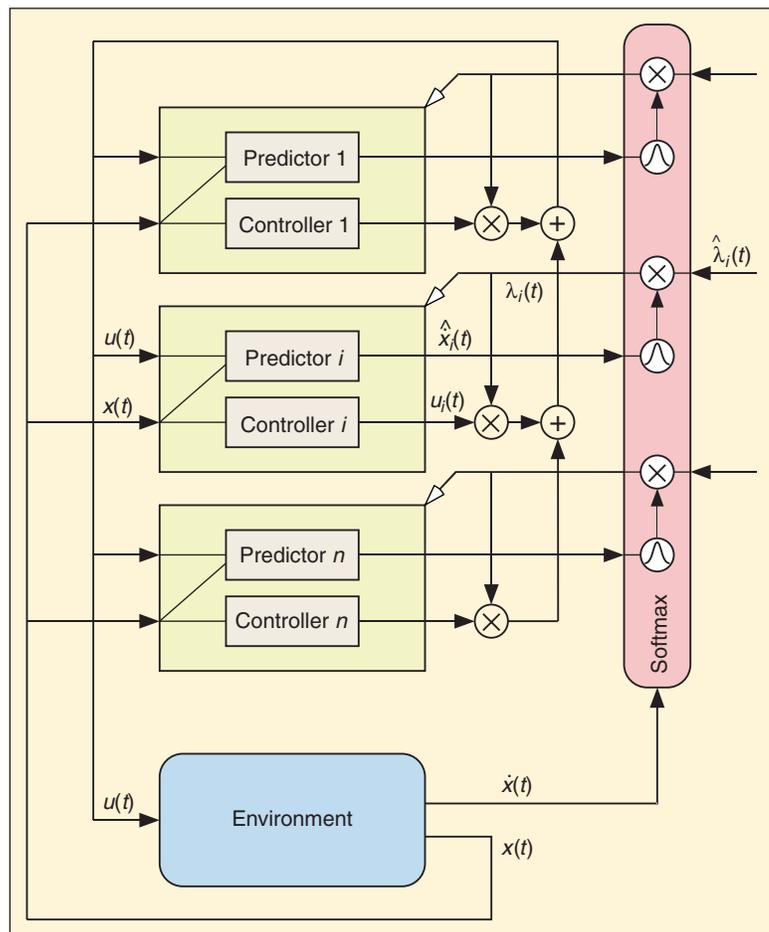


Figure 10. The MOSAIC architecture.

one by one, which takes a lot of time. On the other hand, a set of predictors can be evaluated simultaneously by running them in parallel and comparing their prediction outputs with the output of the real system. This simple fact motivated a modular control architecture in which each controller is paired with a predictor [35].

Fig. 10 shows the *module selection and identification control* (MOSAIC) architecture, based on the prediction error of each module

$$E_i(t) = \|\hat{x}_i(t) - x(t)\|^2,$$

where  $\hat{x}_i(t)$  is the output of the  $i$ th predictor. The *responsibility signal* is given by the soft-max function

$$\lambda_i(t) = \frac{\exp(E_i(t) / \sigma^2)}{\sum_j \exp(E_j(t) / \sigma^2)}.$$

