Dialogues on prediction errors

Yael Niv[1] and Geoffrey Schoenbaum[2]

[1] Center for the Study of Brain, Mind and Behavior and Department of Psychology, Green Hall, Princeton University, Princeton, NJ 08544, USA
[2] Departments of Anatomy and Neurobiology, and Psychiatry, University of Maryland School of Medicine, 20 Penn Street, Baltimore, MD 21201, USA

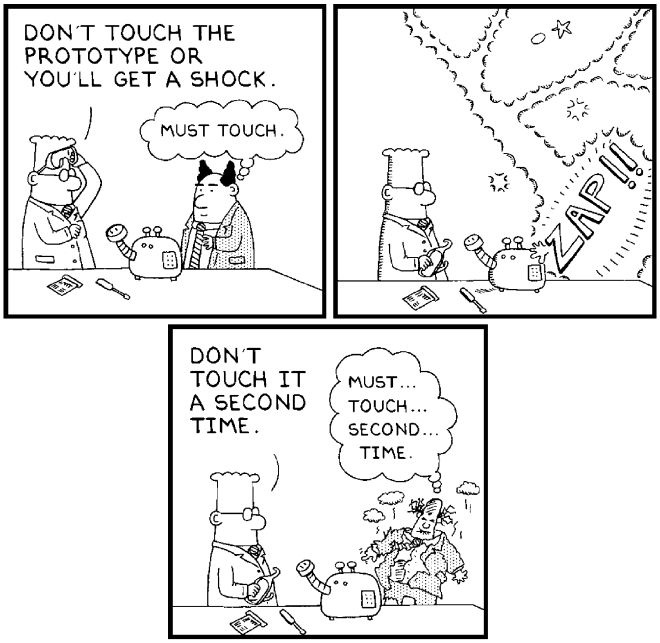
The recognition that computational ideas from reinforcement learning are relevant to the study of neural circuits has taken the cognitive neuroscience community by storm. A central tenet of these models is that discrepancies between actual and expected outcomes can be used for learning. Neural correlates of such prediction-error signals have been observed now in midbrain dopaminergic neurons, striatum, amygdala and even prefrontal cortex, and models incorporating prediction errors have been invoked to explain complex phenomena such as the transition from goal-directed to habitual behavior. Yet, like any revolution, the fast-paced progress has left an uneven understanding in its wake. Here, we provide answers to ten simple questions about prediction errors, with the aim of exposing both the strengths and the limitations of this active area of neuroscience research.

I'M ALWAYS WRONG ABOUT EVERYTHING. WHAT CAN I DO TO FIX THAT?

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Introduction

Arguably, some of the most profound developments in psychology and neuroscience in the last two decades have stemmed from the use of normative ideas from reinforcement learning in thinking about and studying behavior and the brain. Building on a foundation laid by learning theorists, neuroscientists have identified neural substrates that conform to predictions of precise mathematical models of decision making. In this review we focus on a central tenet of reinforcement learning models: temporal difference prediction errors that quantify the discrepancy between what was expected and what is actually observed and that serve to stamp in associations with antecedent events. This theoretical concept was first given neural form in reports that activity in midbrain dopamine neurons met criteria for prediction errors [1,2]. A decade later, correlates of prediction errors are coming up all over the brain [3–9]. However, the exact scope of the temporal difference model and the prediction error hypothesis of dopamine, as well as their implications for neuroscience and psychology, are not always clear. The goal of this 'dialogue style' review (which loosely follows a series of question and answer e-mails between the authors) is to make clear to those not versed in reinforcement learning theory what temporal difference prediction errors are and how this theory interacts with neuroscientific research.


DON'T TOUCH THE PROTOTYPE OR YOU'LL GET A SHOCK.
MUST TOUCH.
ZAP!!
DON'T TOUCH IT A SECOND TIME.
MUST... TOUCH... SECOND... TIME.

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Q1: What is a prediction error and what is it good for?

Learning, in its most basic form, can be seen as the process by which we become able to use past and current events to predict what the future holds. In classical conditioning animals learn to predict what outcomes (e.g. food, water, or foot shock) are contingent on which events (a bell ringing, clouds welling up in the sky, etc.). In instrumental conditioning animals learn to predict the consequences of their actions and can

Corresponding authors: Niv, Y. (yael@princeton.edu); Schoenbaum, G. (schoenbg@schoenbaumlab.org).

1364-6613/$ – see front matter © 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2008.03.006

potentially use this knowledge to maximize the likelihood of rewards and to minimize the occurrence of punishments.

The most intuitive way to learn to predict future reward and punishments is via error correction. The principle here is simple: make the best prediction you can, observe actual events and if your prediction was wrong, update your knowledge-base so that future predictions are more accurate. This is the basis of the extremely influential Rescorla-Wagner [10] model of classical conditioning. For example, imagine trying to predict how good a bottle of wine will be. Opening a bottle of Bordeaux that has been aging in your cellar, you are delighted with its sophisticated flavor. Though you might have hoped for this, you were, presumably, less than 100% certain that the wine had not passed its prime. As a result, there is a difference between your prediction – say, 70% chance of a good bottle of wine – and reality. This error can be used to make your prediction more accurate in the future. Of course not all Bordeaux are alike, so rather than update your prediction to match exactly the current situation – 100% – you might update your prediction to some other probability, say 85%, reflective of the higher likelihood of a good 10-year-old Bordeaux. Through this trial-and-error process of adjustment, over many bottles of wine, you will eventually learn the correct expected reward derived from different types and ages of wine.

