Learning, Action, Inference and Neuromodulation

P Dayan, University College London, London, UK
N D Daw, New York University, New York, NY, USA
Y Niv, Princeton University, Princeton, NJ, USA

Introduction

Neuromodulators, including dopamine, norepinephrine, acetylcholine, and serotonin, exert a wide and diverse collection of effects on a broad range of cortical and subcortical target structures. They are also critically implicated in many aspects of normal behavior and both neurological and psychiatric conditions. With the exception of acetylcholine, they are synthesized by neurons whose somas lie in relatively small subcortical regions such as the ventral tegmental area and substantia nigra pars compacta (for dopamine), the locus coeruleus (for norepinephrine), and the dorsal and median raphe nuclei (for serotonin). However, these neurons can have massive axonal arbors, innervating large areas. The activity of the neuromodulatory neurons and the release of their associated neuromodulators (which is subject to many factors other than the momentary spiking rates) onto the bewildering diversity of their pre- and postsynaptic receptors exhibit a large number of incompletely understood temporal and spatial regularities.

There are obviously huge gaps in our knowledge about neuromodulators, and there are doubtless many as yet undiscovered complexities. However, in some key respects, they appear to admit relatively simple theoretical characterizations associated with learning predictions of affectively important outcomes such as rewards and punishments, learning to choose actions that optimize those outcomes, and regulating and controlling certain general aspects of information processing. These are the most basic functions associated with the survival of mobile, decision-making organisms, and it turns out that a very helpful normative underpinning for these theories can be derived from statistics, operations research, and engineering. It should be acknowledged at the outset, however, that these theories ignore many important functions of the neuromodulators, such as their role in regulating sleep–wake cycles. Nonetheless, they are influential in organizing large bodies of experimental results and also in compelling new empirical approaches.

Under these theoretical views, neuromodulators report on two main quantities: (1) information associated with positive and negative outcomes and (2) information associated with uncertainty. It seems, at least to a crude approximation, that dopamine and serotonin play a special role in positive and negative affect, with norepinephrine and acetylcholine being particularly involved with uncertainty. Further, orthogonal to this distinction, the neuromodulators also have two main classes of effects on neural processing: (1) influencing neural plasticity and learning and (2) regulating different sources of input to neurons and controlling competition between networks of neurons.

This article first provides a somewhat general characterization of issues surrounding learning and then briefly discusses immediate effects of the neuromodulators on inference and regulation.

Learning

Long-term synaptic plasticity, that is, persistent facilitation and depression of the efficacies of synapses, can be characterized at many different levels of biological realism and according to many different theoretical abstractions. Unfortunately, experimental difficulties have hindered efforts to understand completely the range of effects of neuromodulators on plasticity; therefore this article proceeds at a more abstract level appropriate to the relevant normative theories.

The central concept for the learning associated with neuromodulators is prediction. Consider a very simple experiment comprising many trials, in each of which a particular image may or may not be presented and a drop of pleasant-tasting fruit juice may or may not be delivered (regardless of what the subjects do). If the juice is provided more often with than without the image, then subjects learn to expect or predict the delivery of the juice when the image is presented. Such predictions can be measured via behaviors such as anticipatory salivation or licking, which are thought to be triggered more or less directly and automatically by the prediction of the reward. This is a very simple instance of what is known as classical or Pavlovian conditioning. The image is called a conditioned stimulus, the juice is known as an unconditioned stimulus, and the predictive behaviors are termed conditioned responses.

Trial-Based Prediction

Reinforcement learning theories formalize this prediction learning by positing that the presence of the image on trial n is coded by the activity of a unit $x^n = 1$ (with $x^n = 0$ if the image was not presented). This unit makes a connection, via an abstract synapse with strength w^n, to a neuron. This neuron also receives information about the delivery ($r^n = 1$) or nondelivery ($r^n = 0$) of juice on trial n. The value r^n would be scaled on the basis of the appetitive utility of the juice. The idea is that the synaptic connection should be modified according to a learning rule, such that the presence of the stimulus will predict the occurrence of the reward.

One rather general form of such a learning rule is

$$w^{n+1} = w^n + \alpha^n \delta^n x^n \qquad [1]$$

which indicates how w^n changes from one trial to the next. This update is a function of three quantities:

α^n (learning rate): scales the size of the change to the synapse.

