Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Multiplexing signals in reinforcement learning with internal models and dopamine

Hiroyuki Nakahara

A fundamental challenge for computational and cognitive neuroscience is to understand how reward-based learning and decision-making are made and how accrued knowledge and internal models of the environment are incorporated. Remarkable progress has been made in the field, guided by the midbrain dopamine reward prediction error hypothesis and the underlying reinforcement learning framework, which does not involve internal models ('model-free'). Recent studies, however, have begun not only to address more complex decision-making processes that are integrated with model-free decision-making, but also to include internal models about environmental reward structures and the minds of other agents, including model-based reinforcement learning and using generalized prediction errors. Even dopamine, a classic model-free signal, may work as multiplexed signals using model-based information and contribute to representational learning of reward structure.

Addresses
Laboratory for Integrated Theoretical Neuroscience, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Corresponding author: Nakahara, Hiroyuki (hn@brain.riken.jp)

Current Opinion in Neurobiology 2014, **25**:123–129

This review comes from a themed issue on **Theoretical and computational neuroscience**

Edited by **Adrienne Fairhall** and **Haim Sompolinsky**

For a complete overview see the Issue and the Editorial

Available online 23rd January 2014

0959-4388/$ – see front matter,

http://dx.doi.org/10.1016/j.conb.2014.01.001

Introduction

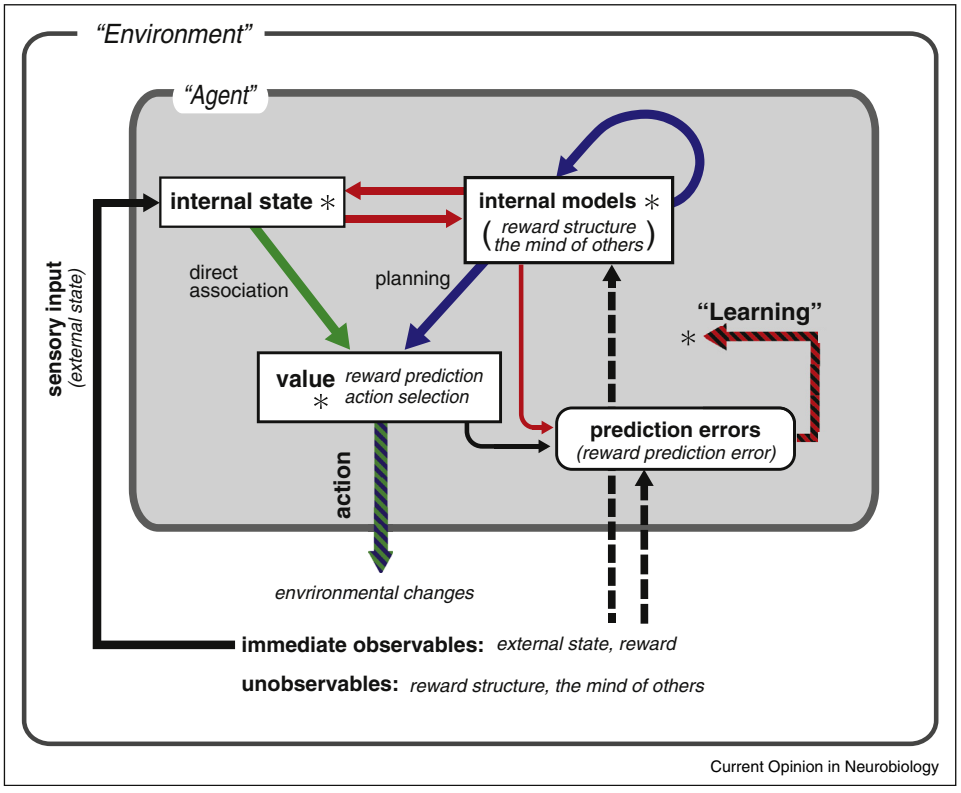
To survive, humans and animals must learn to adapt to their environment to obtain needs such as food and water, the most primary 'rewards.' They must make decisions about their behavior and must continue to adjust their decisions in an ever-changing world. This is a critical ability, making decisions guided by rewards and associated learning, which is termed value-based decision-making. Here, we examine a particular direction of recent progress in the field of value-based decision-making, namely learning and using internal models of the environment (often in relation to rewards, i.e., environmental 'reward structures') to make adaptive decisions. This progress stems from a major breakthrough in the field: the midbrain dopamine (DA) reward prediction error (RPE) hypothesis based on a reinforcement-learning (RL) framework, in particular a 'model-free' RL algorithm called temporal difference (TD) learning. The profound impact is, in the broadest sense, that RL is introduced to the field as a framework that links behavior and neural responses through key computations, allowing us to access subjective variables and internal processes. The model-free RL involves no internal models beyond a reward-contingency association with the most recent sensory input, and is thus particularly transparent in examining neural responses with respect to the computations that have been used successfully in many studies [1]. Encouraged, a number of recent studies have begun to investigate more complex decision-making computations, that is, learning and using internal models [2,3•,4–6]. A selected set of these studies are reviewed here from two perspectives (Figure 1 and Box 1), the first relating to the set-up and variables of RL and the second relating to how multiplexed DA signals may help reward-based learning with the potential use of internal models.