The key computational quantity that drives learning in this example is the discrepancy between predictions and outcomes, that is, the prediction error. Rescorla and Wagner used this quantity in formulating their learning rule [Equation 1]:

$$\begin{aligned} V_{new} &= V_{old} + \eta(outcome - prediction) \\ &= V_{old} + \eta(R - V_{old}) \end{aligned}$$

Here, R is a scalar quantity denoting the goodness of the outcome (a pellet of food or a bottle of wine) and V is the prediction associated with the observed stimulus (a tone that precedes food or the label on the wine bottle), again in units of predicted goodness, derived from past experience with that stimulus. Rescorla and Wagner stipulated that the overall prediction in a certain situation is the sum of all the predictions from all available stimuli $\sum_{stimuli} V_{old}^{stim}$.

In the idealized world of Rescorla and Wagner, the update rule above is applied at the end of each conditioning trial, to all stimuli present in that trial. The learning rate parameter $0 < \eta \leq 1$ determines just how much each specific experience affects the prediction for the future. High learning rates mean that new experience is weighed heavily in the future prediction (thus, learning from new experience is faster, but forgetting the more distant past also is faster), and low learning rates mean that much experience needs to accumulate to profoundly affect predictions.

This simple but powerful learning rule is, perhaps, the most influential model of conditioning to date, successfully explaining phenomena such as blocking [11], overshadowing [12] and conditioned inhibition [13,14] and predicting others not known at the time, such as overexpectation [15,16].

MANAGEMENT TRAINING
WHAT WOULD YOU DO IF YOU MADE A HUGE, INCREDIBLY STUPID MISTAKE?
I WOULD TRY TO LEARN FROM IT.
DID YOU LEARN ANYTHING FROM YOUR ANSWER?

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Q2: Reinforcement learning models of the dopamine system have been associated with a slightly different concept – a temporal difference prediction error. How is this different from the Rescorla-Wagner prediction error?

Ideas about temporal difference (TD) learning and TD prediction errors stem from a line of research on reinforcement learning within the fields of control theory and computer science that was largely motivated by data from classical conditioning (e.g. [17]; see [18] for a comprehensive treatment and [19] for a detailed review). TD learning takes into account that life is not naturally divisible into discrete trials but, rather, consists of a continuous flow of experience. Within this flow, predictive stimuli and rewarding outcomes occur at different points in time, and the goal, at each point in time, is to predict all future outcomes given current and previous stimuli.

To see how this ambitious goal can be achieved, let's start by defining the prediction based on the stimulus at time t (also called the 'value' of this stimulus) as the expected sum of future outcomes:

$$\begin{aligned} & prediction(t) = \\ & = E[outcome(t+1) + outcome(t+2) + outcome(t+3) + \ldots] \\ & = E[outcome(t+1)] + E[outcome(t+2)] + E[outcome(t+3)] \ldots \end{aligned}$$

Of course we can say the same for the prediction based on the stimulus at time $t + 1$:

$$prediction(t+1) = E[outcome(t+2)] + E[outcome(t+3)] + \ldots$$

It then follows directly that:

$$prediction(t) = E[outcome(t+1)] + prediction(t+1),$$

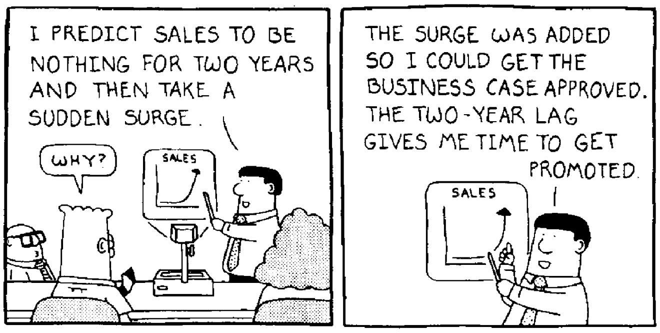
meaning that if our predictions are correct, the prediction based on the stimulus at time t should equal to the sum of two quantities: (i) the expected immediate reward one timestep later (which might be zero) and (ii) the predicted rewards from that time and onward. This also means that if we predict a wrong amount at time t, then at time $t + 1$ we will realize we have a temporal difference prediction error because the immediate reward plus future expectations will be higher or lower than our original prediction. The prediction error δ at time $t + 1$ is, thus [Equation 2],

$$\delta(t+1) = outcome(t+1) + prediction(t+1) - prediction(t).$$

This error can be used to improve the prediction we made based on the stimulus at time t, that is, to learn the value of that stimulus (shorthanded $V(t)$), as in the Rescorla-Wagner model [Equation 3]:

$$\begin{aligned} V(t)_{new} &= \\ &= V(t)_{old} + \eta \cdot \delta(t+1) \\ &= V(t)_{old} + \eta[outcome(t+1) + prediction(t+1) \\ &\quad - prediction(t)]. \end{aligned}$$