δ^n (prediction error): incorporates information about unpredicted or incorrectly predicted outcomes and typically controls the direction of change

x^n (synaptic eligibility): here, limits synaptic change to those trials in which the image is presented ($x^n = 1$).

Different reinforcement learning theories posit different forms for these terms. Many early ideas in this area came from the field of mathematical psychology, where they were invented to describe the development of behavioral responding in conditioning experiments. Their relationship both with the neural data on the actual activities and concentrations of the neuromodulators and with normative computational ideas came substantially later.

One of the most important psychological learning rules is called the Rescorla–Wagner rule. This is based on the idea that the net prediction of juice on trial n is just the total presynaptic input, here $v^n = x^n \times w^n$, and derives learning from the prediction error, the difference between the juice that is provided and the juice that is predicted:

$$\delta^n = r^n - w^n \times x^n \qquad [2]$$

$$= r^n - v^n \qquad [3]$$

In its simplest form, the Rescorla–Wagner rule also suggests that the learning rate $\alpha^n = \alpha$ is constant. Figure 1(a) shows the operation of this rule, indicating how the values of r^n and x^n specify w^n over a whole set of trials.

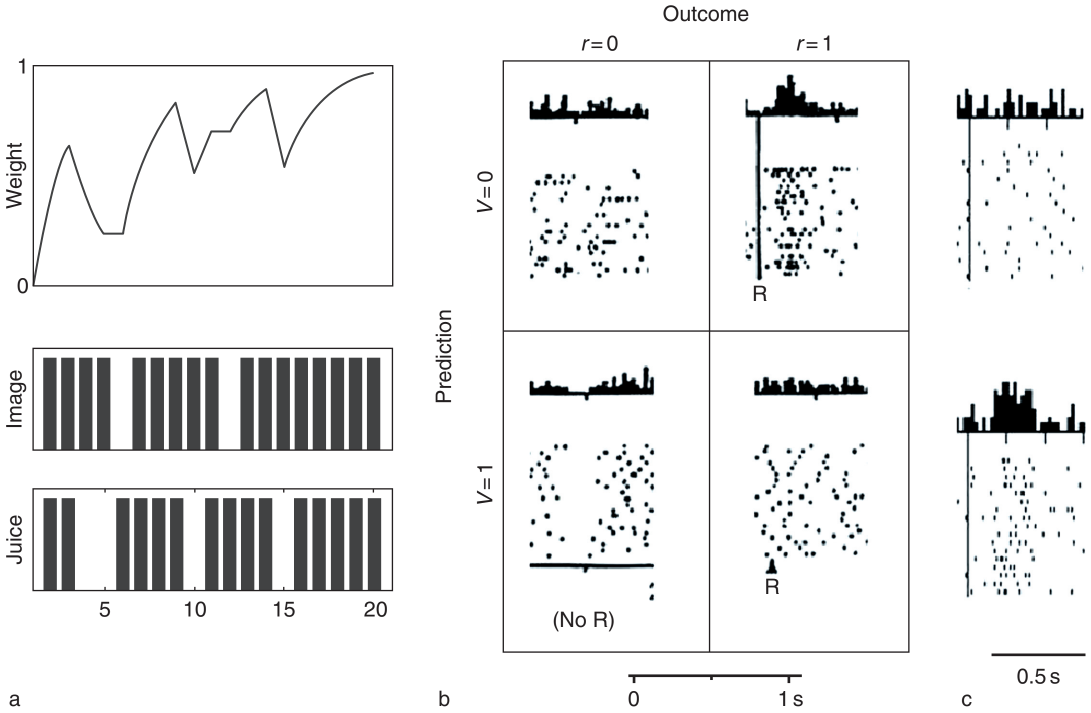

Figure 1 Dopamine and learning rules. (a) The course of learning using the Rescorla–Wagner rule. The top plot shows the evolution of the weight w^n over the course of the 20 trials with an image and juice being provided as in the bottom plot. The learning rate $\alpha = 0.4$. The stochastic nature of the learning to a level around the asymptotic value of 0.8 is apparent. (b) Evidence for a prediction error at the time of reward expectancy is shown in the activity of dopamine cells. The four boxes show rasters and accumulated histograms of the activity of identified dopamine neurons in macaque monkeys (during 20 trials), where the reward is a small drop of juice (provided at the time marked by R). The plots are in accord with the prediction error of eqn [3]. (c) The activity of dopamine neurons in response to the presentation (marked by vertical lines) of a conditioned stimulus which either does not (upper) or does (lower) predict the future delivery of juice. These are consistent only with temporal difference learning and not the Rescorla–Wagner rule. (b) Adapted from Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599. (c) Reprinted, with permission, from the Annual Review of Physiology, Volume 57 © 2006 by Annual Reviews www.annualreviews.org.

Although extremely simple, this rule has quite a number of desirable computational properties and an excellent normative basis in engineering (where it is more usually known as the delta rule). Most important, it leads to correct predictions, at least on average. In particular, if the juice is always provided with the image, then the weight w^n will come to take the value 1; if the juice is provided on half the trials with the image, w^n will take the value of 0.5, and if the juice is never provided, then w^n will become 0. The rule also has a decent evidentiary trail in psychology, capturing many empirical phenomena surrounding the acquisition and extinction of conditioned behaviors in variants of the simple Pavlovian conditioning experiment discussed here.

