Neuroscience Research xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Neuroscience Research

journal homepage: www.elsevier.com/locate/neures

Update article

Learning to represent reward structure: A key to adapting to complex environments

Hiroyuki Nakahara [a,*], Okihide Hikosaka [b]

[a] Laboratory for Integrated Theoretical Neuroscience, RIKEN Brain Science Institute, Wako, Saitama 351-0198, Japan
[b] Laboratory of Sensorimotor Research, National Eye Institute, Bethesda, MD 20892, USA

ARTICLE INFO

Article history:
Received 25 April 2012
Received in revised form 31 August 2012
Accepted 6 September 2012
Available online xxx

Keywords:
Reward
Dopamine
Reinforcement learning
Decision
Value
Salience
Structure

ABSTRACT

Predicting outcomes is a critical ability of humans and animals. The dopamine reward prediction error hypothesis, the driving force behind the recent progress in neural "value-based" decision making, states that dopamine activity encodes the signals for learning in order to predict a reward, that is, the difference between the actual and predicted reward, called the reward prediction error. However, this hypothesis and its underlying assumptions limit the prediction and its error as reactively triggered by momentary environmental events. Reviewing the assumptions and some of the latest findings, we suggest that the internal state representation is learned to reflect the environmental reward structure, and we propose a new hypothesis – the dopamine reward structural learning hypothesis – in which dopamine activity encodes multiplex signals for learning in order to represent reward structure in the internal state, leading to better reward prediction.

1. Introduction

Outcome prediction, along with action selection based on the prediction, underlies motivated and reward-oriented behavior or value-based decision making (Hikosaka et al., 2006; Montague et al., 2006; Rangel et al., 2008; Schultz, 1998). To maximize the gain of outcomes, one should make value-based decisions, not only aiming for the immediate outcome but rather making a balance of outcome predictions between the immediate and temporally distant future. One should also be able to learn appropriate value-based decisions through experience in order to behave adaptively to different circumstances. Finally, one should generate decisions based on the information that is represented in the input (state representation), and this final aspect is the focus of this article.

The reinforcement learning (RL) framework, and temporal difference (TD) learning in particular, can offer a quantitative solution for this balancing and learning. This characteristic has made the theory influential in the recent expansion in our understanding of the value-based decision making process and the underlying neural mechanisms (Montague et al., 1996; Schultz et al., 1997). RL was originally developed in mathematical psychology and operation research (Sutton and Barto, 1990) and remains an active research area in computer science and machine learning (Sutton and Barto, 1998). The intrinsic strength of RL theory is its clear formulation of the issues mentioned above, which can stand on its own with its mathematically defined elements, even without a relationship to any physical entities. However, it is not its intrinsic strength but its clear set of assumptions that made RL influential in the field of neural value-based decision making. These assumptions made it possible to map between the well-defined elements of RL and the underlying neural substrates, thereby allowing us to understand the functions of neural activity and the roles of neural circuits under this theory. A marked example is an ingenious hypothesis about dopamine phasic activity as a learning signal for TD learning (called TD error), which is the strongest example of mapping to date, and is thus a critical driving force behind the progress in this field (Barto, 1994; Houk et al., 1994; Montague et al., 1996; Schultz et al., 1997).

The latest findings from the vanguard of this field, however, have begun to suggest the need for a critical revision of the theory, which is related to the underlying assumptions that map RL to neural substrates and requires a reconsideration of state representation. After providing a brief sketch of RL theory and its assumptions, we first clarify the reward prediction and error of the hypothesis. Using experimental and computational findings on dopamine activity as a primary example, we discuss that the prediction and associated action selection can be significantly enhanced if the structure of rewards are encoded in the state representation for those functions. We propose a new hypothesis in which dopamine activity encodes

* Corresponding author. Tel.: +81 48 467 9663; fax: +81 48 467 9643.
E-mail addresses: hiro@brain.riken.jp (H. Nakahara), oh@lsr.nei.nih.gov (O. Hikosaka).

0168-0102/$ – see front matter © 2012 Elsevier Ireland Ltd and the Japan Neuroscience Society. All rights reserved.
http://dx.doi.org/10.1016/j.neures.2012.09.007

multiplexed learning signals, representing reward structure and leading to improved reward prediction.

2. Background: the reinforcement learning framework

To understand the intrinsic strength of RL, or TD learning, it is useful to first present its mathematical ingredients (Sutton and Barto, 1998) but in an intuitive manner and separately from the assumptions used to map RL to neural substrates. In the TD framework, an abstract entity is first considered that receives an input and then produces an output; this input–output pair causes a transition to the next input, deterministically or probabilistically, and the entity produces an output when given the next input, so that the process continues. Importantly, at each transition, the entity receives a real number, or a numeric, which the entity prefers to be larger. The entity's primary interest is to balance, improve, and ideally maximize the gain of the numeric over the transitions. These are the key concepts of the framework, which can be defined as definite mathematical notions once their definitions, assumptions, and constraints are refined, which we do not attempt here.

