Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Model-based predictions for dopamine

Angela J Langdon[1], Melissa J Sharpe[1,2,3], Geoffrey Schoenbaum[2] and Yael Niv[1]

Phasic dopamine responses are thought to encode a prediction-error signal consistent with model-free reinforcement learning theories. However, a number of recent findings highlight the influence of model-based computations on dopamine responses, and suggest that dopamine prediction errors reflect more dimensions of an expected outcome than scalar reward value. Here, we review a selection of these recent results and discuss the implications and complications of model-based predictions for computational theories of dopamine and learning.

Addresses
[1] Princeton Neuroscience Institute & Department of Psychology, Princeton University, Princeton, NJ 08540, United States
[2] National Institute on Drug Abuse, Baltimore, MD 21224, United States
[3] School of Psychology, University of New South Wales, Australia

Corresponding author: Langdon, Angela J (alangdon@princeton.edu)

Current Opinion in Neurobiology 2017, **49**:1–7

This review comes from a themed issue on **Neurobiology of behavior**

Edited by **Kay Tye** and **Naoshige Uchida**

http://dx.doi.org/10.1016/j.conb.2017.10.006

Introduction

The striking correspondence between the phasic responses of midbrain dopamine neurons and the temporal-difference reward prediction error posited by reinforcement-learning theory is by now well established [1–5]. According to this theory, dopamine neurons broadcast a prediction error — the difference between the learned predictive value of the current state, signaled by cues or features of the environment, and the sum of the current reward and the value of the next state. Central to the normative grounding of temporal-difference reinforcement learning (TDRL) is the definition of 'value' as the expected sum of future (possibly discounted) rewards [6], from whence the learning rule can be derived directly. The algorithm also provides a simple way to learn such values using prediction errors, which is thought to be implemented in the brain through dopamine-modulated plasticity in corticostriatal synapses [7,8] (Figure 1, left). This theory provides a parsimonious account of a number of features of dopamine responses in a range of learning tasks [9–12].

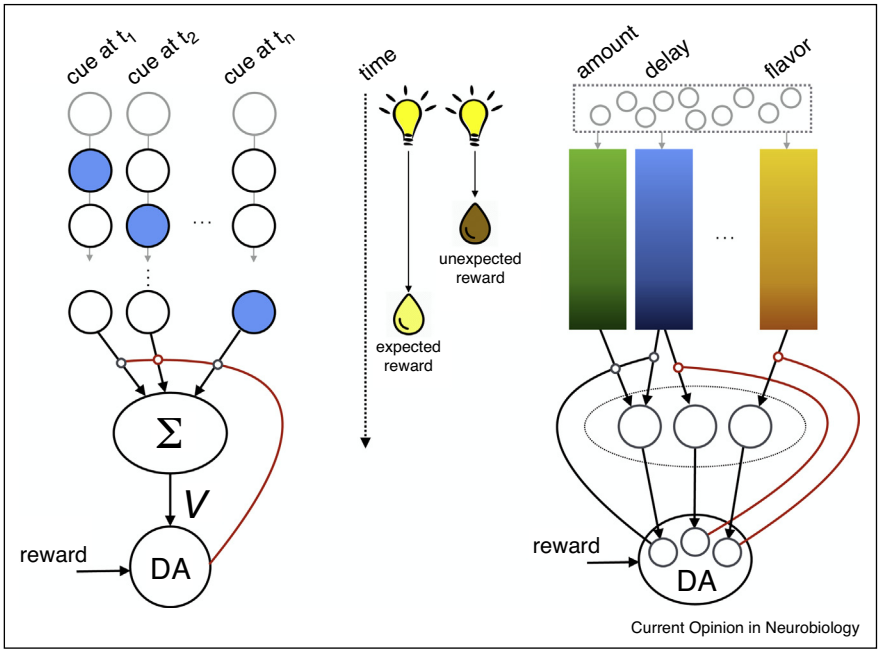
Are model-free dopamine prediction errors a red herring?

A core tenet of TDRL is that it is 'model-free': learned state values are aggregate, scalar representations of total future expected reward, in some common currency [1,13]. That is, the value of a state is a quantitative summary of future reward amount, irrespective of either the specific form of the expected reward (e.g., water, food, a combination of the two), or the sequence of future states through which it will be obtained (e.g., will water be presented before or after food). Critically, model-free TDRL assigns these summed values to temporally-defined states; accordingly, the algorithm binds together predictions about the amount of reward and the expected time of delivery (Figure 1). In many studies, dopamine signals appear to reflect such temporally-precise, unitary value expectations, which also correlate with conditioned responding and choice preferences [14,15]. However, little work has tested this strong hypothesis directly, by, for instance, having a single cue predict several rewards of different types within a single trial, or by testing the effects of changes in type of reward on dopamine signaling, while keeping the reward value constant.