Somewhat remarkably, the error signal δ^n in this form of the rule also appears to capture well some, but crucially not all, of the characteristics of the phasic spiking activity of dopamine cells as recorded electrophysiologically in the midbrain during a learning task like this. Figure 1(b) shows an example of this match. The figure shows a raster plot of the activity of a dopamine neuron in an awake, behaving macaque monkey at the time of the delivery or nondelivery of an unexpected or expected reward. If, indeed, the prediction error occasioned by delivery or nondelivery of juice is reflected by phasic bursts of activity above or below the slow, irregular baseline of neural firing, then from eqn [3], one should expect the following: suprabaseline activity for the delivery of the unexpected reward ($\delta^n = 1 - 0 = 1$; for instance, at the outset of learning, when the participant cannot make an accurate prediction); infrabaseline activity for the nondelivery of the expected reward ($\delta^n = 0 - 1 = {}^-1$); and, crucially, baseline activity when an expected reward is provided ($\delta^n = 1 - 1 = 0$). We would also expect baseline activity for the nondelivery of a nonexpected reward ($\delta^n = 0 - 0 = 0$). The figure, although not showing data collected in exactly this manner, shows just these expectations to be fulfilled. Note, though, that dopamine activity does not offer a completely linear code for δ^n. The potential fidelity of the coding of negative prediction errors is limited by the fairly low baseline activity of the neurons; there is also some evidence that the coding is adaptive, moulding itself to the appropriate statistical contingencies.

In sum, at the time at which reward delivery is expected, dopamine neurons appear to convey a form of prediction error for the delivery of reward, a signal that is exactly appropriate for training predictions (as in eqn [1]). This finding leads to a large range of questions about what happens when the temporal relationship between the image and the juice is manipulated (the data here showing that the Rescorla–Wagner rule is critically flawed) or when there are multiple predictive images or stimuli, how prediction learning and action learning are coupled, and how punishments (e.g., small electric shocks) are processed. There are also some data about the neural structures involved in the various components of this learning rule (i.e., x^n and w^n), notably implicating nuclei in the amygdala and the orbitofrontal cortex, together with their connections to the nucleus accumbens, all structures that receive a strong dopamine projection.

Temporal Prediction

The use of the word prediction hints at the critical role that time plays in conditioning experiments. Broadly speaking, to be a good predictor, it is important that the image precede (rather than coincide with or follow) the juice. This leads to both conceptual and psychological/neural problems. The conceptual problem is that it is necessary to change the goal of prediction – there is no point in predicting the amount of juice that will be provided at the same time as the image, since they are not delivered at the same time; rather, it is necessary to predict the amount that will come in the future. An important step in the field was the suggestion by Richard Sutton and Andrew Barto that participants should try to predict a long-run measure of the reward, such as its sum over a whole trial. Achieving this requires a different prediction error from that in eqn [3]. It also turns out that this new prediction error accounts for more features of the activity of dopamine neurons, specifically their firing patterns at the time of the image presentation.

