

METALEARNING, NEUROMODULATION, AND EMOTION

Kenji Doya

doya@isd.atr.co.jp

Information Sciences Division, ATR International
CREST, Japan Science and Technology Corporation
2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan

Introduction

Recent advances in machine learning and artificial neural networks have made it possible to build robots and virtual agents that can learn a variety of behaviors. However, their learning capabilities are strongly dependent on a number of parameters, such as the learning rate, the degree of exploration, and the time scale of evaluation. These parameters are often called *metaparameters* because they regulate the way detailed parameters of an adaptive system change with learning. The permissible ranges of such metaparameters are dependent on particular tasks and environments, making it necessary for a human expert to tune them usually by trial and error. This is why most learning robots and agents to date can only work in the laboratory.

This is in a marked contrast with learning in even the most primitive animals, which can readily adjust themselves to unpredicted environments without any help from a supervisor. This commonsense observation suggests that the brain has a certain mechanism for *metalearning*, a capability of dynamically adjusting its own metaparameters. A candidate of such a regulatory mechanism in the brain is the neuromodulator systems that project diffusely from the brainstem to the cerebral cortex, the basal ganglia, and the cerebellum (see, e.g., Role and Kelly, 1991; Robbins, 1997; Katz, 1999). Most notable of such neuromodulators are dopamine (DA), serotonin (5-HT), noradrenaline (NA), and acetylcholine (ACh) (see Figure 1).

In order to understand the brain mechanism of behavioral learning, the theory of reinforcement learning (see, e.g., Sutton and Barto, 1998), which has been developed for artificial agents that learn to optimize their behaviors through interaction with the environment, can provide a comprehensive computational framework (Doya, 1999).

This paper first reviews basic algorithms of reinforcement learning and introduces a few metaparameters essential for behavioral learning. Then, based on an extensive body of neurobiological data, the paper proposes hypotheses on how these metaparameters are regulated by neuromodulators. The hypotheses allow us to predict and interpret the interactions between neuromodulators, behaviors, and environments. They may also allow us to develop a computational model of emotional states.

Metaparameters of Reinforcement Learning Agents

Central to the theory of reinforcement learning is the value function of a state:

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

where $r(t)$, $r(t+1)$, $r(t+2)$,... denote the reward acquired by following a certain action policy s starting from the initial state $s(t)$. The discount factor $0 < \gamma < 1$ specifies how far into the future rewards are taken into account. The optimal policy that maximizes the above expectation of a cumulative reward is obtained by solving the Bellman equation:

$$V(s) = \max_a [r(s,a) + \gamma V(s'(s,a))]$$

where $s'(s,a)$ is the state reached by taking an action a at state s (Sutton and Barto, 1998). What this equation says is that when taking an action a , both the immediate reward $r(s,a)$ and the future cumulative reward $V(s'(s,a))$ should be taken into account.

The relative merit of taking an action a at state s

$$\delta(s,a) = r(s,a) + \gamma V(s'(s,a)) - V(s),$$

which is called the *temporal difference* (TD) signal, can be used both for action selection and value function learning. A common way of stochastic action selection to facilitate exploration is the Gibbs sampling method:

$$\text{Prob}(a(t)=a_i) = \exp[\beta \delta(s(t),a_i)] / \sum_j \exp[\beta \delta(s(t),a_j)],$$

where β is a parameter that controls the randomness of the action choice, called the inverse temperature.

The estimate of the value function is updated by

$$V(s(t)) := V(s(t)) + \alpha \delta(s(t),a(t))$$

where α is the learning rate.

Control of Metaparameters by Neuromodulators

Based on a large body of neurobiological data and computational modeling studies, the following hypotheses are proposed:

- 1) The dopaminergic system encodes the relative merit δ .
- 2) The serotonergic system controls the time scale of evaluation γ .
- 3) The noradrenergic system controls the inverse temperature β .
- 4) The acetylcholinergic system controls the learning rate α .

Given below is the experimental evidence supporting these hypotheses.

Dopamine for relative reward prediction

The dopaminergic neurons in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) project extensively to the basal ganglia and the frontal cortex. It has been shown in monkey conditioning experiments that dopaminergic neurons respond initially to rewards but later to reward predicting stimuli (Schultz, 1998). Their activity is well characterized by the TD signal δ in the above formulation (Houk et al, 1995; Schultz et al., 1997). The TD signal δ can be used both for action selection and action reinforcement. This is consistent with the fact that dopamine is involved in action selection (e.g., in Parkinson's disease) and action reinforcement (e.g., with addictive drugs).