The numeric prediction construct and its learning signal are at the heart of the formulation, and they are called the value function and TD error, respectively. The value function defines a solution for the balancing problem, while TD error provides a means for learning ability. The value function solves the balancing problem by summing the numeric over the transitions with the so-called discount factor and thereby discounting the future numeric more strongly; the value of an input, e_i, is given by $V(e_i) = r_i + \gamma r_{i+1} + \gamma^2 r_{i+2} + \ldots$, where r_j refers to the numeric in transition at input e_j and γ is the discount factor, where $0 \leq \gamma \leq 1$. Even if the value function is defined as such, its actual value is unknown, and it is thus learned in the framework as an approximate value. This learning takes advantage of the function's specific form; once it is performed well, $V(e_i) = r_i + \gamma V(e_{i+1})$ should hold on average, and it is thus not well established if both ends of the equation differ. Therefore, it uses the difference as a learning signal or TD error, $\delta(e_i) = r_i + \gamma V(e_{i+1}) - V(e_i)$, as the name indicates (i.e., the temporal difference of values between two consecutive inputs). It adjusts the value in the same direction as the error (either positively or negatively) and also proportional to the magnitude of the error. Using TD error, the entity similarly solves another important issue: learning about output selection or which output to choose with an input. Although there are other types, the formulation sketched here is the most basic type used to solve numeric prediction and output selection in parallel by learning. The majority of studies adopt a linear form for the two functions, which we also follow. By way of an example, the linear-form value function is a multiplication of a vector representation of a given input with a weight vector, and it is improved during learning by changing the weight vector in reference to the input vector.

A simple example of this formulation is that the entity can be regarded as an agent (human or animal) in an environment. The input is a state of the environment and is thus called state; the output is a way for the agent to influence the environment and is thus called action; and the output selection is called action selection. The numeric is an affectively important outcome of the agent, such as reward, and the value function corresponds to reward prediction. Although this example is certainly useful, as it is a major origin of the formulation and often used in the literature (as it is below), understanding the abstract notion is crucial (Sutton and Barto, 1998). In particular, this example is misleading if it is taken to imply that the TD learning framework demands that the entity must be a "whole" agent, so that the state of the environment must be the input to the entity. Instead, the abstract notion defines only that a given entity should implement functions of TD learning, or the reward prediction and action selection, given its inputs. Specifically, an entity can be a part of the agent; when considering that TD learning is a part of brain function, it is more appropriate to consider that the entity is a functional part of the brain, so that the input to the entity should be based not only on the input from the environment, but also on the information generated internally in the brain (Singh et al., 2005).

3. Versatility and limitations of the reward prediction error hypothesis

The hypothesis that dopamine (DA) phasic activity corresponds to TD error, called the reward prediction error hypothesis, has facilitated transparent mapping between the computational notions of TD and the underlying neural substrates (Barto, 1994; Houk et al., 1994; Montague et al., 1996; Schultz et al., 1997). This transparent mapping has helped to drive the field's progress since the proposal of this hypothesis, and it has been observed as the correspondence between "canonical" DA responses and the TD error of the hypothesis (Schultz et al., 1997). DA exhibits phasic activity in response to the delivery of an unexpected reward. Once the pair of a reward-predicting cue (CS) and reward (US) has been presented with sufficient repetition (as in a Pavlovian conditioning task), DA displays phasic activity to the CS but ceases to respond to the US; if the US is omitted, DA demonstrates a suppressive response at the time of US omission. Furthermore, several other notable characteristics of DA have made the hypothesis more plausible and attractive (Schultz, 1998), only a few of which are now mentioned. DA is known to act as a modulator of synaptic plasticity, thus being attractive as a learning signal (Reynolds and Wickens, 2002). A major proportion of DA neurons originating from the midbrain, especially the ventral tegmental area (VTA) and substantianigra pars compacta (SNc), have massive, diffuse projections not only to the basal ganglia (e.g., striatum and nucleus accumbens) but also to the overall cerebral cortex; such a projection pattern seems ideal to concordantly modulate the functions of different areas in TD learning. Given the available experimental evidence when the hypothesis was proposed, DA phasic activity was considered to be largely homogeneous in the VTA and SNc, except for some minor variability in the responses ("noisy" responses). Thus, assigning an important, single role to DA made sense, and TD error is quite attractive as a unifying theory, especially given the well-documented but still sought-after roles of DA in motivated and addictive behaviors.

Two assumptions of the hypothesis enabled transparent mapping for clarity (Schultz et al., 1997). The first is a state assumption. The hypothesis practically uses the agent-environment example, described in the previous section, as the basis for its construction. Accordingly, the state is taken to be the equivalent of a momentary external event or the event's sensory input to the agent (Fig. 1A); in the CS–US case described above, the CS itself is a state. The second is a time assumption. In the original, mathematical setting, although there are transitions between the inputs, they are, in principle, not related to the physical passage of time (Nakahara and Kaveri, 2010); however, in the real world, there are often intervals between external events. For example, after the brief presentation of a CS, a time delay may occur before the next clear external event or US. In the hypothesis, time is divided into small constant time bins (e.g., 200-ms bins) and each bin corresponds to each state. For bins with clear external events, the states correspond to the events. For bins with no external events, state representations are filled in, which are assumed to be generated by the most recent past event as a time trace (called stimulus-time compound representation) (Sutton and Barto, 1990). For example, it is the time assumption that allows the TD error of the hypothesis to indicate a suppressive response to

Please cite this article in press as: Nakahara, H., Hikosaka, O., Learning to represent reward structure: A key to adapting to complex environments. Neurosci. Res. (2012), http://dx.doi.org/10.1016/j.neures.2012.09.007