This is used for weighting the outputs of multiple controllers, i.e.,

$$u(t) = \sum_i \lambda_i(t) u_i(t),$$

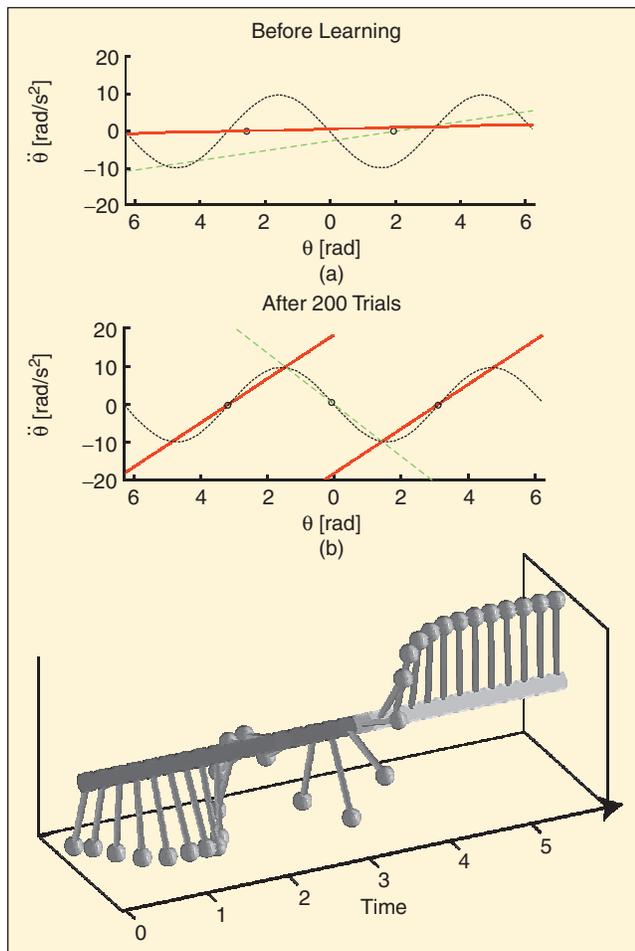
where  $u_i(t)$  is the output of the  $i$ th controller. The responsibility signal is also used to weight the learning rates of the predictors and the controllers of the modules, which causes modules to be specialized for different situations. The parameter  $\sigma$ , which controls the sharpness of module selection, is initially set large to avoid suboptimal specialization of modules.

Fig. 11 shows an example of using this scheme in a simple nonlinear control task of swinging up a pendulum [36]. Each module learns a locally linear dynamic model and a locally quadratic reward model. Based on these models, the value function and the corresponding control policy for each module are derived by solving a Riccati equation, which makes learning much faster than by iterative estimation of the value function. After about 100 trials, each module successfully approximated the sinusoidal nonlinearity in the dynamics in either the bottom or top half of the state space. Accordingly, a controller that destabilizes the stable equilibrium at the bottom and another controller that stabilizes the unstable equilibrium at the top were derived. They were successfully switched based on the responsibility signal.

The scheme has also been shown to be applicable to a nonstationary control task. The results suggest the usefulness of this biologically motivated modular control architecture in decomposing nonlinear and/or nonstationary control tasks in space and time based on the predictability of the system dynamics. A careful theoretical study assessing the conditions in which modular learning and control methods work reliably is still required.

Imamizu and colleagues performed a series of visuo-motor adaptation experiments using a computer mouse with rotated pointing direction (e.g., the cursor moves to the right when the mouse is moved upward). Initially in learning, a large part of the cerebellum was activated. However, as the subject became proficient in the rotated mouse movement, small spots of activation were found in the lateral cerebellum, which can be interpreted as the neural correlate of the internal model of the new tool [37]. Furthermore, when the subject was asked to use two different kinds of an unusual computer mouse, two different sets of activation spots were found in the lateral cerebellum (Fig. 12) [38].

Experiments of multiple sequential movement in monkeys have shown that neurons in the supplementary motor area (SMA) are selectively activated during movements in particular sequences. Furthermore, in an adjacent area called pre-SMA, some neurons were activated when the monkey



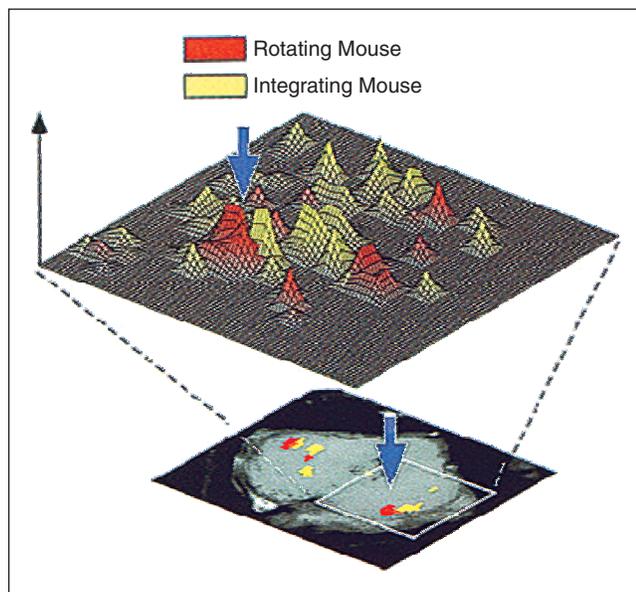
**Figure 11.** A result of learning to swing up an underpowered pendulum using a reinforcement learning version of the MOSAIC architecture [36]. (a) The sinusoidal nonlinearity of the gravity term in the angular acceleration was approximated by two linear models, shown in different colors. In each module, a local quadratic reward model was also learned and a linear feedback policy was derived by solving a Riccati equation. (b) The resulting policy for the first module made the downward position unstable. As the pendulum moved upward, based on the relative prediction errors of the two prediction models, the second module was selected and stabilized the upward position.