Reinforcement learning and use of internal models

Here, following a brief summary of RL and the model-free RL, we highlight three lines of research related to the use of internal models: studies on sequential decision-making (model-based RL), studies on value in foraging, and studies on learning signals generalized from the RPE.

Value and RPE are key variables in RL, most simply described using a model-free RL without involving sequential decision-making. An 'agent' observes a 'state' in a given environment and selects an 'action' among those available to obtain a 'reward.' Value is the predicted reward that guides action selection, and the RPE (the difference between predicted and actual reward) functions as a learning signal to improve future predictions (and selections). For instance, prediction learning is aimed at reducing error. In a model-free framework, only prediction and selection are required to make decisions and are learned as a direct association to the state, without considering any internal models of the environment (Figure 1, green arrow). This description is completed when the issue of 'time' is included, as in TD learning. Time can be appreciated as real-time, that is, the passage of time from observing the state to obtaining the reward [7–9]. Over a longer timescale, that is, for sequential decision-making, the agent may rather go through several states, deciding an action per a state, to collect rewards at visited states. To maximize reward acquisition, the agent

Figure 1



Schematic for computations of value-based decision-making. An agent (i.e., a decision-maker) receives information of immediate observable elements in the environment via sensory input (external state). Through experience (dashed arrow), however, the agent also constructs internal models that collectively reflect unobservable elements in the environment such as reward structures and the minds of others. Value guides decisions, involving reward prediction and action selection. Model-free RL computes value by direct association to the internal state (green arrow). Here, the model-free RL with the 'sensory input' state definition further equates sensory input to the internal state. Model-based RL computes value by using internal models (blue arrow.) Internal models 'simulate' the external world, possibly approximately or partially (recursive blue arrow) and then make decisions with regard to planning. In complex cases such as foraging, a recursive loop may exist between internal models and value before a final decision is made and overt action is taken (not shown in the figure). The red arrows emphasize the issues described in this article: the internal state is an integral of sensory input with information from internal models, which may lead to better model-free RL algorithms than the original model-free RL; the information from internal models can also be used to produce generalized prediction errors; and generalized prediction errors are used to learn internal state, internal model, and value (*).

should pursue not only the immediate reward, but also ensure a balance between immediate and future rewards. TD learning addresses these issues, forming a value that takes into account the balance over states (time) and TD error, which is a sophisticated version of the RPE. These descriptions may be appreciated more, being explicitly linked to mathematical terms, even in brief: first, denoting time by t, value (given a state s_t) is defined by $V(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$, where r refers to reward at specific times, and γ (being $0 \leq \gamma \leq 1$) is the discount factor, which takes care of the balance; and second, using the constraint that $V(s_t) = r_t + \gamma V(s_{t+1})$ should hold on average when learning converges, TD error is defined by $\delta(s_t) = r_t + \gamma V(s_{t+1}) - V(s_t)$. Then, we have the DA RPE hypothesis, which posits that DA activity signals TD error, based on original experimental observations (e.g., [10]), combined with a model-free RL or a TD that addresses the passage of time [7,8]. The TD learning model has two assumptions: that only information regarding the current state, without past information, is sufficient to make optimal decisions; and that the most recent sensory events and their time traces are the 'states.' The first assumption makes theoretical analyses easier even though it could be violated in the real world (see the next section), and the second is a simplification, so that in a given experimental setting, we usually consider value to be generated by the association to such 'sensory event' states.

The first group of studies refer to model-based RL, another class of RL algorithms related to sequential decision-making [1,3•], and demonstrate the power of model-based RL approaches particularly when compared with the model-free RL of the DA RPE hypothesis. In general, a model-based RL system learns, at least approximately, internal models about either or both the