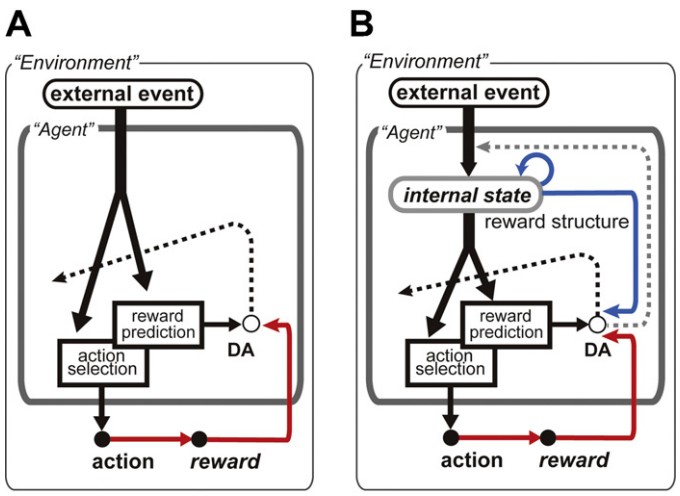

Fig. 1. (A) Reward prediction learning according to the reward prediction error hypothesis. Dopamine (DA) activity encodes the reward prediction error, which is the difference between the actual reward (red arrow) and the predicted reward that is expected based on momentary external event information (black arrow), and then contributes to learning in reward prediction and action selection (indicated by the dashed line intersecting the external state inputs to the two functions). Under the reward prediction error hypothesis, DA activity is considered to encode a specific reward prediction error signal $\delta(e_i) = r_i + \gamma V(e_{i+1}) - V(e_i)$ wherein the input e_j is equivalent to external events (or their time traces), say E_i, and then $e_i = E_i$ in the hypothesis. (B) Schematic of a new hypothesis, the reward structural learning hypothesis. With input reflecting the structure of the rewards (blue arrow toward DA), DA activity encodes multiplexed learning signals: signals for learning to represent the reward structure in the internal state (gray dashed arrow) and improved reward prediction error signals, i.e., signals for learning better reward prediction and action selection (black dashed arrow). Here, "internal state" in the figure refers to the neural, internal representation acquired by the reward structural learning, which is then used as input to generate reward prediction and action selection. Under the reward structural learning hypothesis, DA activity may encode two types of signals. One type of signal is a reward prediction error signal (mostly in the black dashed arrow but possibly also in the gray dashed arrow). The input e_j for $\delta(e_i) = r_i + \gamma V(e_{i+1}) - V(e_i)$ is not necessarily E_i if, say, s_i (i.e., e_i=s_i); s_i is learned to better reflect reward structure, e.g., taking account of past and future events, actions and outcomes. The other type of signal facilitates the learning of s_i (in the gray dashed arrow). For example, a variety of DA signals discussed in the text, e.g., "salient", "alerting", "initiating", "uncertainty", "information-seeking", and "history-dependent" signals, could underlie this type of learning signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

an unexpected reward omission (as the TD error of the bin changes with no reward occurrence), similarly to the canonical DA response in that case.

Together with these assumptions, the overall setting of the TD learning framework, reviewed in the previous section, determines the two crucial characteristics of reward prediction and its error postulated by the hypothesis (Fig. 1A). First, the prediction and error are produced reactively to external events. In essence, external events are the states of the TD in the hypothesis. Therefore, the reward prediction of the hypothesis depends directly on the most recent external event, or indirectly via a time trace triggered by the event (before the next external event happens). As both reward prediction and action selection are computed as soon as the state arrives (e.g., multiplication between the state and weight vectors in the linear form), their outputs are produced reactively to the momentary external event (or the momentary time trace of the event). The TD error of the hypothesis is also produced reactively to such states because it is computed by using the actual outcome and the values of the "current" and "next" states only after observing the "next" state.

Second, the predictive nature of reward prediction and error (and also action selection) is limited in a specific way under the hypothesis. Generally, in TD learning, while reward prediction and action selection acquire a predictive nature via learning with TD error, TD error sets a limit on the prospective information that reward prediction and action selection can access during learning, and the predictive nature of TD error comes from being generated as the temporal difference of reward predictions or value function that is defined to sum outcomes over transitions. As the hypothesis assumes external events to be states of the TD, the state representation limits the information available as only that contained in the momentary external event (or momentary time trace). Consequently, the reward prediction of the hypothesis could be learned and generated to an extent that is based on the information provided by the momentary external event, accordingly inducing a specific TD error.

Thus, the essential elements of the hypothesis include the fact that the states are external events, and the corresponding reward prediction and error. These are frequently regarded in the field as a default value-based decision-making process. Under the hypothesis, DA activity is the specific reward prediction error, i.e., the signal for learning the reward prediction of the default process. Moreover, in the literature, further neural functions are often investigated or discussed as additional components to the default process.

Therefore, the proposition that DA activity encodes the error of the default process needs to be critically examined. As the default process is defined by the choice of the states as external events, a representational question is central to this examination. The reward prediction error hypothesis practically abandons this question, as it equates momentary external events (or their time traces) with "internal state representation", which serves as input for generating reward prediction and action selection (Fig. 1A).

4. Reward structure useful for prediction: does dopamine activity reflect reward structure?

Do DA neurons really encode the specific reward prediction error (the specific TD error) of the reward prediction error hypothesis? In fact, we found that DA activity can encode the reward prediction error better than the specific error of the hypothesis (Nakahara et al., 2004). Critically, this prediction error encoded by DA activity is the error that could be generated only when the structure of rewards was acquired in internal state representation.

