Current Opinion in
Neurobiology

Vol 17 No 2 April 2007

Tobias Bonhoeffer & Marc Tessier-Lavigne, Editors

April 2007

Cognitive neuroscience

Edited by Takeo Watanabe and Keiji Tanaka

June 2007 Signalling mechanisms
August 2007 Sensory systems
October 2007 Neuronal and glial cell biology • New technologies
December 2007 Motor systems • Neurobiology of behaviour
February 2008 Development

ScienceDirect Access CONB articles online up to one month before they appear in your print journal www.sciencedirect.com

CURRENT OPINION

Full text provided by www.sciencedirect.com

Efficient reinforcement learning: computational theories, neuroscience and robotics

Mitsuo Kawato[1] and Kazuyuki Samejima[2]

Reinforcement learning algorithms have provided some of the most influential computational theories for behavioral learning that depends on reward and penalty. After briefly reviewing supporting experimental data, this paper tackles three difficult theoretical issues that remain to be explored. First, plain reinforcement learning is much too slow to be considered a plausible brain model. Second, although the temporal-difference error has an important role both in theory and in experiments, how to compute it remains an enigma. Third, function of all brain areas, including the cerebral cortex, cerebellum, brainstem and basal ganglia, seems to necessitate a new computational framework. Computational studies that emphasize meta-parameters, hierarchy, modularity and supervised learning to resolve these issues are reviewed here, together with the related experimental data.

Addresses
[1] ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
[2] Brain Science Research Center, Tamagawa University Research Institute, 6-1-1, Tamagawa-gakuen, Machida, Tokyo 194-8610, Japan

Corresponding author: Kawato, Mitsuo (kawato@atr.jp)

Current Opinion in Neurobiology 2007, **17**:205–212

This review comes from a themed issue on
Cognitive neuroscience
Edited by Keiji Tanaka and Takeo Watanabe

Available online 19th March 2007

0959-4388/$ – see front matter
© 2007 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.conb.2007.03.004

Introduction

Reinforcement learning algorithms, as proposed and developed by Andy Barto and Rich Sutton in the early 1980s [1,2], have provided some of the most influential computational theories in neuroscience. In a class of reinforcement learning algorithms called 'temporal-difference learning', the temporal-difference error has a central role in guiding learning of the predicted reward and behaviors. Wolfram Schultz and colleagues [3,4] provided epoch-making neurophysiological data to suggest that activities of dopaminergic neurons encode temporal-difference error; this provoked several theorists [5,6] to propose computational learning models of basal ganglia and dopaminergic neurons based on reinforcement learning algorithms. Much experimental support of these models has been obtained over the past five years from different levels of organisms (e.g. cell, circuit, brain and behavioral levels) and use of different neuroscience techniques (e.g. slice experiments, extracellular unit recordings in behaving animals, lesion studies and human imaging). This paper briefly reviews the resulting data, and then describes the theoretical difficulties of using a plain reinforcement learning algorithm as a brain model. Several computational studies and relevant experimental data are introduced to deal with these issues.

Experimental data that support reinforcement learning theory

Various neurophysiological studies have focused on the role of dopaminergic neurons in decision making [7–9,10•,11], on the influences of these neurons on the striatum [12,13,14•] and on activities in the striatum that predict reward expectation [15,16••], reward expectation error [17,18••] or action–reward association [16••,18••]. Reward prediction-related neuronal activity in association learning has also been reported in prefrontal cortical areas [19,20]. Prefrontal and parietal areas of the cortex exhibit neuronal activities that are correlated with reward values during decision-making tasks in which there is a stochastic reward [21,22]. Neuronal activities in the lateral intraparietal region [22,23] were directly compared with computational representations in a dynamic stochastic-matching model of choice behavior [24,25], and the matching behavior could be explained by a reinforcement learning algorithm [26•]. Human functional magnetic resonance imaging (fMRI) studies [27,28••,29–36,37•,38] have also revealed neural correlates of putative computational variables such as discounted future reward [33,34], risk [35] and ambiguity [36] in decision-making tasks. Neural correlates of reward expectation error have been reported in the dorsal and ventral striatum [30,37•] and in the orbitofrontal cortex [31,38].

Most reinforcement learning algorithms possess meta-parameters. These include the learning rate (which determines the effects of experienced stimulus, action and reward on current estimates of values and resulting behaviors), the inverse temperature (which determines the randomness of behavioral choice), and the discount factor for future rewards (an important parameter that balances future and immediate rewards). Schweighofer and Doya [39] proposed their meta-learning hypothesis based on the previous theory [40] that different neuromodulators represent different meta-parameters. With regard to the learning rate, fast adaptation and long-term accumulation of experience should be assessed in stochastic environments because utilization of experience for current estimates of values is an important issue for creatures that live in dynamic environments. Choice behavior of monkeys in a matching task [22] can be explained by competition between reward values that are obtained by integrating past reward history with double exponential discounting, with a steep curve of forgetting for near-past experience and a long-tailed accumulation for experience in the far past [25]. A lesion to the medial prefrontal cortex affects length of time for which the experience of an outcome can be used in handle-movement choice tasks that switch the action–outcome contingency [41••]. With regard to the inverse temperature, the balance of exploitation and exploration is one of the most important theoretical issues in reinforcement learning, and it should be determined based on the uncertainty of environments and the information gained by exploration [42]. Neural activities that relate to the information gained are observed in the dorsal premotor cortex in monkeys [43], and exploratory behaviors activate anterior prefrontal cortical areas in humans [44••,45]. Uncertainty or information gained from external environments could be used for behavioral control by multiple decision systems within cortical and subcortical structures [46•]. With regard to the discount factor, discounting of future rewards has been studied in connection with impulsivity and addiction [33,34,47–49,50••]. The reinforcement learning theories have been extended to model the strategies of other people, game playing and subjective values [21,51–54]. It now seems safe to state that at least some aspects of learning that depend on reward and penalty can be understood within the framework of reinforcement learning theory.