The amplitude of the TD signal δ also indicates the relevance of the current action and the state change in terms of the future reward. This is in accordance with the fact that dopamine facilitates working memory in the frontal cortex (Sawaguchi and Goldman-Rakic, 1994)

Serotonin for evaluation time scale

In many control tasks, the time scale γ of evaluation determines the weighting for future rewards compared to the immediate action cost. Accordingly, setting γ too small can make “doing nothing” the best solution.

The level of serotonin is higher in awake states and lower during sleep. A high level of serotonin generally stabilize behaviors, while a low level of serotonin can lead to impulsive behaviors (Buhot, 1997). These facts are consistent with the hypothesis that serotonin controls the time scale γ of evaluation. For example, serotonin-enhancing drugs (e.g., Prozac) can reduce depression and anxiety by making immediate negative rewards less significant with a longer evaluation time scale.

Noradrenaline for exploration and optimization

Stochastic action selection is helpful for long-term learning by facilitating wide exploration, while deterministic, greedy action selection is favored in making best use of what has already been learned. Thus the randomness in action selection should be actively tuned in reference to the progress of learning and the urgency of the situation, known as the exploration-exploitation problem.

Noradrenaline has been known to be involved in the control of arousal and relaxation. The noradrenergic neurons in the locus coeruleus (LC) are activated in urgent situations (e.g., with aversive stimuli). It was recently shown in monkeys that the LC neuron activity is correlated closely with the accuracy of action selection (Usher et al., 1999). Furthermore, noradrenaline sharpens the response tuning of neurons by increasing the threshold and the gain (Servan-Schreiber et al., 1990).

These facts suggest that noradrenaline regulates the randomness in action selection, similar to the inverse temperature β in the above formulation.

Acetylcholine for learning rate

Acetylcholine is known to modulate the synaptic plasticity in the hippocampus and the cerebral cortex (Hasselmo, 1999). Depletion of acetylcholine leads to memory disorders like Alzheimer’s disease. These facts point to the possibility that

acetylcholine controls the learning rate α , which determines when to learn something new and when to retain what has been memorized.

Interactions of the Neuromodulators

The appropriate setting of one of these metaparameters depends on the settings of the other metaparameters as well as the environmental setting and the progress of learning. The above hypotheses on the roles of neuromodulators as metaparameters of learning enable us to predict how the levels of neuromodulators should affect each other and change with the environment and the behavior.

For example, the equation for the TD signal

$$\delta(s,a) = r(s,a) + \gamma V(s'(s,a)) - V(s)$$

specifies how the activity of dopamine, δ , should depend on the activity of serotonin, γ . It suggests that serotonin can have both facilitatory and inhibitory effects on dopamine, depending on whether the expected future reward is positive or negative. A higher level of serotonin favors long-term prediction over short-term outcome when they are in conflict. This can explain serotonin's differential effects on dorsal and ventral striatal dopaminergic pathways (De Deurwaerdère et al., 1998) by assuming that the two are involved in long- and short-term reward prediction, respectively.

Conclusion

Neurobiological studies on emotion have so far focused on the role of emotion as the “emergency programs” of behaviors, such as escaping and freezing. However, the role of emotion in modulating cognitive and behavioral learning systems is highly important; many affective or mental disorders occur as a result of the “runaway” of learning systems. Consideration of emotion as a metalearning system enables a novel computational approach in which studies on learning theory, autonomous agents, and neuromodulatory systems can be bound together.

References

- Buhot, M.-C. (1997) Serotonin receptors in cognitive behaviors. *Current Opinion in Neurobiology*, 7:243-254.
- De Deurwaerdère, P., Stinus, L, and Sampinato, U. (1998) Opposite changes of in vivo dopamine release in the rat nucleus accumbens and striatum that follows electrical stimulation of dorsal raphe nucleus: role of 5-HT₃ receptors. *Journal of Neuroscience*, 18:6528-6538.
- Doya, K. (1999) What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, 12:961-974.
- Hasselmo, M.E. (1999) Neuromodulation: acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, 3:351-359.
- Houk, J.C., Adams, J.L., and Barto, A.G. (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J.C. Houk, J.L. Davis, and D.G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*, MIT Press, Cambridge, MA, USA, pp. 249-270.
- Katz, P.S. (1999) *Beyond Neurotransmission: Neuromodulation and its importance for information processing*. Oxford University Press, Oxford, UK.
- Robbins, T.W. (1997) Arousal systems and attentional processes. *Biological Psychology*, 45:57-71.
- Role, L.W. and Kelly, J.P. (1991) The brain stem: Cranial nerve nuclei and the monoaminergic systems. In E.R. Kandel, J.H. Schwartz, and T.M. Jessel, (Eds.), *Principles of Neural Science*, third edition, Appleton & Lange, Norwalk, CT, USA pp. 683-699.
- Sawaguchi, T. and Goldman-Rakic (1994) The role of D1 dopamine receptor in working memory: Local injections of dopamine antagonists into the prefrontal cortex of rhesus monkeys performing an oculomotor delayed-response task. *Journal of Neurophysiology*, 71:515-528.
- Schultz, W., Dayan, P., and Montague, R.P. (1997) A neural substrate of prediction and reward. *Science*, 275:1593-1599.