The study aimed to address whether DA activity, a putative reward prediction error signal, could access information beyond that of momentary external events (or their time traces). In the study, an instructed saccade task was used in which correct saccades to instructed cues were accompanied with different outcomes (in short, reward or no reward). A pseudo-random procedure was used to determine a sequence of task trials; the rewarded and non-rewarded cues were randomly permuted within each sub-block of trials so that the pre-determined, average probability of the rewarded and non-rewarded cues was maintained within a pre-fixed number of trials, or a block of trials. This procedure induced a reward probability that was embedded in the past sequence of outcomes over trials. This history-dependent reward probability changed over trials, and it was a more precise measure for the prediction of coming cues (or outcomes) in the next trial than the average reward probability. The reward prediction and TD error by the reward prediction error hypothesis would correspond to those produced using the average reward probability. On the contrary, we found that the phasic response of DA to the instruction cue matched the TD error using the history-dependent reward probability, which could be modeled by adding the representation of the sequential reward structure as internal states to the TD learning framework. The DA response emerged only after extensive experience with the task. Additionally, the findings were somewhat concordant with the findings of other studies (Bayer and Glimcher, 2005;

Please cite this article in press as: Nakahara, H., Hikosaka, O., Learning to represent reward structure: A key to adapting to complex environments. Neurosci. Res. (2012), http://dx.doi.org/10.1016/j.neures.2012.09.007

Bromberg-Martin et al., 2010b; Enomoto et al., 2011; Satoh et al., 2003). Overall, they demonstrated that DA activity can encode a better TD error, as if an appropriate state representation is acquired beyond the external events and then used for reward prediction.

Indeed, similarly to the case described above, reward prediction and/or action selection can be improved in many situations by a better state representation than those used in the value-based decisions of the reward prediction error hypothesis. The above case is only an example of the situations in which one should adjust the reward prediction considering the sequence of past outcomes, rather than just to try and learn the reward expectation given the momentary external cue; for example, in foraging, one should adjust the expectation as one acquires fruit from the same tree (Hayden et al., 2011). More generally, we can consider a classification of such situations based on what types of information may be useful to be included in the state representation (Table 1). First, configurational information within a momentary event is potentially beneficial, compared to cases in which the event is encoded plainly without representing the configuration. Different coordinate-specific representations may lead to different learning speeds (Hikosaka et al., 1999; Nakahara et al., 2001). Such within-the-moment information could also exist in other factors. Encoding the relationship among rewards in the state is potentially useful (Acuna and Schrater, 2010; Gershman and Niv, 2010; Green et al., 2010). The action could also be represented at different levels, e.g., effector-independent versus effector-specific, and this would result in different learning speeds or differently converging selection (Bapi et al., 2006; Gershman et al., 2009; Nakahara et al., 2001; Palminteri et al., 2009). Second, useful information could also exist in the temporal sequence of these factors. As described above, DA activity or TD error could benefit from encoding information from past outcomes into the state (Nakahara et al., 2004). Similarly, encoding the information of a sequence or any combination of external events, actions, and outcomes, in some ways or even partially, can be beneficial for improving reward predictions (Kolling et al., 2012). Action selection can similarly benefit; an action may be selected more accurately by taking into account a series of events before or even after the momentary external event (Hikosaka et al., 1999; Nakahara et al., 2001), e.g., sequence-dependent action or motor control, possibly using different coordinate-specific representations.

5. Dopamine activity for learning the reward structure

We thus suggest that learning the reward structure is indispensable for learning the reward prediction and propose a new hypothesis, termed the dopamine reward structural learning hypothesis (Fig. 1B), in which DA activity encodes multiplexed learning signals. These signals include those for learning the structure of a reward in internal state representation ("representation learning"; gray dashed arrow in Fig. 1B), together with signals for learning to predict the reward ("prediction learning"; black dashed arrow in Fig. 1B), as signals of an improved reward prediction error supported by representation learning.

Several findings support the view that a variety of DA activities is helpful for learning the reward structure. First, DA activity modulates the cortical re-representation of external events, or re-mapping of auditory cues (Bao et al., 2001), and, more broadly, is considered to play a major role in reward-driven perceptual learning (Seitz and Dinse, 2007; Zacks et al., 2011). Second, a subset of DA activity can respond in an excitatory manner to aversive stimuli (CS and/or US) in a similar way to appetitive stimuli, which is opposite to the presumably inhibitory response posited by the reward prediction error hypothesis. This observation was noted in behaving awake monkeys (Joshua et al., 2009; Matsumoto and Hikosaka, 2009) and in rodents (Brischoux et al., 2009; Cohen et al., 2012). Although further delineation is required (Frank and Surmeier, 2009; Glimcher, 2011), such DA activity may encode the saliency signal (Bromberg-Martin et al., 2010b; Matsumoto and Hikosaka, 2009), which is important for knowing what information is crucial, even though it does not code for the "direction" of importance (i.e., being positive or negative for appetitive and aversive stimuli, respectively, as the TD error does). Third, a subset of DA activity can also encode signals that alert or initiate a sequence of external events that are evoked by an initiating external event or aligned with a self-initiated motor act (Bromberg-Martin et al., 2010b; Costa, 2011; Redgrave and Gurney, 2006). A group of DA activities is hypothesized to contain a novelty signal or signals for exploration (Daw et al., 2005; Kakade and Dayan, 2002). Indeed, DA activity is also shown to encode "uncertainty" signals (Fiorillo et al., 2003) or "information-seeking" signals (Bromberg-Martin and Hikosaka, 2009). These signals can be important for forming a representation that reflects a useful portion of external events. Fourth, a subset of DA activity has been shown to add information on the action choice or task structure to the reward prediction error (Morris et al., 2006; Roesch et al., 2007), suggesting that an interplay between representation learning and prediction learning is reflected in DA activity. Fifth, even DA tonic activity was found to be modulated by information on this relationship within a block of trials and even between blocks (Bromberg-Martin et al., 2010a), further supporting the reflection of temporal structure information in DA activity. Thus, these findings indicate that DA activity is not quite as homogeneous as originally thought or implicitly presumed in the reward prediction error hypothesis, but it is rather heterogeneous. Notably, all of these DA activities described above can assist representation learning in principle.