Difficulties with plain reinforcement learning and hierarchy

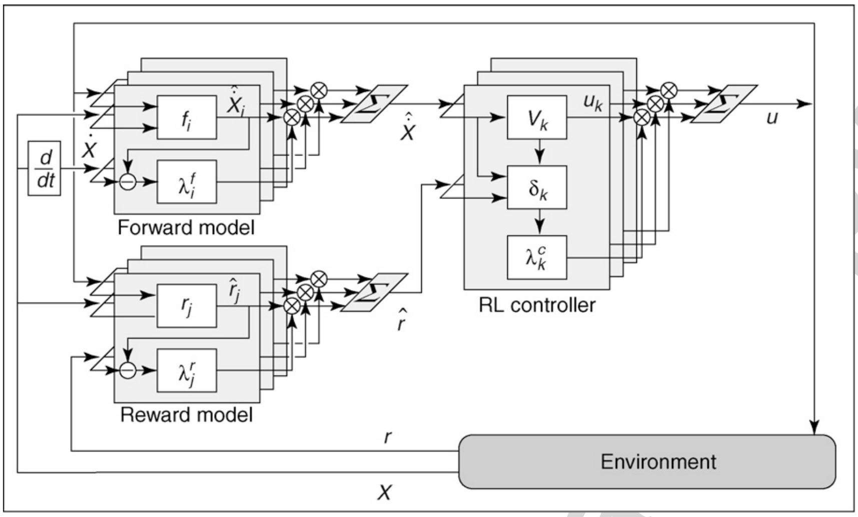
However, at least the following three difficult issues remain unresolved. First, the plain reinforcement learning algorithm is so slow when applied to practical problems such as robot control or financial decision that it cannot be considered a realistic model of brain learning. Second, although it is highly probable that dopaminergic neurons encode the temporal-difference error, the neural circuits and neural mechanisms that are used to compute this error are still unknown. Third, although some functional aspects of dopaminergic and striatal neural circuits might be understood by plain reinforcement learning algorithms, much more complicated algorithms seem to be required for behavioral learning that depends on the cerebral cortex and cerebellum in addition to the basal ganglia and brainstem.

The aforementioned three difficulties of plain reinforcement learning — that is, slowness, computation mechanism for temporal-difference error, and incompatibility with the brain structure — all point to specific structures and neural mechanisms for efficient learning algorithms. One possible solution to the problem of slowness is to put hierarchy into the algorithms [55–59]. In the upper part of such a hierarchy, because the state space describing a body and an external world is coarsely represented by a reasonable number of discrete regions, cells or neurons (e.g. up to one million neurons; a reasonable number for one function), an approximate optimal solution could be found in a reasonable time. In the lower part of this hierarchy, the state space is finely represented and the approximate solution that is obtained at the upper level can constrain the possible range of exploration space. Thus, the problem could be solved within a reasonable amount of time and with the required accuracy. In most studies, coarse representations have been specified in the upper hierarchy or to predetermine motor primitives and/or subgoals in the lower hierarchy. This recourse is nearly equivalent to assuming the existence of a homunculus or that those factors are all predetermined genetically — both assumptions that are unacceptable from the viewpoint of neuroscience. Most studies have dealt with discrete-state and discrete-time examples such as a maze, which is artificial from neuroscience viewpoint; one of the rare exceptions that adopted continuous time and continuous state-space examples was a standing-up task carried out by a multi-joint robot [60], which is computationally equivalent to human stand-up behaviors.

Modular reinforcement learning

One of the common strategies for tackling complicated and large-scale problems is 'divide and conquer'. A crucial and the most difficult issue in such modular reinforcement learning [61] is how to divide or modularize the original problem into subproblems [59]. One promising approach is modular selection and identification control (MOSAIC), which was originally proposed for supervised learning [62–64] and then extended to reinforcement learning [65]. In the most advanced form of reinforcement MOSAIC (N Sugimoto, PhD thesis, Nara Institute of Science and Technology, 2006), three different sets of expert networks, each of which is specialized for each subproblem, self-organize (Figure 1). One of these network sets is composed of predictors of state transitions, which are formally called internal forward models. An internal forward model receives information about the current state and the current motor command and then predicts the next time-step state. The two other sets approximate actual rewards and reinforcement learning controllers that compute motor commands for optimal control. The responsibility signal for the internal forward model network is determined by prediction error in dynamics, and the responsibility signal for the other two networks is determined by approximation error in reward and in reward prediction error; these signals are used to weight outputs from experts. These signals also gate the learning of each expert network by regulating the learning rate. Consequently, modularity emerges without ad hoc manual tuning by researchers or a homunculus. Switching and weighting by approximation and prediction errors are key features of MOSAIC. Medial prefrontal cortical areas might have an important role in detecting