was instructed to change the movement sequence [39]. These results suggest the possibility that modular organization of internal models like MOSAIC is realized in a circuit including the cerebellum and the cerebral cortex.

## Conclusion

We reviewed three topics in motor control and learning: prediction of future rewards, the use of internal models, and modular decomposition of a complex task using multiple models. We have seen that these three aspects of motor control are related to the functions of the basal ganglia, the cerebellum, and the cerebral cortex, which are specialized for reinforcement, supervised, and unsupervised learning, respectively [1], [2].

Our brain undoubtedly implements the most efficient and robust control system available to date. However, how it really works cannot be understood just by watching its activity or by breaking it down piece by piece. It was not until the development of reinforcement learning theory that a clear light was shed on the function of the basal ganglia. Theories of adaptive control and studies of artificial neural networks were essential in understanding the function of the cerebellum. Such understanding provided new insights for the design of efficient learning and control systems. The theory of adaptive systems and the understanding of brain function are highly complementary developments.



**Figure 12.** The activity in the cerebellum for two different kinds of computer mouse: a rotating mouse (red), in which the direction of the cursor movement is rotated, and an integrating mouse (yellow), in which the mouse position specifies the velocity of the cursor movement trajectory on the screen, alternately using the two different mouse settings. Large areas in the cerebellum were activated initially. After several hours of training, activities were seen in limited spots in the lateral cerebellum, which were different for different types of mouse.

## Acknowledgments

We thank Raju Bapi, Hiroaki Gomi, Okihide Hikosaka, Hiroshi Imamizu, Jun Morimoto, Hiroyuki Nakahara, and Kazuyuki Samejima for their collaboration on this article. Studies reported here were supported by the ERATO and CREST programs of Japan Science and Technology (JST) Corp.