Representation learning yields better prediction learning than that described in the reward prediction error hypothesis. Once representation learning enriches the internal state representation with information on the reward structure, reward prediction and action selection can be significantly improved, even if they are generated reactively. The reward prediction error is also naturally improved, as it uses better reward predictions (Nakahara et al., 2004). Additionally, the error of the reward structural learning hypothesis can acquire a proactive nature because it can reflect changes in internal states, or temporal evolution of internal states, which can be distinct from the external events (Nakahara et al., 2001). This feature also applies to reward prediction and action selection. Even with the same external event, differences in the internal state could allow those functions to produce different outputs (Doya, 1999; Nakahara et al., 2001). During time delays with no explicit external events, the internal state could allow those functions to be evoked before the actual occurrence of an external event, leading to anticipatory reward prediction and action.

Representation learning is multi-faceted: it works to synthesize useful information from different sources in order to support and improve reward prediction. Sequential information, or information on task structure, can, in principle, be utilized in two ways (Hikosaka et al., 2006; Nakahara et al., 2004; Ribas-Fernandes et al., 2011): retrospectively and prospectively with respect to a momentary external event. In the retrospective scheme, the internal state should compactly represent information on preceding event sequences in addition to the event information via learning. In the prospective scheme, it should include the information on future event sequences that have not yet occurred. This can be performed either as the direct learning of future events in the representation (Dayan, 1993) or as an active process (recursive blue arrow with internal state in Fig. 1B). One mechanism for the prospective scheme using the active process would be to use a recall that starts after the event, evoking future likely events (also actions or outcomes) and imposing their information into the representation.

Please cite this article in press as: Nakahara, H., Hikosaka, O., Learning to represent reward structure: A key to adapting to complex environments. Neurosci. Res. (2012), http://dx.doi.org/10.1016/j.neures.2012.09.007

Table 1
Structure of rewards, useful to be encoded in state representation.

Factors	*Examples*
Class: **Configurational** Acquiring information latent within a moment into state representation.	
External event—association of a pattern or subset of an external event with the outcome or appropriate action.	A specific visual pattern configuration may be a key for reward prediction (e.g., in board games). Encoding the configuration in the state can drastically change the learning and execution of prediction and action selection.
Reward—relationships of reward delivery, or their absence, with actions or events.	Reward delivery to one choice may imply reward absence to the other (e.g., among numbers in a roulette game) or could be independent of the other (among people). Encoding the dependence or independence in the state may drastically change learning and execution.
Action—appropriate levels to choose an action, more specific or general.	Action to indicate choosing an option on the "left" can be expressed in different specific ways (e.g., by hand, eye, or chin), but also in a general form as being "left." The appropriate level encoding the action in the state changes the TD learning of action selection.
Class: **Sequential** Acquiring information over moments into state representation.	
Retrospective—adding information of a sequence of past events, rewards, and/or actions in a compact form, and typically recent past ones, to the information of a momentary external event.	Foraging among fruit trees. One should not keep increasing the expectation of rewards on a tree as one collects fruit from the tree, but rather decrease the expectation because obtaining more fruit from the tree means less remaining fruits. TD learning with momentary external events (e.g., looking at the tree) as the states cannot immediately take account of such a reward structure, as its reward prediction is learned to be an average (discounted) value of fruit with the state.
Prospective—adding information of likely future events, outcomes, or actions to the information of a momentary external event.	Moving to where a puck would go. In ice hockey, we should not just go to where a puck currently is, but rather move, considering where a puck is likely to be. By contrast, TD learning with momentary external events as states can learn reward prediction and action selection only reactively with respect to the events.

Other neural functions that are debated in reference to the original setting of the reward prediction error hypothesis are mostly related to this type of recall because those functions are defined to invoke additional processes after the event, beyond the default value-based decision-making process. For example, active recall after the event has also been applied to extract configurational information as a complementary process to the default process (Courville et al., 2006; Daw et al., 2006; Gershman and Niv, 2010; Green et al., 2010; Rao, 2010; Redish et al., 2007) (see below). Another mechanism for the prospective scheme would be anticipatory recall before the event to encode future likely events (along with actions or outcomes) in the representation. This mechanism would make information on future events available before any event starts, therefore rendering value-based decisions very flexible.

While DA activity would exert effects on representation learning primarily through DA modulation of synaptic plasticity (gray dashed arrow in Fig. 1B), it may, additionally, directly affect the internal state representation with its effect on membrane excitability (for which the gray dashed arrow in Fig. 1B could additionally be considered to represent direct modulation). For example, DA activity may change or gate that which is maintained as the internal state, e.g., in working memory or sustained neural activity (Gruber et al., 2006; Montague et al., 2004; Todd et al., 2009). In concert with the prospective mechanism and the anticipatory recall discussed above, the immediate effect of DA activity on the internal states may provide an additional mechanism to adaptively select the internal states. Presumably, the DA-mediated synaptic learning mechanism is better equipped to extract useful information by superimposing reward-related events over a long time, while the DA-mediated immediate mechanism is equipped to adjust to changes in the environment over a short time. In a broader perspective, the immediate mechanism is also a part of representation learning, i.e., setting an improved state for reward prediction and action selection.