Figure 1



Overall organization of the architecture used for combinatorial module-based reinforcement learning (CMRL). The architecture consists of forward models, reward modules and reinforcement learning (RL) controllers. The forward models (f_i) predict the environmental dynamics ($\dot{X}$) from the current state and action (x and u, respectively) and segregate the environmental dynamics into several subdynamics based on the prediction error of the dynamics. Each reward module (r_j) represents the local reward function conditioned on the state. Then an RL controller, receiving outputs from forward and reward modules, approximates local value functions (V_k) and outputs the local greedy action (u_k). Finally, the CMRL computes the sum of these local actions weighted by the responsibility signal of each RL module (λ_k^c), a posterior probability obtained from the squared temporal-difference errors (δ_k^2) of each RL module (likelihood) and the usage history of component forward and reward modules (priors). Prior probabilities of selecting a specific forward model or a reward model are generally useful because contextual cues often indicate probabilistically in which part of the environment animals are situated and what task requirements are there. These priors can be again computed as the responsibility signals of forward and reward modules (λ_i^f and λ_j^r) by applying Bayes rule to the prediction error of the environmental dynamics and reward approximation error, respectively. Please note that the architecture does not select only one forward model, reward model or RL controller, but instead utilizes many of them and blends their outputs, which are weighted by the responsibility signals.

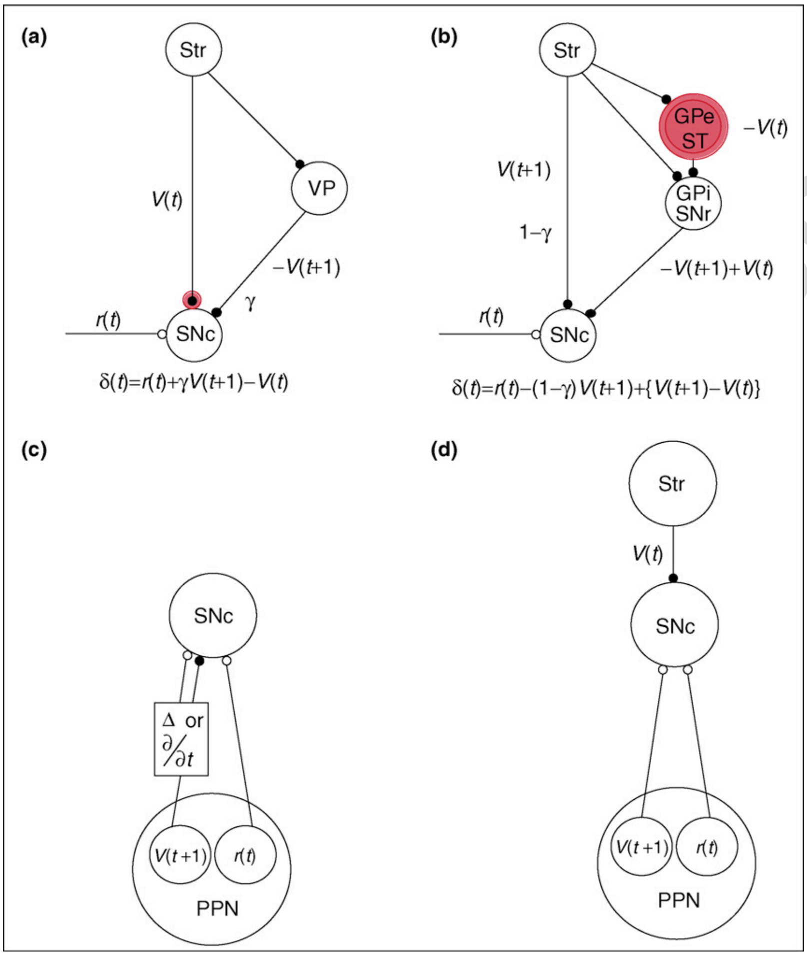
environmental change by accumulating prediction error, as suggested by the corresponding electroencephalogram (EEG) negativity in conflict monitoring or error detection. Lesion of the monkey medial prefrontal cortex affects the length of accumulating reward prediction error in contextual switching behavior [41••]. The responsibility signal could be interpreted as posterior probability of Bayesian inference, assuming that there are multiple linear systems that have Gaussian noise. In a human imaging study [38], ventromedial prefrontal activities were correlated with Bayesian update errors of probability, which are differences between the posterior and prior probabilities, in a stochastic reversal task.

Circumstantial support of hierarchy and modularity in the context of reinforcement learning was obtained mainly from neuroimaging and lesion studies. The activities of many brain areas — including those of the prefrontal cortex, orbitofrontal cortex, premotor cortex, supplementary motor area, cerebellum and basal ganglia — are correlated with important variables in reinforcement learning, such as accumulated reward, learned behavior, short-term reward and behavioral learning, even in a simple Markov decision process with monetary reward [32]. Furthermore, although activity in the human putamen is correlated with reward prediction that depends on selected actions, which is consistent with a study in monkeys [16••], activity in the caudate is correlated more closely with reward prediction error [18••,30,37•]. Topographic representations that have different parameters of the discount factor for reward prediction were found in the medial prefrontal cortex and the insular cortex, and topographic representations for reward prediction error were found in the basal ganglia [33].