To understand this intuitively, imagine a situation in which several predictive events might follow one another. For instance, going back to our bottle of wine, you might have predicted a superb wine based on the age and label. However, suppose you then notice that the cork is dry and crumbling. Even though you have not even opened the bottle, you can use this new state of affairs to change your prediction. This is not possible using the Rescorla-Wagner learning rule because it only looks at the difference between predictions and outcomes at the end of the trial. By contrast TD learning is based on the difference between consecutive predictions (99% superb wine upon seeing the label, and merely a hopeful 30% upon realizing that it might have turned to vinegar due to the bad cork). In Equation (3), *V(t)* is the prediction at time point t (say, when seeing the label), *outcome* ($t + 1$) is the outcome (still no wine) at the next time point (trying to open the cork) and *prediction* ($t + 1$) is the prediction at that time point (probably bad wine).


I PREDICT SALES TO BE NOTHING FOR TWO YEARS AND THEN TAKE A SUDDEN SURGE.
WHY?
SALES
THE SURGE WAS ADDED SO I COULD GET THE BUSINESS CASE APPROVED. THE TWO-YEAR LAG GIVES ME TIME TO GET PROMOTED.
SALES

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Q3: What defines how far into the future our predictions should go?

Of course, there must be some way to carve up time – it makes no sense to try to predict the sum of all future rewards from now until the day we die. There are two ways to deal with this:

(i) Sometimes, as in the case of trying different bottles of wine, it does make sense to break up the sequence of events into discrete trials. For instance, in a laboratory experiment in which the subject is instructed to treat each trial independently (or leads us to believe that this is what she is doing), TD learning can be applied with the goal of predicting the sum of all rewards in a trial. Still, TD learning is more general than the Rescorla-Wagner model because it accounts properly for the timing of multiple stimuli and rewards within the trial.

(ii) More generally, research shows that animals and humans discount, or view as less valuable, rewards that are far away in time. TD learning can account for such temporal discounting by changing Equation (2) to:

$$\delta(t+1) = \gamma[outcome(t+1) + prediction(t+1)] - prediction(t),$$

with $0 < \gamma \leq 1$ as an (exponential) discounting factor. Discounting the value of future rewards limits the horizon of prediction: rewards two years down the line are worth nil and, so, contribute nothing to the sum of future rewards that we want to predict, and the steeper the discounting, the shorter the prediction horizon. Similarly, the past predictive information is limited by the steepness of the discounting function: cues too far back cannot predict rewards obtained at present.

Q4: What are the crucial differences between TD learning and Rescorla-Wagner learning? Are there situations when one reduces to the other?

The crucial difference lies in the distinction between a discrete trial-level model and a continuous-time model of learning. In Rescorla and Wagner's model, the goal is to predict the outcome of a trial and prediction errors only arise at a trial's end. In TD models the goal is to predict the future and any new information can lead to prediction errors.

Indeed, when only one or several simultaneous cues predict reward, the prediction error at the time of the reward will be similar in both models. However, even in this case TD does not entirely reduce to the Rescorla-Wagner model. This is because in Pavlov's famous [bell → food] experiment, Rescorla and Wagner would have it that there are prediction errors only at the time of food arrival, whereas TD learning posits that there also is a prediction error when the bell is sounded, that is, at the time of the predictive cue. This is because before hearing the bell there was no expectation of food [$V(t) = 0$], whereas after hearing it, food is expected [$V(t + 1) = 1$]. This difference between successive values is sufficient to generate a prediction error. Moreover, the presence of this signal means that if another stimulus precedes the bell (say, Pavlov entering the room or a light preceding the bell as in second-order conditioning), the animal can learn to use that stimulus to predict the occurrence of the bell (and thereby the food reward). Second- (or higher) order conditioning is important because it is the basis for the motivating effects of money and other proxies for primary reward. Furthermore, this characteristic of the TD prediction error as compared with a Rescorla-Wagner prediction error will become crucial when we compare the predictions of these models to neural activity.

Thus, the TD model of learning (i.e. learning according to Equation 3) can explain not only all those phenomena that were explained by the Rescorla-Wagner rule but also phenomena such as second-order conditioning and conditioned reinforcement, as well as within-trial effects of the temporal relationship between stimuli and outcomes [17].

Q5: The influential reward prediction error hypothesis of dopamine arose from comparing monkey electrophysiological data to the characteristics of a TD prediction error [1]. Today, other forms of neural recordings have targeted this same signal. What are the basic criteria for establishing that a recorded signal is, indeed, a TD prediction error?