Our dopamine reward structural learning hypothesis provides important insight into a dichotomy of decision making: the so-called model-free and model-based RL mechanisms (Acuna and Schrater, 2010; Balleine et al., 2008; Daw et al., 2011, 2005; Dayan and Niv, 2008; Doya, 2007; Funamizu et al., 2012; Gläscher et al., 2010; Suzuki et al., 2012; Wunderlich et al., 2012). In these studies, both mechanisms use the external events as the state in the same way that is assumed for the reward prediction error hypothesis. However, they differ in what they are designed to learn and how they are designed to makes decisions. The model-free RL is the default process described earlier. It learns values that are directly associated with states (which are mediated by DA activity) and then makes decisions by comparing the values. On the other hand, the model-based RL directly learns the transitions across states and the ways in which the reward is given in the transition, and it makes decisions by simulating future changes in the environment and comparing the simulated values. Thus, the model-free RL is more economical in computational labor, but it is less flexible (or 'habitual'), whereas the model-based RL requires heavier computations, but it is more flexible. By contrast, our hypothesis suggests that internal states, acquired by representation learning, would provide a better default process, and this default process can work as an improved model-free RL mechanism. Compared with the 'original' model-free RL, the new model-free RL may be more optimal, compactly representing useful information beyond the immediate past event and yielding to better reward predictions, for example. It may also be more flexible, possibly combined with the prospective mechanism or anticipatory recall. On the other hand, it involves heavier learning, which is learning the internal state. Compared with the 'original' model-based RL, the new model-free RL can work faster and more preemptively in decision making and may be potentially more economical. However, it may not achieve the same ultimate degree of optimality and flexibility as the original model-based RL could because the original model-based RL involves more exhaustive learning and "recall after the event" computations for making decisions. Thus, the new model-free RL may account for some behaviors or functions that have been ascribed to the original model-based RL. More importantly, our reward structural learning indicates a potentially more ideal mechanism for value-based decision making, balancing among economy, optimality and flexibility.

Please cite this article in press as: Nakahara, H., Hikosaka, O., Learning to represent reward structure: A key to adapting to complex environments. Neurosci. Res. (2012), http://dx.doi.org/10.1016/j.neures.2012.09.007

6. Future directions

The dopamine reward structural learning hypothesis raises a number of questions that need to be addressed. For example, what are the computational processes that underlie the learning of reward structures in internal state representations, or representation learning? As noted above, several experimental studies indicate that different forms of reward structures may be learned in internal representation during different tasks. A pressing computational question seeks to find the relationship between unified representation learning and DA activity, or the form or aspect of representation learning to which DA activity contributes. Studies of reward-driven perceptual learning address interactions between representation learning and prediction learning, and their progress will provide insights (Nomoto et al., 2010; Reed et al., 2011; Seitz and Dinse, 2007). Progress related to learning the reward structure in representation has been ongoing in other fields apart from neuroscience, such as machine learning, by using predictive states, extracting or approximating features that represent states, or using other types of time traces (Daw et al., 2006; Gershman et al., 2012; Ludvig et al., 2008; Nakahara and Kaveri, 2010; Parr et al., 2007; Sutton et al., 2009, 2011). Interestingly, they suggest different ways to improve state representation, and future research can benefit from their use (Wan et al., 2011).

Which neurophysiological and behavioral experiments can allow us to further examine representation learning of reward structure? A useful experiment is to systematically probe the specific information that is useful for value-based decisions, hidden within a moment or over moments, that can be reflected in DA responses, and whether such DA responses change through the experience of trials, concordantly with behavioral choices. For example, few studies have systematically addressed the use of extracting and learning temporal structure information for value-based decision making. To dissect the roles of DA activity or activity in other related areas in learning, it is desirable to be able to inactivate DA neurons or the activity of other neurons in a reversible manner.

Which neural circuits underlie the concurrent processes of representation and prediction learning? Insights may be gained by considering their relationships for computations and circuits together. First, the areas that generate internal representation should be located upstream from those that generate reward prediction and action selection (Fig. 1B). A clear possibility is a combination of cortical and basal ganglia areas that receive heavy DA innervation; for example, the prefrontal cortical areas may act primarily for learning the reward structure in internal states (McDannald et al., 2011, 2012; Rushworth et al., 2012), whereas the striatum may act primarily for learning the reward prediction (and action selection). Second, representation learning would require more detailed learning signals than prediction learning, so that areas receiving heterogeneous DA signals, such as salient signals, are more likely to be involved in representation learning. Areas that receive projections from DA neurons in the dorsolateral SNc, in which DA neurons that encode salient signals tend to be located, include the dorsolateral prefrontal cortex, dorsal striatum, and nucleus accumbens (core) (Bromberg-Martin et al., 2010b; Lammel et al., 2008; Matsumoto and Hikosaka, 2009). Areas that have neural activity that is akin to salient signals may also be a part of the circuit for representation learning, such as the basolateral amygdala and anterior cingulate cortex (Hayden et al., 2010; Roesch et al., 2010). In summary, synthesizing the original success of the reward prediction error hypothesis and the discrepancies found in recent experimental evidence, the reward structural learning hypothesis can help to guide future research for understanding neural value-based decision making.

Acknowledgement

This work is partly supported by KAKENHI grants 21300129 and 24120522 (H.N.).