How can temporal-difference error be computed?

The most important role of temporal-difference error is in solving the temporal credit assignment problem in reinforcement learning theory. Houk, Adams and Barto [5] proposed an explicit neural circuit model for computing the temporal-difference error (Figure 2a); later, Doya [6] revised this model (Figure 2b). In both models, the primary reward information $r(t)$ is assumed to be carried to dopaminergic neurons in the substantia nigra pars compacta (SNc) by direct excitatory inputs. By contrast, the reward expectation of the next time step $V(t+1) = \{r(t+1) + r(t+2)+\ldots\}$ minus the reward expectation at the current time step $-V(t) = -E\{r(t) + r(t+1)+\ldots\}$ is assumed to be computed within the basal ganglia network; these two

Figure 2



Models for computing temporal-difference error. **(a,b)** Possible mechanisms for computations of the temporal-difference error from the value function, as represented in the basal ganglia by the Houk model (a) [5] and the Doya model (b) [6]. Small white circles represent excitatory connections and small black circles represent inhibitory connections. Red circles represent sources of possible time delay. Abbreviations: GPe, globus pallidus pars externa; GPi, globus pallidus pars interna; SNc, substantia nigra pars compacta; SNr, substantia nigra pars reticulata; ST, subthalamic nucleus; Str, striatum; VP, ventral pallidum. **(c,d)** Two possible neural circuit models for computing the temporal-difference error, with emphasis on excitatory inputs from the pedunculopontine tegmental nucleus (PPN) to the SNc. Please note that in models of (a) and (b), the striatum represents $V(t + 1)$, but in the model of (d), the striatum represents $V(t)$ and PPN represents $V(t + 1)$. What unit of time is represented by '1', and how time advance or time delay is neurally implemented, still remain an enigma in reinforcement learning models for neuroscience.

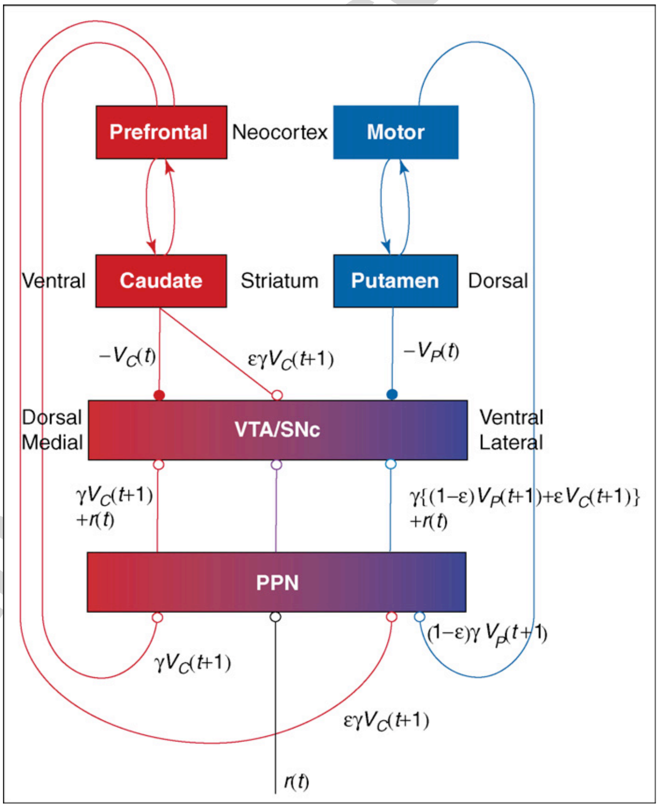
expectations utilize double-inhibitory and inhibitory connections, respectively, from the common source of the value function $V(t + 1)$ represented in the striatum. These models satisfy the most basic assumption of a reinforcement learning algorithm — that is, the Bellman consistency (approximately $V(t) = V(t + 1) + r(t)$ while neglecting discounting) as the source of the error signal — and they utilize the known neural network within the basal ganglia to compute this equation. However, no direct physiological support has been obtained for these models. Furthermore, double inhibition might not be able to generate burst firing of dopaminergic neurons in response to a cue signal that predicts the future reward, owing to the membrane properties of dopaminergic neurons that were revealed in slice experiments.

The SNc receives major excitatory synaptic inputs from the pedunculopontine tegmental nucleus (PPN) [66]. Recent neurophysiological studies [67] further suggest that excitatory inputs from the PPN to the SNc are central to computing temporal-difference error. Monkeys performed saccade tasks in which variable reward amounts were indicated by the shape of a fixation spot. Two populations of neurons whose firing rates varied with reward amount were observed: one population of neurons that seemed to encode the primary or current reward $r(t)$, and another population that started to fire at the presentation of the fixation spot and maintained firing until saccades or delivery of the primary reward ended, even after the fixation spot was extinguished. Furthermore, the firing rates of this second population of neurons predict the amount of future reward, thus seeming to encode the predicted reward or the value function $V(t+1)$. These remarkable findings suggest two possible neural mechanisms (Figure 2c,d) for computation of temporal-difference error. In Figure 2c, some intranuclear circuits within the PPN or SNc, or some membrane properties of dopaminergic neurons, execute either temporal difference or differentiation (box in Figure 2c). By contrast, the model in Figure 2d predicts that the primary reward information $r(t)$ and the expected reward at the next time step $V(t+1)$ are carried by excitatory inputs from the PPN to the SNc, whereas the inhibitory input from the striatum conveys the subtracted predicted reward information at the current time $-V(t)$. The model in Figure 2d seems more plausible, first because neither cellular nor circuit mechanisms for temporal difference or differentiation is known in the PPN or SNc, and second because the inhibitory inputs from the striatum to the SNc are well established — it would be odd if they did not contribute to computing the temporal-difference error at all.