- Schultz, W. (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1-27.
- Servan-Schreiber, D., Printz, H., and Cohen, J.D. (1990) A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, 249:892-895.
- Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA.
- Usher, M., Cohen, J.D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999) The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283:549-554.

Figure 1

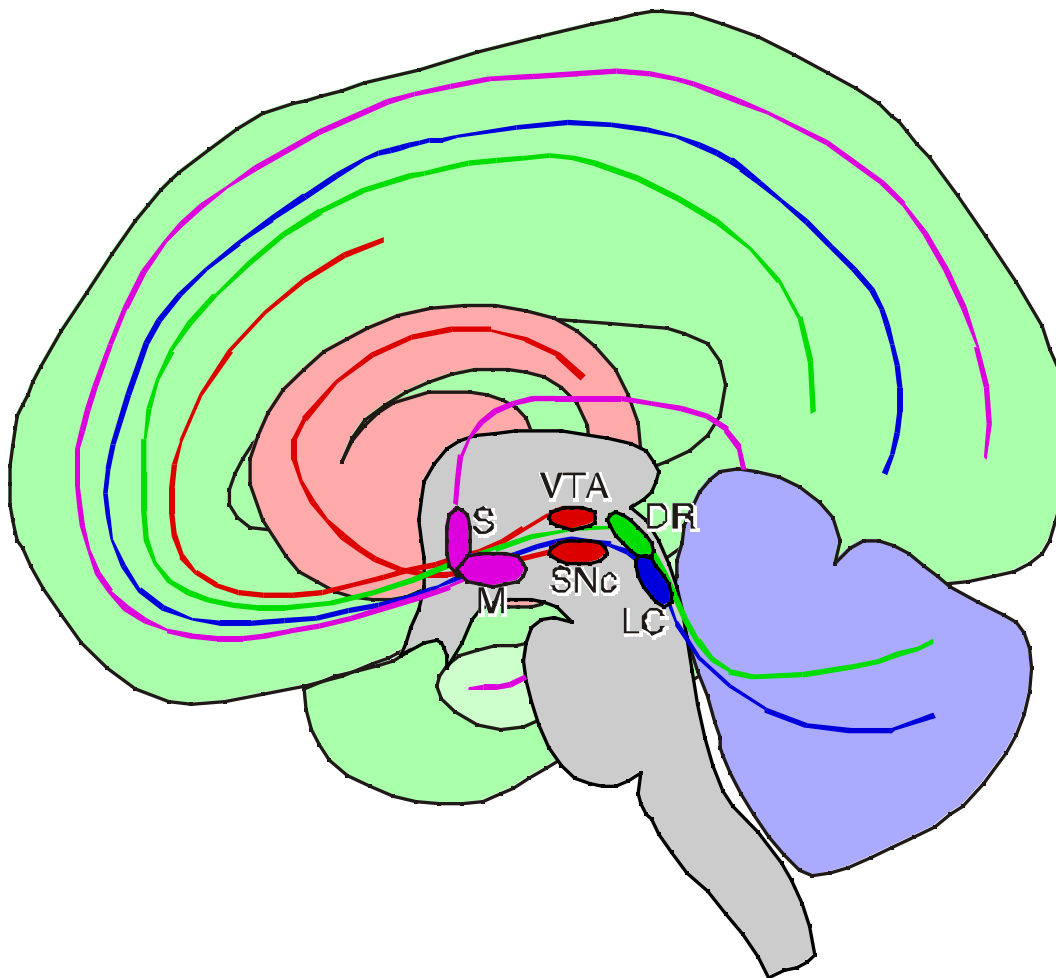


Figure 1: The projection of neuromodulators from the brain stem to the cerebral cortex, the basal ganglia, and the cerebellum. Dopaminergic projection from substantia nigra, pars compacta (SNc) and ventral tegmental area (VTA). Serotonergic projection from dorsal raphe nucleus (DR). Noradrenergic projection from locus coeruleus (LC). Acetylcholinergic projection from septum (S) and Meynert nucleus (M).