## References

- [1] K. Doya, "What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex," *Neural Net.*, vol. 12, pp. 961-974, 1999.
- [2] K. Doya, "Complementary roles of basal ganglia and cerebellum in learning and motor control," *Current Opinion in Neurobiology*, vol. 10, pp. 732-739, 2000.
- [3] R.S. Sutton and A.G. Barto, *Reinforcement Learning*. Cambridge, MA: MIT Press, 1998.
- [4] G. Tesauro, "TD-Gammon, a self-teaching backgammon program, achieves master-level play," *Neural Comput.*, vol. 6, pp. 215-219, 1994.
- [5] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.
- [6] K. Doya, "Reinforcement learning in continuous time and space," *Neural Comput.*, vol. 12, pp. 219-245, 2000.
- [7] W. Schultz, P. Dayan, and P.R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593-1599, 1997.
- [8] J.C. Houk, J.L. Adams, and A.G. Barto, "A model of how the basal ganglia generate and use neural signals that predict reinforcement," in *Models of Information Processing in the Basal Ganglia*, J.C. Houk, J.L. Davis, and D.G. Beiser, Eds. Cambridge, MA: MIT Press, 1995, pp. 249-270.
- [9] P.R. Montague, P. Dayan, and T.J. Sejnowski, "A framework for mesencephalic dopamine systems based on predictive Hebbian learning," *J. Neurosci.*, vol. 16, pp. 1936-1947, 1996.
- [10] J.N. Kerr and J.R. Wickens, "Dopamine D-1/D-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro," *J. Neurophysiol.*, vol. 85, pp. 117-124, 2001.
- [11] H. Nakahara, K. Doya, and O. Hikosaka, "Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuo-motor sequences—A computational approach," *J. Cognitive Neurosci.*, vol. 13, no. 5, 2001.
- [12] R.E. Suri and W. Schultz, "Temporal difference model reproduces anticipatory neural activity," *Neural Comput.*, vol. 13, pp. 841-862, 2001.
- [13] J. Brown, D. Bullock, and S. Grossberg, "How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues," *J. Neurosci.*, vol. 19, pp. 10502-10511, 1999.
- [14] H. Gomi and M. Kawato, "Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement," *Science*, vol. 272, pp. 117-120, 1996.
- [15] D.M. Wolpert, R.C. Miall, and M. Kawato, "Internal models in the cerebellum," *Trends in Cognitive Sciences*, vol. 2, pp. 338-347, 1998.
- [16] M. Kawato, "Internal models for motor control and trajectory planning," *Current Opinion in Neurobiology*, vol. 9, pp. 718-727, 1999.
- [17] D. Marr, "A theory of cerebellar cortex," *J. Physiol.*, vol. 202, pp. 437-470, 1969.
- [18] J.S. Albus, "A theory of cerebellar function," *Math. Biosci.*, vol. 10, pp. 25-61, 1971.
- [19] M. Ito, M. Sakurai, and P. Tongroach, "Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells," *J. Physiol.*, vol. 324, pp. 113-134, 1982.
- [20] Y. Kobayashi, K. Kawano, A. Takemura, Y. Inoue, T. Kitama, H. Gomi, and M. Kawato, "Temporal firing patterns of Purkinje cells in the cerebellar ven-

tral paraflocculus during ocular following responses in monkeys. II. Complex spikes," *J. Neurophysiol.*, vol. 80, pp. 832-848, 1998.

[21] S. Kitazawa, T. Kimura, and P.-B. Yin, "Cerebellar complex spikes encode both destinations and errors in arm movements," *Nature*, vol. 392, pp. 494-497, 1998.

[22] M. Kawato, K. Furukawa, and R. Suzuki, "A hierarchical neural network model for control and learning of voluntary movement," *Biol. Cybern.*, vol. 57, pp. 169-185, 1987.

[23] M. Kawato, "The feedback-error-learning neural network for supervised motor learning," in *Neural Network for Sensory and Motor Systems*, R. Eckmiller, Ed. Amsterdam: Elsevier, 1990, pp. 365-372.

[24] K. Doya, H. Kimura, and A. Miyamura, "Motor control: Neural models and system theory," *Appl. Math. Comput. Sci.*, vol. 11, pp. 101-128, 2001.

[25] A. Miyamura and H. Kimura, "Stability of feedback error learning scheme," submitted for publication.

[26] M. Lotze, P. Montoya, M. Erb, E. Hülsmann, H. Flor, U. Klose, N. Birbaumer, and W. Grodd, "Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fMRI study," *J. Cognitive Neurosci.*, vol. 11, pp. 491-501, 1999.

[27] O. Hikosaka, H. Nakahara, M.K. Rand, K. Sakai, X. Lu, K. Nakamura, S. Miyachi, and K. Doya, "Parallel neural networks for learning sequential procedures," *Trends Neurosci.*, vol. 22, pp. 464-471, 1999.

[28] J.H. Gao, L.M. Parsons, J.M. Bower, J. Xiong, J. Li, and P.T. Fox, "Cerebellum implicated in sensory acquisition and discrimination rather than motor control," *Science*, vol. 272, pp. 545-547, 1996.

[29] S.J. Blakemore, D.M. Wolpert, and C.D. Frith, "Central cancellation of self-produced tickle sensation," *Nature Neurosci.*, vol. 1, pp. 635-640, 1998.

[30] M. Ito, "Movement and thought: Identical control mechanisms by the cerebellum," *Trends Neurosci.*, vol. 16, pp. 448-450, 1993.

[31] Y.P. Shimansky, "Spinal motor control system incorporates an internal model of limb dynamics," *Biol. Cybern.*, vol. 83, pp. 379-389, 2000.