Heterarchical reinforcement learning model

In the models of Figure 2a–c, the temporal-difference error is computed from the same source as the value function, thus conforming to the most basic principle of temporal-difference learning: the Bellman consistency. By contrast, $V(t)$ and $V(t+1)$ come from different sources in Figure 2d, and thus doubt is thrown on the fidelity of this model to the Bellman consistency; however, this deviation might lend efficiency and power to the model. If subtraction can be computed at the SNc between the PPN and striatal inputs, some supervised learning aspects could be introduced into the reinforcement learning: the learner $V(t)$ might partially approximate the teaching signal $r(t) + (t+1)$, and the error signal for this supervised learning might be conveyed by dopaminergic neurons. The model shown in Figure 2d led to the 'heterarchical' reinforcement learning model shown

Figure 3



A basic heterarchical reinforcement learning model consisting of a caudate–prefrontal loop (red) and a motor–putamen loop (blue). Also involved are the ventral tegmental area (VTA), substantia nigra pars compacta (SNc) and pedunculopontine tegmental nucleus (PPN). White and colored circles show excitatory (disinhibitory) and inhibitory projections, respectively. The double inhibitory connection (open circle) from the caudate to VTA/SNc is realized by neural circuits depicted in Figure 2 (a) and (b). V_C and V_P represent a coarse value function in the caudate nucleus and a fine value function in the putamen, respectively. r is a reward, γ is a discount factor for future rewards, and ε is a time-varying weighting factor between V_C and V_P.

in Figure 3 [68••], which has two almost independent closed loops: a cognitive loop involving the prefrontal cortex, caudate nucleus, ventral tegmental area and PPN; and a sensorimotor loop involving the motor cortex, putamen, SNc and PPN. If these two closed loops are entirely independent, then they conform to the Bellman consistency. However, the heterarchical structure enables the cognitive-loop value function to spread to that of the sensorimotor loop, by neural connections from the caudate to the ventrolateral part of the SNc, and also connections from the prefrontal cortex to the part of the PPN that innervates the ventrolateral SNc, which are supported by the spiral connections of corticostriatal and striatonigral loops [69]. This heterarchical model captures some aspects of the hierarchical reinforcement learning models [55–60] but does not have any strict hierarchy — hence the name 'heterarchy'. Representations of the state and actions in the cognitive loop are much coarser than those for the sensorimotor loop, so reinforcement learning progresses much more quickly in the cognitive loop. Then, the earlier learned value function in the cognitive loop $V_C(t+1)$ acts as a teaching signal for the sensorimotor loop. However, a much more fine-grained value function will be ultimately learned by the proper reinforcement learning that is implemented by the sensorimotor loop. Thus, fast but accurate learning is possible, and no distinct boundaries need to be set by a homunculus. PPN neural firing has been characterized as encoding intrinsic motivation or internal rewards [70]. Because the PPN receives strong inputs from brainstem nuclei, the future expected reward $V(t+1)$ could be represented on the basis of relatively coarse and endogenous factors, almost to the exclusion of sensorimotor or cognitive inputs; however, even the coarse future expected reward might still be very helpful for fine-tuning value functions on the basis of sensorimotor and cognitive inputs, similar to a computational technique known as 'reward shaping' [71]. Consequently, the spiral and heterarchical structure might provide a coherent resolution to the three theoretical difficulties of reinforcement learning — that is, slowness, computation of temporal-difference error, and global neural networks — in addition to a clue for refining computational understanding of intrinsic motivation.

Conclusions

Reinforcement learning theory has gained much support from experiments conducted at different levels within organisms and using different techniques. Important variables such as the reward prediction error and value functions that depend on state or on both state and action have been found to correlate with neural activities and/or fMRI blood oxygen level dependent (BOLD) signals of various brain areas. Hyper-parameters in the reinforcement learning algorithms, such as the learning rate, the inverse temperature and the discount rate, are now known to be important in guiding experimental paradigms and interpreting experimental data. However, difficult theoretical issues remain to be explored in combination with new experimental paradigms. First, the plain reinforcement learning is much too slow. Second, how can temporal-difference error be computed? Third, how do global brain networks function in reward-dependent behavioral learning? Hierarchy, modularity and solutions blended with supervised learning have been studied in computational fields to cope with these issues. The PPN might provide the most important information in computing temporal-difference error, and a heterarchical reinforcement learning model based on this hypothesis might coherently resolve the theoretical difficulties.

Acknowledgements

We thank Norikazu Sugimoto and Masahiko Haruno for preparing the figures and their legends.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Barto AG, Sutton RS, Anderson CW: **Neuron-like elements that can solve difficult learning control problems**. *IEEE Trans Syst Man Cybern* 1983, **13**:835-846.