Three criteria can be considered the ‘fingerprint’ of a reward prediction error signal: a phasic increase to unexpected rewards (a positive prediction error), no change to predicted rewards and a phasic decrease (a negative prediction error) when an expected reward is omitted (or vice-versa – decreases to positive errors and increases to negative errors, e.g. [20]). Of course, to establish that a signal is a TD prediction error rather than a Rescorla-Wagner prediction error, one should also show that after learning, the unexpected presentation of a stimulus that predicts future reward elicits a prediction error signal. These characteristics have been demonstrated repeatedly for midbrain dopamine neurons. Indeed, the presentation of a predictive cue elicits a burst of activity from these neurons, the size of which is proportional to the magnitude, probability and even delay of the predicted reward [21–23].

Note that prediction errors should arise only at the boundaries between situations that predict different amounts of reward. For instance, you might experience a prediction error when you see the label on a wine bottle, but there are no further errors while you are looking for the wine opener and getting ready to open the bottle. This is because in the absence of new information, the predictions at times t and $t + 1$ would be identical. Other quantities of interest might not be phasic: the expected value $V(t)$ (that quantifies the amount of reward predicted) should change from 0 to the cue-related value once the label is seen and stay at that level until the final consumption of the wine (Figure 1).

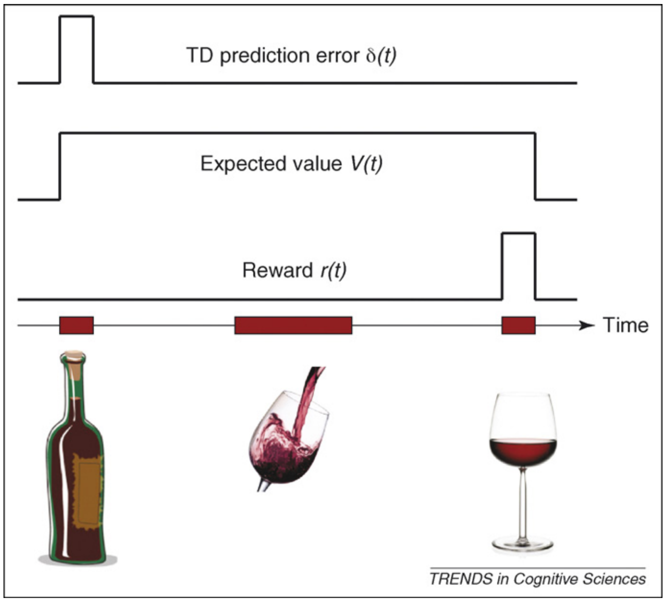


Figure 1. The time course of the reward, value and prediction error signals in the TD model. The first predictive stimulus is the label on the wine bottle, after which wine is poured into the glass and finally consumed. Cue-related phasic neural signals whose magnitude reflects the future predicted reward can be called prediction error signals, but sustained neural signals corresponding to the value of the predicted reward throughout the trial are designated value signals.

Q6: The TD theory can account for both appetitive and aversive outcomes by simply assuming that an aversive event has a negative utility whereas a positive event has a positive utility. Is this what is seen in dopaminergic firing patterns?

That’s a good question that has been at the heart of an interesting debate regarding the dopamine signal and has yet to be resolved convincingly. According to one popular view, dopamine serves to direct attention to salient events, and it only seems like a prediction error because rewards and reward-predicting stimuli are salient events (see, e.g., [24,25]). To test this hypothesis all eyes (or, should we say, electrodes) have turned to aversive events such as shocks. These are clearly salient, but carry a negative prediction, so the salience hypothesis predicts a positive phasic response to such stimuli, whereas the prediction error hypothesis suggests the opposite.

Initial evidence showed that some dopaminergic neurons responded to aversive events with a burst of firing [26]. However, recent work in anesthetized animals suggests that midbrain neurons that fire in response to aversive events are not dopaminergic (although their location and the shape of their action potentials are deceptively similar to that of dopaminergic neurons), whereas dopaminergic neurons respond to aversive events with a dip in firing [27,28]. As for awake animals, recordings in monkeys have been inconclusive because only weakly aversive stimuli such as air-puffs have been used [29]. Recent advances in dopamine recordings in awake rats, when more strongly aversive stimuli can be applied, might help resolve this controversy.

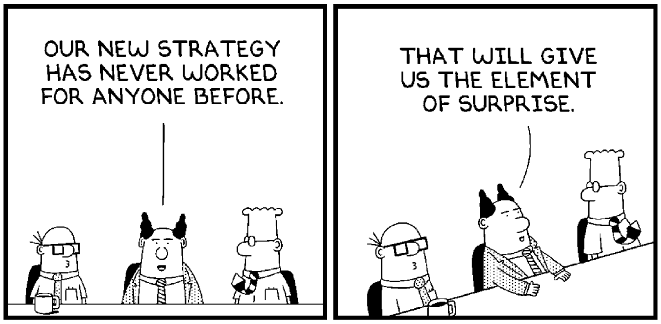
From a practical point of view, it is unclear whether dips below the baseline firing rate of dopamine neurons – either caused by omission of expected reward or by delivery of unexpected punishment – can be used as a reliable signal for learning. This is because the baseline firing rate of dopamine neurons is low (3–8 spikes per second), leaving little dynamic range for encoding of events ranging from a slap on the wrist to the threat of a hungry lion by suppression of firing. An analogous aversive prediction error, perhaps signaled by serotonin, has been proposed [30], although recent data suggest that negative prediction errors can be signaled by the duration of the suppression of dopaminergic firing [31].