Another important feature of model-free learning (including TDRL) is that it posits that scalar state values are accrued solely through experiencing the relationship between the current state and the (possibly rewarded) state that follows [6,16]. That is, state values are learned through experience and 'cached' for future use. This is in contrast to model-based decision making [17], where values are computed anew each time a state is encountered by mentally simulating possibly distant futures using a learned internal 'world model', which captures the sequences of transitions between non-adjacent states and their associated rewards (but see below for some more nuanced distinctions).

Although phasic dopamine signals have predominantly been interpreted as model-free temporal difference prediction errors, a growing number of studies leveraging complex behavioral tasks, alongside novel optogenetic and imaging techniques, are revealing an increasingly detailed picture of dopamine reward prediction errors during learning, and the multiple dimensions of reward prediction on which they are based. Intriguingly, several of these studies have demonstrated a significant degree of

Figure 1



Multiple dimensions of prediction in dopamine prediction errors. Consider a simple task in which a brief presentation of a light cue is repeatedly followed by a drop of vanilla milk after some fixed delay (middle). What would happen on a trial in which the light is followed by a drop of equally-preferred chocolate milk after a shorter delay? Model-free TDRL with a complete serial compound stimulus representation proposes that the cue triggers a discrete sequence of activity that represents sequential time points after the presentation of the cue (left; a number of neurons are depicted horizontally; their activity at different timepoints is portrayed vertically). At each timepoint, summation of this weighted representation produces a scalar estimate of future value (V), which dopamine neurons (DA) compare to obtained reward to compute a prediction error signal. The prediction error is then broadcast widely (red) and used to modify the weights for neurons that were recently active (circles on arrows). When an unexpectedly early, chocolate-flavored reward is delivered, the prediction error signals the difference in time-discounted value, and modifies the weights for the part of the representation that is active when the prediction error is signaled. By contrast, we propose that dopamine neurons have access to (and maybe aid in learning) dimensions of prediction other than scalar value, and these are used for computation and signaling of prediction errors (right). For example, after the presentation of the cue, multiple features of the predicted next event (in this case, a liquid reward) may be represented by (perhaps overlapping) populations of neurons through time (color gradient), including the predicted amount (e.g., one drop), the delay to reward delivery (it will arrive after several seconds) and the flavor of the reward (vanilla milk). At the time of reward delivery, violations of the prediction along any of these dimensions may elicit a phasic response from dopamine neurons, though different neurons may be specialized for prediction errors corresponding to different dimensions. In this case, at the early presentation of a drop of chocolate milk, prediction errors are elicited for the timing of reward delivery as well as for flavor (red) but no prediction error arises for amount (black).

heterogeneity in dopaminergic responses during learning, suggesting greater complexity in these signals than previously appreciated. Below we review evidence from these recent studies, asking what is the nature of dopamine signals? Do they reflect an aggregate (scalar) error, or a vector-based signal that includes not only the magnitude of deviation from predictions, but also the identity of the deviation (did I get more food than expected, or water instead of food)? And how might these signals be incorporated into learning algorithms implemented throughout the brain?

Temporal representation and dopamine

One notable property of dopamine prediction errors is that they are temporally precise: if an expected reward is omitted, the phasic decrease in dopamine neuron activity appears just after the time the reward would have occurred [2]. It is this phenomenon that inspired the TDRL algorithm, which models such temporally precise predictions by postulating sequences of time-point states that are triggered by a stimulus (known as the 'complete serial compound,' CSC stimulus representation, or 'tapped delay line'; Figure 1), each of which separately accrues value through experience [6]. However, when a reward is delivered unexpectedly early, dopamine neurons do not display a phasic decrease in activity at the original expected time of reward, as would be implied by the CSC, in which a prediction error updates the value of the current, and not subsequent, timepoint states [18,19]. Reset mechanisms, in which reward delivery terminates the CSC representation, have been proposed to address this [19], but other challenges suggest that the CSC is perhaps not as viable an explanation for learned timing. Specifically, prediction errors are only slightly enhanced to temporally variable rewards, suggesting that under some conditions reward predictions may have low