2. Sutton RS, Barto AG: *Reinforcement learning*. The MIT Press; 1998

3. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward**. *Science* 1997, **275**:1593-1599.

4. Schultz W, Dickinson A: **Neuronal coding of prediction errors**. *Annu Rev Neurosci* 2000, **23**:473-500.

5. Houk JC, Adams JL, Barto AG. **Models of information processing in the basal ganglia.** Edited by Houk JC, Davis JL, Beiser DG. The MIT Press; 1995.

6. Doya K: **Complementary roles of basal ganglia and cerebellum in learning and motor control**. *Curr Opin Neurobiol* 2000, **10**:732-739.

7. Satoh T, Nakai S, Sato T, Kimura M: **Correlated coding of motivation and outcome of decision by dopamine neurons**. *J Neurosci* 2003, **23**:9913-9923.

8. Takikawa Y, Kawagoe R, Hikosaka O: **A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping**. *J Neurophysiol* 2004, **92**:2520-2529.

9. Fiorillo CD, Tobler PN, Schultz W: **Discrete coding of reward probability and uncertainty by dopamine neurons**. *Science* 2003, **299**:1898-1902.

10. • Tobler PN, Fiorillo CD, Schultz W: **Adaptive coding of reward value by dopamine neurons**. *Science* 2005, **307**:1642-1645.

When a monkey receives unexpected reward, phasic change of activity of dopaminergic neurons can be explained as encoding the temporal-difference error. These authors added new findings regarding the scaling of this coding: this error signal is scaled by variance of reward distribution within a given context. Thus, the activity of dopaminergic neurons is scaled by the range of reward distribution, and could represent expected 'risk' of reward in the context.

11. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H: **Midbrain dopamine neurons encode decisions for future action**. *Nat Neurosci* 2006, **9**:1057-1063.

12. Wickens JR, Reynolds JN, Hyland BI: **Neural mechanisms of reward-related motor learning**. *Curr Opin Neurobiol* 2003, **13**:685-690.

13. Kawagoe R, Takikawa Y, Hikosaka O: **Reward-predicting activity of dopamine and caudate neurons — a possible mechanism of**

motivational control of saccadic eye movement. *J Neurophysiol* 2004, **91**:1013-1024.

14. Nakamura K, Hikosaka O: **Role of dopamine in the primate**
• **caudate nucleus in reward modulation of saccades**. *J Neurosci* 2006, **26**:5360-5369.

This study demonstrates the role of dopamine for sensorimotor adaptation in the striatum. The authors injected two antagonists of dopamine receptors into the monkey caudate during a visually guided saccade task with asymmetrical reward delivery. The saccade latency of rewarded direction was shorter than that of non-rewarded direction. When the rewarded direction was switched to the opposite direction, monkey behavior followed that change. An antagonist of D_1 dopamine receptors attenuated the reward-dependent change of eye movement, whereas an antagonist of D_2 receptors enhanced it. The dopamine-dependent plasticity in corticostriatal synapses could modulate sensorimotor learning dependent on reward.

15. Cromwell HC, Hassani OK, Schultz W: **Relative reward processing in primate striatum**. *Exp. Brain Res* 2005, **162**:520-525.

16. Samejima K, Ueda Y, Doya K, Kimura M: **Representation of**
•• **action-specific reward values in the striatum**. *Science* 2005, **310**:1337-1340.

This study demonstrates that a considerable proportion of dorsal striatal neural activity represents the 'action value' predicted by Doya [6]. The authors then directly compared striatal neuronal activities with dynamically changing model parameters in a trial-by-trial manner, using a sophisticated estimation method, and found a good fit.

17. Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H: **Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons**. *Neuron* 2004, **43**:133-143.

18. Williams ZM: **Eskandar EN: Selective enhancement of**
•• **associative learning by microstimulation of the anterior caudate**. *Nat Neurosci* 2006, **9**:562-568.

This paper provides direct evidence that the striatum contributes to stimulus–action–reward association learning. The authors recorded caudate neuronal discharge that is temporarily reinforced during learning. They also applied electrical microstimulation to the caudate nucleus while the monkey learned arbitrary association between a visual stimulus and movement guided by reward feedback, and the microstimulation selectively enhanced the association learning.

19. Pasupathy A, Miller EK: **Different time courses of learning-related activity in the prefrontal cortex and striatum**. *Nature* 2005, **433**:873-876.

20. Matsumoto K, Suzuki W, Tanaka K: **Neuronal correlates of goal-based motor selection in the prefrontal cortex**. *Science* 2003, **301**:229-232.

21. Barraclough DJ, Conroy ML, Lee D: **Prefrontal cortex and decision making in a mixed-strategy game**. *Nat Neurosci* 2004, **7**:404-410.

22. Sugrue LP, Corrado GS, Newsome WT: **Matching behavior and the representation of value in the parietal cortex**. *Science* 2004, **304**:1782-1787.

23. Sugrue LP, Corrado GS, Newsome WT: **Choosing the greater of two goods: neural currencies for valuation and decision making**. *Nat Rev Neurosci* 2005, **6**:363-375.

24. Lau B, Glimcher PW: **Dynamic response-by-response models of matching behavior in rhesus monkeys**. *J Exp Anal Behav* 2005, **84**:555-579.