Q7: What about stimulus–stimulus learning? Is this driven by prediction errors as well, perhaps of a different type (event prediction errors rather than reward prediction errors)?

A central premise of the TD model is that predictions are about the sum of future rewards. This provides a powerful framework for understanding many aspects of conditioning and decision making, but it is also severely limited. One limitation is that the TD model (and similarly the Rescorla-Wagner model) does not address learning of relationships between events that do not have an affective component. Instead, such stimulus–stimulus learning is addressed by 'unsupervised learning' models.

To illustrate why TD learning cannot account for stimulus–stimulus learning, assume the situation in Figure 1 in which stimulus A (wine label) is followed by stimulus B (dark red liquid in a glass) and then by the consumption of good wine (reward utility = 1). After training the values of each of the stimuli (in terms of predicted future reward) will be 1, and so there will be a prediction error only when A appears unexpectedly (for instance, when you find that long-lost bottle in your cellar). Now assume that in separate 'conditioning', stimulus C (cold, frothy pale liquid) predicts a nice refreshing beer (utility = 1). The value of C after training will be the same as that of A and B. Now imagine a 'swap' trial in which you see A (wine label) followed by the pouring of C (pale liquid) and finally the flavor of beer. Most people would notice the unexpected occurrence of C – they might react behaviorally to it and might subsequently expect beer rather than wine. However, such a sequence of events would theoretically generate no TD reward prediction-error at the time of C or at the time of the reward because C predicts a reward with a scalar utility and a temporal delay that is equal to that predicted by B.

That an obviously surprising and unpredicted event fails to generate a prediction error exposes the limitation of the TD model: this model aims to predict the scalar utility of future rewards, not the identity of the reward (wine or beer) or the precise stimuli (B or C) that precede it. Because animals and humans presumably can learn which specific reward or stimulus follows which, it is clear that TD learning must be only one of several mechanisms for learning in the brain [32]. It is important to judge the TD learning mechanism according to its goals: it is a good way to learn to predict the value of future rewards, but only that.


OUR NEW STRATEGY HAS NEVER WORKED FOR ANYONE BEFORE.
THAT WILL GIVE US THE ELEMENT OF SURPRISE.

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Q8: Correlates of prediction errors in functional imaging studies are frequently found not in the midbrain but, rather, in areas such as the striatum, amygdala, and orbitofrontal cortex [5,6,9,33]. Do all these areas signal prediction errors?

This is a tricky issue that has caused much confusion. Imaging studies have indeed found blood-oxygen-level-dependent (BOLD) signals that correlate with a precise, computationally derived TD prediction error in a variety of brain areas. Furthermore, a handful of single-unit recording studies have reported that activity in other brain areas – amygdala, striatum, orbitofrontal cortex and elsewhere – is reliably modulated by whether rewards or punishments are expected [3,4,7,20,34,35] (also see [8] for an excellent review of these data).

However, current thought has it that the BOLD signal does not directly reflect firing activity in an area but, rather, correlates with the local field potential and local processing, which are driven by subthreshold activity and synaptic inputs to the area [36–38]. Thus, perhaps it is appropriate to view the imaging results as reflecting the information that an area is receiving and processing, whereas single-unit activity reflects the information that an area is transmitting to downstream regions. Moreover, dopamine can influence BOLD signals directly through its effects on local blood vessels [38,39] and on neuronal oscillations in target areas [40].

These considerations can explain why BOLD signals in striatal and prefrontal cortical areas, which are the primary recipients of dopaminergic inputs, resemble a prediction error. However, care must be taken when inferring that this signal reflects dopaminergic activity: striatal BOLD correlates just as well with prediction errors for aversive outcomes, with no sign change (i.e. a positive prediction error to an unexpected aversive event!) [41–43]. This is not what is seen in dopaminergic recordings, reminding us that the BOLD signal can be modulated by all afferent activity, be it dopaminergic or otherwise.

Accordingly, prediction errors are generally not seen in single-unit activity in areas that correlate with prediction errors in imaging studies. For example, neurons in ventral striatum seem to signal the value of rewards and cues that predict reward rather than prediction errors [44–48]. Similarly, activity in orbitofrontal cortex shows no evidence of prediction error signaling in a task that was effective at demonstrating such signaling in rat dopamine neurons[*].

A second issue is that neural activity can differ for expected and unexpected outcomes for a host of reasons other than signaling of reward prediction errors. For example, attention declines as events become expected. Changes in neural activity might reflect attentional modulation rather than prediction error signaling [49–51]. The real test is whether the neural activity meets all of the criteria laid out above for prediction error signaling. To date, single-unit studies that have examined encoding of prediction errors in nondopaminergic areas have only demonstrated a subset of the criteria, normally in a very small percentage of the neural population.