[32] E. Nakano, H. Imamizu, R. Osu, Y. Uno, H. Gomi, T. Yoshioka, and M. Kawato, "Quantitative examinations of internal representations for arm trajectory planning: Minimum commanded torque change model," *J. Neurophysiol.*, vol. 81, pp. 2140-2155, 1999.

[33] R.S. Bapi, K. Doya, and A.M. Harner, "Evidence for effector independent and dependent representations and their differential time course of acquisition during motor sequence learning," *Experimental Brain Res.*, vol. 132, pp. 149-162, 2000.

[34] J. Morimoto and K. Doya, "Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning," in *17th Int. Conf. Machine Learning*, 2000, pp. 623-630.

[35] D.M. Wolpert and M. Kawato, "Multiple paired forward and inverse models for motor control," *Neural Net.*, vol. 11, pp. 1317-1329, 1998.

[36] K. Doya, K. Samejima, K. Katagiri, and M. Kawato, "Multiple model-based reinforcement learning," Japan Sci. and Technol. Corp., Kawato Dynamic Brain Project Tech. Rep. KDB-TR-08, 2000.

[37] H. Imamizu, S. Miyauchi, T. Tamada, Y. Sasaki, R. Takino, B. Pütz, T. Yoshioka, and M. Kawato, "Human cerebellar activity reflecting an acquired internal model of a new tool," *Nature*, vol. 403, pp. 192-195, 2000.

[38] H. Imamizu, S. Miyauchi, Y. Sasaki, R. Takino, B. Pütz, and M. Kawato, "Separated modules for visuomotor control and learning in the cerebellum: A

functional MRI study," in *NeuroImage: Third International Conference on Functional Mapping of the Human Brain*, vol. 5, A.W. Toga, R.S.J. Frackowiak, and J.C. Mazziotta, Eds. Copenhagen, Denmark, 1997, pp. S598.

[39] K. Shima, H. Mushiake, N. Saito, and J. Tanji, "Role for cells in the presupplementary motor area in updating motor plans," in *Proc. Nat. Acad. Sci.*, vol. 93, pp. 8694-8698, 1996.

**Kenji Doya** received the Ph.D. in engineering from the University of Tokyo in 1991. He was a Research Associate at the University of Tokyo in 1986, at the University of California, San Diego, in 1991, and at Salk Institute in 1993. He has been a Senior Researcher at ATR International since 1994, and the Director of Metalearning, Neuromodulation, and Emotion Research, CREST, at JST, since 1999. He serves as an action editor of *Neural Networks* and *Neural Computation* and as a board member of the Japanese Neural Network Society. His research interests include reinforcement learning, the functions of the basal ganglia and the cerebellum, and the roles of neuromodulators in metalearning.

**Hidegori Kimura** received the Ph.D. in engineering from the University of Tokyo in 1970. He was appointed a faculty member at Osaka University in 1970, a Professor with the Department of Mechanical Engineering for Computer-Controlled Machinery, Osaka University, in 1987, and a Professor in the Department of Mathematical Engineering and Information Physics, University of Tokyo, in 1995. He has been working on the theory and application of robust control and system identification. He received the IEEE-CSS outstanding paper award in 1985 and the distinguished member award of the IEEE Control Systems Society in 1996. He is an IEEE Fellow.

**Mitsuo Kawato** received the Ph.D. in engineering from Osaka University in 1981. He became a faculty member at Osaka University in 1981, a Senior Researcher at ATR Auditory and Visual Processing Research laboratories in 1988, a department head at ATR Human Information Processing Research Laboratories in 1992, and the leader of the Computational Neuroscience Project at Information Sciences Division, ATR International, in 2001. He has been the project leader of the Kawato Dynamic Brain Project, ERATO, JST, since 1996. He received an outstanding research award from the International Neural Network Society in 1992 and an award from the Ministry of Science and Technology in 1993. He serves as a co-editor-in-chief of *Neural Networks* and a board member of the Japanese Neural Network Society. His research interests include the functions of the cerebellum and the roles of internal models in motor control and cognitive functions.