25. Corrado GS, Sugrue LP, Seung HS, Newsome WT: **Linear–nonlinear–Poisson models of primate choice dynamics**. *J Exp Anal Behav* 2005, **84**:581-617.

26. Sakai Y, Okamoto H, Fukai T: **Computational algorithms and**
• **neuronal network models underlying decision processes**. *Neural Netw* 2006, **19**:1091-1105.

This review provides computational explanations of perceptual decision and matching behavior, which have been studied in the field of behavioral psychology, and their relationship to optimal learning theories and algorithms of reinforcement learning.

27. Montague PR, King-Casas B, Cohen JD: **Imaging valuation models in human choice**. *Annu Rev Neurosci* 2006, **29**:417-448.

28. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD:
•• **Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans**. *Nature* 2006, **442**:1042-1045.

Using combinations of pharmacological, imaging and computational modeling techniques, the authors directly demonstrate that fMRI signals in the striatum and orbitofrontal cortex are modulated by dopamine.

29. Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS: **Temporal difference models describe higher-order learning in humans**. *Nature* 2004, **429**:664-667.

30. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ: **Dissociable roles of ventral and dorsal striatum in instrumental conditioning**. *Science* 2004, **304**:452-454.

31. Tobler PN, O'Doherty JP, Dolan RJ, Schultz W: **Human neural learning depends on reward prediction errors in the blocking paradigm**. *J Neurophysiol* 2006, **95**:301-310.

32. Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, Imamizu H, Kawato M: **A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task**. *J Neurosci* 2004, **24**:1660-1665.

33. Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S: **Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops**. *Nat Neurosci* 2004, **7**:887-893.

34. Tanaka SC, Samejima K, Okada G, Ueda K, Okamoto Y, Yamawaki S, Doya K: **Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics**. *Neural Netw* 2006, **19**:1233-1241.

35. Preuschoff K, Bossaerts P, Quartz SR: **Neural differentiation of expected reward and risk in human subcortical structures**. *Neuron* 2006, **51**:381-390.

36. Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF: **Neural systems responding to degrees of uncertainty in human decision-making**. *Science* 2005, **310**:1680-1683.

37. Haruno M, Kawato M: **Different neural correlates of reward**
• **expectation and reward expectation error in the putamen and caudate nucleus during stimulus–action–reward association learning**. *J Neurophysiol* 2006, **95**:948-959.

This study is an example of recent research trends of computational-model-based neuroimaging. Subjects' choice behavior was modeled by a simple reinforcement learning algorithm (Q-learning) while the model was given the same sensory stimuli and received the same rewards. The model reproduced subjects' behavior reasonably well. Internal representations within the model, such as the reward expectation and the reward expectation error, were found to correlate differentially with activities in the putamen and caudate, respectively.

38. Hampton AN, Bossaerts P, O'Doherty JP: **The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans**. *J Neurosci* 2006, **26**:8360-8367.

39. Schweighofer N, Doya K: **Meta-learning in reinforcement learning**. *Neural Netw* 2003, **16**:5-9.

40. Doya K: **Metalearning and neuromodulation**. *Neural Netw* 2002, **15**:495-506.

41. Kennerley SW, Walton ME, Behrens TE, Buckley MJ,
•• Rushworth MF: **Optimal decision making and the anterior cingulate cortex**. *Nat Neurosci* 2006, **9**:940-947.

This study shows that the anterior cingulate cortex has an important role in utilizing integrated past history of action and reward experiences for reward-based decision making. The authors demonstrate that lesion of the monkey anterior cingulate cortex did not impair performance just after error trials, but it rendered the monkeys unable to maintain optimal choice. The authors also found that in a matching task, monkeys that had lesions of the anterior cingulate cortex took longer to attain the optimum choice ratio.

42. Ishii S, Yoshida W, Yoshimoto J: **Control of exploitation–exploration meta-parameter in reinforcement learning**. *Neural Netw* 2002, **15**:665-687.

43. Nakamura K: **Neural representation of information measure in the primate premotor cortex**. *J Neurophysiol* 2006, **96**:478-485.

44. Yoshida W, Ishii S: **Resolution of uncertainty in prefrontal**
•• **cortex**. *Neuron* 2006, **50**:781-789.

In many situations, we face difficulty finding ascertaining where we are from a limited set of available sensory inputs. The problem of finding an optimal strategy with only partial information of a current state is called a 'partially observable Markov decision process'. The authors modeled how humans resolve this problem in a task of pathfinding through a maze of which there was only a limited view. The anterior prefrontal BOLD signal was found to correlate with the uncertainty of a current belief state.

45. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ: **Cortical substrates for exploratory decisions in humans**. *Nature* 2006, **441**:876-879.

46. Daw ND, Niv Y, Dayan P: **Uncertainty-based competition**
• **between prefrontal and dorsolateral striatal systems for behavioral control**. *Nat Neurosci* 2005, **8**:1704-1711.

The authors propose an interesting computational model consisting of two parallel reinforcement learning modules that could be implemented in prefrontal and subcortical basal ganglia circuits. The prefrontal circuit implements 'model-based reinforcement learning', whereas the cortico-basal ganglia circuit including dorsolateral striatum implements 'model-free reinforcement learning'. The former achieves goal-directed behaviors using a tree-search algorithm by simulating the action and state transitions even before actual execution of the action. By contrast, the latter maintains a 'cash' of action value for each state, and it is updated by feedback of reward, step by step through actual action execution.