Of course other brain areas might ultimately be found to signal prediction errors or contribute to signaling of prediction errors by dopamine neurons. Clearly, dopamine

[*] Calu, D.J. *et al.* (2007) Orbitofrontal cortex does not signal reward prediction errors. *Society for Neuroscience Abstracts* 749.16/FFF32.

neurons base the computation of a prediction error on information coming from various sources. For example, information about expected values *V(t)* has been proposed to be conveyed to dopamine neurons by striatal afferents, whereas information regarding primary rewards might arise from the pedunculopontine nucleus [52] as well as the lateral habenula [20,53,54].

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

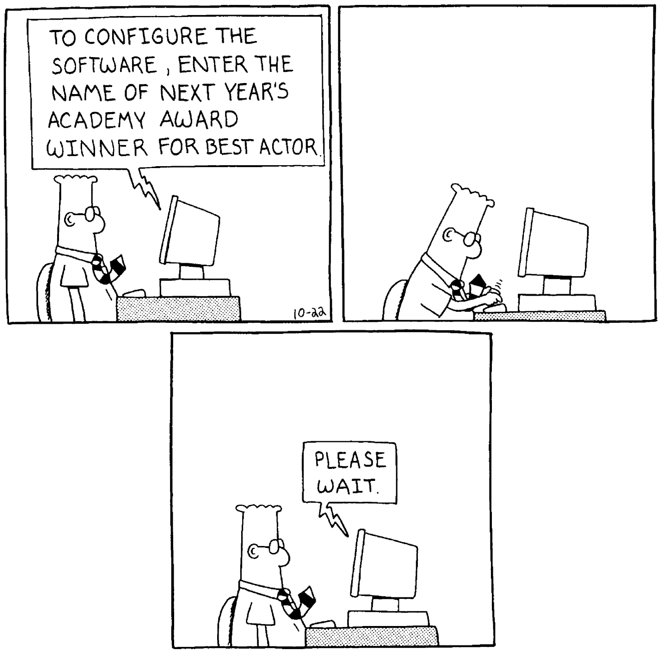
Q9: We have been talking about using TD errors to learn to predict the future, but these predictions would be rather useless if they could not influence behavior that preempts these future consequences. What is the relationship between prediction errors and action selection?

Although Pavlovian prediction learning is important (for instance, it supports innate approach behaviour to appetitive stimuli and withdrawal from aversive stimuli, which are no doubt helpful for survival), instrumental action selection is, in some ways, the hallmark of intelligence. This is because instrumental action is aimed at bringing about those rewarding outcomes that are available in the environment. Fortunately, TD prediction errors can assist in that as well. By supplying positive or negative evaluative signals even long before an actual outcome is realized, TD errors can solve the ‘credit-assignment problem’ of how to correctly apportion credit for reward to the different past actions. That is, the error signal can be used in lieu of the actual reward signal to ‘reinforce’ actions that lead to better states of the environment (in terms of future predicted rewards) and ‘punish’ those that lead to worse states [55,56].

Several suggestions exist for how action selection and prediction errors can be combined. According to one popular model, a critic learns to evaluate situations by using TD learning, whereas an actor maintains an action plan or policy in which the tendency to perform different actions is increased or decreased based on the prediction error signal that follows each action [57,58]. Other models suggest that predictive values are learned not for stimuli but, rather, for actions taken in the presence of different stimuli. The distinctions between these models are subtle, but they do have different properties. For instance, they suggest slightly different forms for the TD prediction error. Researchers have only recently started to examine which one of these models, if any, is implemented in the brain, and results so far are equivocal [23,59].

Q10: So, what does all this mean regarding the role of TD learning and dopamine in the brain? Is all learning and action selection dependent on these?

Definitely not! Although dopamine has an important role in conditioning, the kind of reinforcement learning that it is thought to support is most strongly associated with habitual learning and action selection. Goal-directed behavior, which probably uses representations of contingencies and rewards for forward-looking planning, might not be dependent on learning via dopamine and TD prediction errors [32,60,61]. Furthermore, wholly different categories of learning, such as perceptual, stimulus–stimulus and episodic learning, do not use TD prediction errors. So, the bottom line is that dopamine clearly holds an important role in learning and behavior. By using precise computational models, we can appreciate fully what this role is, as well as what it isn’t (Box 1).


TO CONFIGURE THE SOFTWARE, ENTER THE NAME OF NEXT YEAR'S ACADEMY AWARD WINNER FOR BEST ACTOR.
10-22
PLEASE WAIT.

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Box 1. What are some of the outstanding questions regarding prediction error encoding in the brain?

Do dopamine neurons really signal temporal difference prediction errors?

The evidence for this idea is strong; however, this is only a simplified model of dopamine function, and it is almost certainly ultimately incomplete. First, dopamine signals might serve more than one function and might have different roles in different target areas. Furthermore, there are alternatives to the view presented here, based in part on apparent exceptions in which dopamine neurons respond to novel cues or seem to generalize their responses to cues that do not predict reward. Finally, basic experimental predictions still remain untested. For example, transfer of a cue-evoked response from a primary-conditioned stimulus to a second-order cue in the absence of primary rewards remains to be shown. This would validate the idea that cue-evoked signaling has the same properties as firing triggered by primary reward. Outcome-substitution studies such as the beer/wine swap described earlier can test whether the prediction error is really only for the value but not the identity of the outcome. Seminal studies using blocking or similar paradigms also remain to be repeated with strongly aversive outcomes.