References

Acuna, D.E., Schrater, P., 2010. Structure learning in human sequential decision-making. PLoS Comput. Biol. 6 (12), e1001003.
Balleine, B.W., Daw, N., O'Doherty, J.P., 2008. Multiple Forms of Value Learning and the Function of Dopamine, Neuroeconomics Decision Making and the Brain. Elsevier, Amsterdam, pp. 367–387.
Bao, S., Chan, V.T., Merzenich, M.M., 2001. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. Nature 412 (6842), 79–83.
Bapi, R.S., Miyapuram, K.P., Graydon, F.X., Doya, K., 2006. fMRI investigation of cortical and subcortical networks in the learning of abstract and effector-specific representations of motor sequences. NeuroImage 32 (2), 714–727.
Barto, A., 1994. Adaptive critics and the basal ganglia. In: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in the Basal Ganglia. , pp. 12–31. Q2
Bayer, H., Glimcher, P., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47 (1), 129–141.
Brischoux, F., Chakraborty, S., Brierley, D.I., Ungless, M.A., 2009. Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proc. Natl. Acad. Sci. U.S.A. 106 (12), 4894–4899.
Bromberg-Martin, E.S., Hikosaka, O., 2009. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron 63 (1), 119–126.
Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O., 2010a. Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. Neuron 67 (1), 144–155.
Bromberg-Martin, E.S., Matsumoto, M., Nakahara, H., Hikosaka, O., 2010b. Multiple timescales of memory in lateral habenula and dopamine neurons. Neuron 67 (3), 499–510.
Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., Uchida, N., 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482 (7383), 85–88.
Costa, R.M., 2011. A selectionist account of de novo action learning. Curr. Opin. Neurobiol. Q3
Courville, A., Daw, N., Touretzky, D., 2006. Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. 10 (7), 294–300.
Daw, N.D., Courville, A.C., Touretzky, D.S., 2006. Representation and timing in theories of the dopamine system. Neural Comput. 18 (7), 1637–1677.
Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69 (6), 1204–1215.
Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8 (12), 1704–1711.
Dayan, P., 1993. Improving generalization for temporal difference learning: the successor representation. Neural Comput. 5, 613–624.
Dayan, P.G., Niv, Y., 2008. Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. 18 (2), 185–196. Q4
Doya, K., 1999. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. 12 (7–8), 961–974.
Doya, K., 2007. Reinforcement learning: computational theory and biological mechanisms. HFSP J. 1 (1), 30–40.
Enomoto, K., Matsumoto, N., Nakai, S., Satoh, T., Sato, T.K., Ueda, Y., Inokawa, H., Haruno, M., Kimura, M., 2011. Dopamine neurons learn to encode the long-term value of multiple future rewards. Proc. Natl. Acad. Sci. U.S.A.
Fiorillo, C.D., Tobler, P.N., Schultz, J., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299, 1898–1902.
Frank, M.J., Surmeier, D.J., 2009. Do substantia nigra dopaminergic neurons differentiate between reward and punishment? J. Mol. Cell Biol. 1 (1), 15–16.
Funamizu, A., Ito, M., Doya, K., Kanzaki, R., Takahashi, H., 2012. Uncertainty in action-value estimation affects both action choice and learning rate of the choice behaviors of rats. Eur. J. Neurosci. 35 (7), 1180–1189.
Gershman, S.J., Moore, C.D., Todd, M.T., Norman, K.N., Sederberg, P.B., 2012. The successor representation and temporal context. Neural Comput. 24, 1–16.
Gershman, S.J., Niv, Y., 2010. Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. 20 (2), 251–256.
Gershman, S.J., Pesaran, B., Daw, N.D., 2009. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J. Neurosci. 29 (43), 13524–13531.
Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66 (4), 585–595.
Glimcher, P.W., 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. U.S.A. 108 (Suppl. 3), 15647–15654.
Green, C.S., Benson, C., Kersten, D., Schrater, P., 2010. Alterations in choice behavior by manipulations of world model. Proc. Natl. Acad. Sci. U.S.A. 107 (37), 16401–16406.

Please cite this article in press as: Nakahara, H., Hikosaka, O., Learning to represent reward structure: A key to adapting to complex environments. Neurosci. Res. (2012), http://dx.doi.org/10.1016/j.neures.2012.09.007