47. Redish AD: **Addiction as a computational process gone awry**. *Science* 2004, **306**:1944-1947.

48. Schoenbaum G, Roesch MR, Stalnaker TA: **Orbitofrontal cortex, decision-making and drug addiction**. *Trends Neurosci* 2006, **29**:116-124.

49. Roesch MR, Taylor AR, Schoenbaum G: **Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation**. *Neuron* 2006, **51**:509-520.

50. Rudebeck PH, Walton ME, Smyth AN, Bannerman DM,
•• Rushworth MF: **Separate neural pathways process different decision costs**. *Nat Neurosci* 2006, **9**:1161-1168.

This rat lesion study shows that two different types of decision costs — the effort cost and the temporal cost in waiting reward — are represented by different prefrontal circuits — the anterior cingulate cortex and orbitofrontal cortex, respectively. A T-maze with an obstacle in the goal arm and temporal delay for reward delivery were used to examine the effort and temporal costs, respectively. Anterior cingulate cortex lesions affected how much effort rats decided to invest for reward. Orbitofrontal cortex lesions affected how long rats decided to wait for reward. This study is very important in understanding how decision is made depending not only on reward but also on multiple types of cost.

51. Lee D, Conroy ML, McGreevy BP, Barraclough DJ: **Reinforcement learning and decision making in monkeys during a competitive game**. *Brain Res Cogn Brain Res* 2004, **22**:45-58.

52. Soltani A, Lee D, Wang XJ: **Neural mechanism for stochastic behaviour during a competitive game**. *Neural Netw* 2006, **19**:1075-1090.

53. Dorris MC, Glimcher PW: **Activity in posterior parietal cortex is correlated with the relative subjective desirability of action**. *Neuron* 2004, **44**:365-378.

54. Padoa-Schioppa C, Assad JA: **Neurons in the orbitofrontal cortex encode economic value**. *Nature* 2006, **441**:223-226.

55. Singh S: **Transfer of learning by composing solutions of elemental sequential tasks**. *Mach Learn* 1992, **8**:323-339.

56. Wiering M, Schmidhuber J: **HQ-learning**. *Adapt Behav* 1997, **6**:219-246.

57. Parr R, Russell S: **Reinforcement learning with hierarchies of machines.Advances in Neural Information Processing Systems**. MIT press; 1997:1043-1049.

58. Sutton RS: **Planning by incremental dynamic programming**. *Eighteenth International Workshop on Machine Learning; San Mateo, CA*. Morgan Kaufmann; 1991:353-357.

59. Sutton RS, Precup D, Singh S: **Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning**. *Artif Intell* 1999, **112**:181-211.

60. Morimoto J, Doya K: **Hierarchical reinforcement learning for motion learning: learning 'stand up' trajectories**. *Adv Robot* 1999, **13**:267-268.

61. Samejima K, Doya K, Kawato M: **Inter-module credit assignment in modular reinforcement learning**. *Neural Netw* 2003, **16**:985-994.

62. Wolpert DM, Kawato M: **Multiple paired forward and inverse models for motor control**. *Neural Netw* 1998, **11**:1317-1329.

63. Kawato M: **Internal models for motor control and trajectory planning**. *Curr Opin Neurobiol* 1999, **9**:718-727.

64. Haruno M, Wolpert DM, Kawato M: **Mosaic model for sensorimotor learning and control**. *Neural Comput* 2001, **13**:2201-2220.

65. Doya K, Samejima K, Katagiri K, Kawato M: **Multiple model-based reinforcement learning**. *Neural Comput* 2002, **14**:1347-1369.

66. Mena-Segovia J, Bolam JP, Magill PJ: **Pedunculopontine nucleus and basal ganglia: distant relatives or part of the same family?** *Trends Neurosci* 2004, **27**:585-588.

67. Kobayashi Y, Okada K, Inoue Y. **Reward predicting activity of pedunculopontine tegmental nucleus neurons during visually guided saccade tasks.** In *2002 Abstract Viewer and Itinerary Planner Online* (http://sfn.scholarone.com/). Society for Neuroscience; 2002: Program No. 890.5.

68. Haruno M, Kawato M: **Heterarchical reinforcement-learning**
•• **model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning**. *Neural Netw* 2006, **19**:1242-1254.

In this study, a heterarchical reinforcement learning model is proposed and supporting fMRI data are presented. The interplay between the model and the experiments could resolve theoretical difficulties of the plain reinforcement learning algorithm.

69. Haber SN: **The primate basal ganglia: parallel and integrative networks**. *J Chem Neuroanat* 2003, **26**:317-330.

70. Kobayashi Y, Inoue Y, Yamamoto M, Isa T, Aizawa H: **Contribution of pedunculopontine tegmental nucleus neurons to performance of visually guided saccade tasks in monkeys**. *J Neurophysiol* 2002, **88**:715-731.

71. Ng AY, Harada D, Russell S: **Policy invariance under reward transformations: theory and application to reward shaping**. In *Proceedings of the Sixteenth International Conference on Machine Learning: 1999*. Morgan Kaufmann; 1999:278-287.