How is dopaminergic activity related to action selection?
The vast majority of the single-unit studies showing error encoding have used Pavlovian tasks (in which there is no response requirement) or imperative tasks (in which only a single response is available). Yet reinforcement learning is really about selecting actions, making choices and guiding behavior. Thus, a crucial area that needs to be addressed is how error signaling by dopamine neurons changes in these settings.

How are prediction errors computed in the brain?
On the one hand dopamine neurons do not seem to signal all prediction errors. On the other hand there is evidence that other brain areas might signal prediction errors. Moreover, it is currently unclear what is the basic network that allows the computation of prediction errors to take place. It is imperative that other suspect areas be investigated in awake, behaving animals using behavioral tasks that are optimal for identifying error signaling or its necessary building blocks.

Author note

This manuscript grew out of a set of e-mail questions Geoff sent Yael after they met at a reward circuits meeting held at Lake Arrowhead in California in 2006. The answers to the questions were so useful that they were often printed and passed around the laboratory. This generated further discussion and ideas and, ultimately, more questions. Because these dialogues were extraordinarily useful to the authors, they thought they also might be helpful to others.

Acknowledgements

The authors thank D. Calu, M. Delgado, Q. Huys and D. Schiller for helpful comments on this manuscript. This work was supported by the Human Frontiers Science Program (Y.N.) and by the National Institute on Drug Abuse and National Institute on Aging (G.S.).

References

1 Schultz, W. *et al.* (1997) A neural substrate for prediction and reward. *Science* 275, 1593–1599

2 Montague, P.R. *et al.* (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.* 16, 1936–1947

3 Feierstein, C.E. *et al.* (2006) Representation of spatial goals in rat orbitofrontal cortex. *Neuron* 51, 495–507

4 Belova, M.A. *et al.* (2007) Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron* 55, 970–984

5 O'Doherty, J. *et al.* (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454

6 Tobler, P.N. *et al.* (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. *J. Neurophysiol.* 95, 301–310

7 Matsumoto, M. *et al.* (2007) Medial prefrontal cell activity signaling prediction errors of action values. *Nat. Neurosci.* 10, 647–656

8 Schultz, W. and Dickinson, A. (2000) Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* 23, 473–500

9 Nobre, A.C. *et al.* (1999) Orbitofrontal cortex is activated during breaches of expectation in tasks of visual attention. *Nat. Neurosci.* 2, 11–12

10 Rescorla, R.A. and Wagner, A.R. (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (Black, A.H. and Prokasy, W.F., eds), pp. 64–99, Appleton-Century-Crofts

11 Kamin, L.J. (1969) Predictability, suprise, attention, and conditioning. In *Punishment and Aversive Behavior* (Campbell, B.A. and Church, R.M., eds), pp. 242–259, Appleton-Century-Crofts

12 Reynolds, G.S. (1961) Attention in the pigeon. *J. Exp. Anal. Behav.* 4, 203–208

13 Konorski, J. (1948) *Conditioned Reflexes and Neuron Organization.* Cambridge University Press

14 Rescorla, R.A. and LoLordo, V.M. (1968) Inhibition of avoidance behavior. *J. Comp. Physiol. Psychol.* 59, 406–412

15 Rescorla, R.A. (1970) Reduction in effectiveness of reinforcement after prior extinction conditioning. *Learn. Motiv.* 1, 372–381

16 Kremer, E.F. (1978) Rescorla-Wagner model - losses in associative strength in compound conditioned stimuli. *J. Exp. Psychol. Anim. Behav. Process.* 4, 22–36

17 Sutton, R.S. and Barto, A.G. (1990) Time-derivative models of Pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (Gabriel, M. and Moore, J., eds), pp. 497–537, MIT Press

18 Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An introduction.* MIT Press

19 Niv, Y. and Montague, P.R. (2008) Theoretical and empirical studies of learning. In *Neuroeconomics: Decision-Making and the Brain* (Glimcher, P.W. *et al.*, eds), pp. 329–349, Academic Press

20 Matsumoto, M. and Hikosaka, O. (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* 447, 1111–1115

21 Fiorillo, C.D. *et al.* (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898–1902

22 Tobler, P.N. *et al.* (2003) Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *J. Neurosci.* 23, 10402–10410

23 Roesch, M.R. *et al.* (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* 10, 1615–1624

24 Redgrave, P. *et al.* (1999) Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.* 22, 146–151

25 Berridge, K.C. (2007) The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl.)* 191, 391–431

26 Horvitz, J.C. (2000) Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* 96, 651–656

27 Ungless, M.A. *et al.* (2004) Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science* 303, 2040–2042

28 Coizet, V. *et al.* (2006) Nociceptive responses of midbrain dopaminergic neurons are modulated by the superior colliculus in the rat. *Neuroscience* 139, 1479–1493