Gruber, A.J., Dayan, P., Gutkin, B.S., Solla, S.A., 2006. Dopamine modulation in the basal ganglia locks the gate to working memory. J. Comput. Neurosci. 20 (2), 153–166.
Hayden, B.Y., Heilbronner, S., Pearson, J., Platt, M.L., 2010. Neurons in anterior cingulate cortex multiplex information about reward and action. J. Neurosci. 30 (9), 3339–3346.
Hayden, B.Y., Pearson, J.M., Platt, M.L., 2011. Neuronal basis of sequential foraging decisions in a patchy environment. Nat. Neurosci.
Hikosaka, O., Nakahara, H., Rand, M.K., Sakai, K., Lu, X., Nakamura, K., Miyachi, S., Doya, K., 1999. Parallel neural networks for learning sequential procedures. Trends Neurosci. 22 (10), 464–471.
Hikosaka, O., Nakamura, K., Nakahara, H., 2006. Basal ganglia orient eyes to reward. J. Neurophysiol. 95 (2), 567–584.
Houk, J.C., Adams, J.L., Barto, A., 1994. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in the Basal Ganglia. , pp. 249–252.
Joshua, M., Adler, A., Rosin, B., Vaadia, E., Bergman, H., 2009. Encoding of probabilistic rewarding and aversive events by pallidal and nigral neurons. J. Neurophysiol. 101 (2), 758–772.
Kakade, S., Dayan, P., 2002. Dopamine: generalization and bonuses. Neural Netw. 15 (4–6), 549–559.
Kolling, N., Behrens, T.E., Mars, R.B., Rushworth, M.F., 2012. Neural mechanisms of foraging. Science 336 (6077), 95–98.
Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B., Roeper, J., 2008. Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. Neuron 57 (5), 760–773.
Ludvig, E.A., Sutton, R.S., Kehoe, E.J., 2008. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Comput. 20 (12), 3034–3054.
Matsumoto, M., Hikosaka, O., 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459 (7248), 837–841.
McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y., Schoenbaum, G., 2011. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J. Neurosci. 31 (7), 2700–2705.
McDannald, M.A., Takahashi, Y.K., Lopatina, N., Pietras, B.W., Jones, J.L., Schoenbaum, G., 2012. Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. Eur. J. Neurosci. 35 (7), 991–996.
Montague, P., Dayan, P., Sejnowski, T., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16 (5), 1936–1947.
Montague, P.R., Hyman, S.E., Cohen, J.D., 2004. Computational roles for dopamine in behavioural control. Nature 431 (7010), 760–767.
Montague, P.R., King-Casas, B., Cohen, J.D., 2006. Imaging valuation models in human choice. Annu. Rev. Neurosci. 29, 417–448.
Morris, G., Nevet, A., Arkadir, D., Vaadia, E., Bergman, H., 2006. Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. 9 (8), 1057–1063.
Nakahara, H., Doya, K., Hikosaka, O., 2001. Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences – a computational approach. J. Cogn. Neurosci. 13 (5), 626–647.
Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., Hikosaka, O., 2004. Dopamine neurons can represent context-dependent prediction error. Neuron 41, 269–280.
Nakahara, H., Kaveri, S., 2010. Internal-time temporal difference model for neural value-based decision making. Neural Comput. 22 (12), 3062–3106.
Nomoto, K., Schultz, W., Watanabe, T., Sakagami, M., 2010. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J. Neurosci. 30 (32), 10692–10702.
Palminteri, S., Boraud, T., Lafargue, G., Dubois, B., Pessiglione, M., 2009. Brain hemispheres selectively track the expected value of contralateral options. J. Neurosci. 29 (43), 13465–13472.
Parr, R., Painter-Wakefield, C., Li, L., Littman, M., 2007. Analyzing feature generation for value-function approximation. New York, NY, USA. p. 737–744.
Rangel, A., Camerer, C., Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. 9 (7), 545–556.
Rao, R.P., 2010. Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front. Comput. Neurosci. 4, 146.
Redgrave, P., Gurney, K., 2006. The short-latency dopamine signal: a role in discovering novel actions? Nat. Rev. Neurosci. 7 (12), 967–975.
Redish, A.D., Jensen, S., Johnson, A., Kurth-Nelson, Z., 2007. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol. Rev. 114 (3), 784–805.
Reed, A., Riley, J., Carraway, R., Carrasco, A., Perez, C., Jakkamsetti, V., Kilgard, M.P., 2011. Cortical map plasticity improves learning but is not necessary for improved performance. Neuron 70 (1), 121–131.
Reynolds, J.N., Wickens, J.R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15 (4–6), 507–521.
Ribas-Fernandes, J.J.F., Solway, A., Diuk, C., McGuire, J.T., Barto Andrew, G., Niv, Y., Botvinick, M.M., 2011. A neural signature of hierarchical reinforcement learning. Neuron 71 (2), 370–379.
Roesch, M.R., Calu, D.J., Esber, G.R., Schoenbaum, G., 2010. Neural correlates of variations in event processing during learning in basolateral amygdala. J. Neurosci. 30 (7), 2464–2471.
Roesch, M.R., Calu, D.J., Schoenbaum, G., 2007. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. Neurosci. 10 (12), 1615–1624.
Rushworth, M.F., Kolling, N., Sallet, J., Mars, R.B., 2012. Valuation and decision-making in frontal cortex: one or many serial or parallel systems? Curr. Opin. Neurobiol.
Satoh, T., Nakai, S., Sato, T., Kimura, M., 2003. Correlated coding of motivation and outcome of decision by dopamine neurons. J. Neurosci. 23 (30), 9913–9923.
Schultz, W., 1998. Predictive reward signal of dopamine neurons. J. Neurophysiol. 80, 1–27.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275 (5306), 1593–1599.
Seitz, A.R., Dinse, H.R., 2007. A common framework for perceptual learning. Curr. Opin. Neurobiol. 17 (2), 148–153.
Singh, S., Barto, A.G., Chentanez, N., 2005. Intrinsically Motivated Reinforcement Learning. Vancouver, B.C., Canada.
Sutton, R., Barto, A.G., 1998. Reinforcement Learning: An Introduction.
Sutton, R.S., Barto, A.G., 1990. Time-derivative models of pavlovian reinforcement. In: Gabriel, M., Moore, J. (Eds.), Learning and Computational Neuroscience: Foundations of Adaptive Networks. The MIT Press, pp. 497–537.
Sutton, R.S., Maei, H.R., Precup, D., Bhatnagar, S., Silver, D., Szepesvari, C., Wiewiora, E., 2009. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In: ICML-09, Montreal, Canada, pp. 993–1000.
Sutton, R.S., Modayil, J., Delp, M., Degris, T., Pilarski, P.M., White, A., 2011. Horde: A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction.
Suzuki, S., Harasawa, N., Ueno, K., Gardner, J.L., Ichinohe, N., Haruno, M., Cheng, K., Nakahara, H., 2012. Learning to simulate others' decisions. Neuron 74, 1125–1137.
Todd, M., Niv, Y., Cohen, J.D., 2009. Learning to use working memory in partially observable environments through dopaminergic reinforcement. NIPS, 1–8.
Wan, X., Nakatani, H., Ueno, K., Asamizuya, T., Cheng, K., Tanaka, K., 2011. The neural basis of intuitive best next-move generation in board game experts. Science 331 (6015), 341–346.
Wunderlich, K., Dayan, P., Dolan, R.J., 2012. Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci., 1–19.
Zacks, J.M., Kurby, C.A., Eisenberg, M.L., Haroutunian, N., 2011. Prediction error associated with the perceptual segmentation of naturalistic events. J. Cogn. Neurosci. 23 (12), 4057–4066.

Please cite this article in press as: Nakahara, H., Hikosaka, O., Learning to represent reward structure: A key to adapting to complex environments. Neurosci. Res. (2012), http://dx.doi.org/10.1016/j.neures.2012.09.007