29 Mirenowicz, J. and Schultz, W. (1996) Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* 379, 449–451

30 Daw, N.D. *et al.* (2002) Opponent interactions between serotonin and dopamine. *Neural Netw.* 15, 603–616

31 Bayer, H.M. *et al.* (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. *J. Neurophysiol.* 98, 1428–1439

32 Daw, N.D. *et al.* (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711

33 O'Doherty, J.P. *et al.* (2003) Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. *Neuron* 38, 329–337

34 Ramus, S.J. and Eichenbaum, H. (2000) Neural correlates of olfactory recognition memory in the rat orbitofrontal cortex. *J. Neurosci.* 20, 8199–8208

35 Tremblay, L. and Schultz, W. (2000) Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *J. Neurophysiol.* 83, 1864–1876

36 Logothetis, N.K. *et al.* (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157

37 Logothetis, N.K. and Wandell, B.A. (2004) Interpreting the BOLD signal. *Annu. Rev. Physiol.* 66, 735–769

38 Attwell, D. and Iadecola, C. (2002) The neural basis of functional imaging signals. *Trends Neurosci.* 25, 621–625

39 Krimer, L.S. *et al.* (1998) Dopaminergic regulation of cerebral cortical microcirculation. *Nat. Neurosci.* 1, 286–289

40 Costa, R.M. *et al.* (2006) Rapid alterations in corticostriatal ensemble coordination during acute dopamine-dependent motor dysfunction. *Neuron* 52, 359–369

41 Jensen, J. *et al.* (2007) Separate brain regions code for salience versus valence during reward prediction in humans. *Hum. Brain Mapp.* 28, 294–302

42 Seymour, B. *et al.* (2004) Temporal difference models describe higher order learning in humans. *Nature* 429, 664–667

43 Menon, M. *et al.* (2007) Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: effects of dopaminergic modulation. *Biol. Psychiatry* 62, 765–772

44 Carelli, R.M. (2002) Nucleus accumbens cell firing during goal-directed behaviors for cocaine vs. 'natural' reinforcement. *Physiol. Behav.* 76, 379–387

45 Carelli, R.M. and Wondolowski, J. (2003) Selective encoding of cocaine versus natural rewards by nucleus accumbens neurons is not related to chronic drug exposure. *J. Neurosci.* 23, 11214–11223

46 Roitman, M.F. *et al.* (2005) Nucleus accumbens neurons are innately tuned for rewarding and aversive taste stimuli, encode their predictors, and are linked to motor output. *Neuron* 45, 587–597

47 Nicola, S.M. *et al.* (2004) Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative task. *J. Neurophysiol.* 91, 1840–1865

48 Setlow, B. *et al.* (2003) Neural encoding in ventral striatum during olfactory discrimination learning. *Neuron* 38, 625–636

49 Gallagher, M. *et al.* (1990) The amygdala central nucleus and appetitive Pavlovian conditioning: lesions impair one class of conditioned behavior. *J. Neurosci.* 10, 1906–1911

50 Holland, P.C. and Gallagher, M. (1993) Amygdala central nucleus lesions disrupt increments, but not decrements, in conditioned stimulus processing. *Behav. Neurosci.* 107, 246–253

51 Han, J.S. *et al.* (1997) The role of an amygdalo-nigrostriatal pathway in associative learning. *J. Neurosci.* 17, 3913–3919

52 Kobayashi, Y. and Okada, K. (2007) Reward prediction error computation in the pedunculopontine tegmental nucleus neurons. *Ann. N. Y. Acad. Sci.* 1104, 310–323

53 Ji, H. and Shepard, P.D. (2007) Lateral habenula stimulation inhibits rat midbrain dopamine neurons through a GABA(A) receptor-mediated mechanism. *J. Neurosci.* 27, 6923–6930

54 Christoph, G.R. *et al.* (1986) Stimulation of the lateral habenula inhibits dopamine containing neurons in the substantia nigra and ventral tegmental area of the rat. *J. Neurosci.* 6, 613–619

55 Barto, A.G. (1995) Adaptive critics and the basal ganglia. In *Models of Information Processing in the Basal Ganglia* (Houk, J.C. *et al.*, eds), pp. 215–232, MIT Press

56 Sutton, R.S. (1988) Learning to predict by the method of temporal difference. *Mach. Learn.* 3, 9–44

57 Houk, J.C. *et al.* (1995) A model of how basal ganglia generate and use neural signals that predict reinforcement. In *Models of Information Processing in the Basal Ganglia* (Houk, J.C. *et al.*, eds), MIT Press

58 Joel, D. *et al.* (2002) Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* 15, 535–547

59 Morris, G. *et al.* (2006) Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* 9, 1057–1063

60 Dickinson, A. and Balleine, B.W. (1994) Motivational control of goal-directed action. *Anim. Learn. Behav.* 22, 1–18

61 Dayan, P. and Balleine, B.W. (2002) Reward, motivation, and reinforcement learning. *Neuron* 36, 285–298