To understand this new prediction error, one needs a new variable t to count time within each trial. For instance, the image might be presented at time $t = 2$, the juice at time $t = 5$, with the trial ending at $t = T$. Now the reward on trial n is indexed by time $r^n(t)$ with, for instance, $r^n(5) = 1$ on trials on which the juice is given. The new task of the participants is to predict $v^n(t)$ at time t within a trial as the sum of all the future rewards that will be provided on average during that trial after t. Assuming for the moment that rewards are delivered deterministically, this means that the following relationship should hold (for which is used the symbol $\triangleq$):

$$v^n(t) \triangleq r^n(t) + r^n(t+1) + r^n(t+2) + \ldots + r^n(T) \qquad [4]$$

The obvious trouble with this definition as a goal for prediction is that the right-hand side, measuring the actual delivery of juice, cannot be assessed until the end of the trial. The solution to this problem, which underlies the new prediction error, is to observe that the sum can be split into its first and subsequent terms:

$$v^n(t) \triangleq r^n(t) + [r^n(t+1) + r^n(t+2) + \ldots + r^n(T)] \quad [5]$$

so if the prediction at time $t + 1$, $v^n(t + 1)$, is, as it should be, equal to $r^n(t + 1) + r^n(t + 2) + \ldots r^n(T)$, then,

$$v^n(t) \triangleq r^n(t) + v^n(t+1) \quad [6]$$

The new prediction error becomes the discrepancy between the left- and right-hand sides and can be computed after waiting just one timestep, without waiting for the end of the trial:

$$\delta^n(t) = r^n(t) + v^n(t+1) - v^n(t) \quad [7]$$

This is called the temporal difference prediction error because of the key role played by the change in the prediction over consecutive timesteps ($v^n(t + 1)- v^n(t)$). Even though, during learning, it cannot be guaranteed that $v^n(t + 1) = r^n(t + 1) + r^n(t + 2) + \ldots + r^n(T)$, it turns out that a learning rule like eqn [1], using $\delta^n(t)$ to update the weights associated with the prediction made at time t will also ultimately arrive at the correct values. Learning of this sort is said to involve bootstrapping because the predictions at one time are learned partly on the basis of predictions at the subsequent time.

The most important practical difference between the previous and present prediction errors (i.e., between eqns [3] and [7]) is the value of the new prediction error associated with the first presentation of the image (in this case, $\delta^n(1) = r^n(1) + v^n(2) - v^n(1) = 0 + v^n(2) - 0 = v^n(2)$). If the image is always followed by juice, then, once learning is complete, the prediction changes from $v^n(1) = 0$ to $v^n(2) = 1$ when the image is provided, resulting in a transient prediction error at time $t = 1$, $\delta^n(1) = 1$. The prediction then remains at 1 until the reward is actually delivered, so there is no difference between $v^n(t)$ and $v^n(t + 1)$, and therefore $\delta^n(t) = 0$ for all t preceding the time of reward. At the reward time, (here $t = 5$), no further reward is expected in the trial, so $v^n(6) = 0$, and the error from eqn [7] is $\delta^n(5) = r^n(5) - v^n(5)$. Hence, this prediction error behaves similarly to that in the Rescorla–Wagner rule for all the situations shown in Figure 1(b).

Figure 1(c) shows that the activity of dopamine neurons at the time of the image follows exactly the pattern expected from eqn [7]. It is as if the activity that at the beginning of training happens at the time of the unexpected reward moves backward in time across the trial so that it is instead initiated by the earlier reliable predictor of that reward, namely, the image. The transient activity at the time of the image can play a number of important functions, notably, allowing a predictor of reward to act like a surrogate reward itself, an effect known in psychology as conditioned reinforcement.

The neural problem associated with temporal prediction is exactly that the predictors need to be able to count time; in this case, the time since the image was presented. The well-timed depression in the activity of the dopamine neurons when an expected reward is not delivered is clear evidence that exactly this happens (and incidentally argues strongly against certain other construals of the dopamine activity, for instance that it flags all 'salient' events without regard to their appetitive or aversive valence). However, although there are various suggestions about the involvement of the cerebellum, the hippocampus, the striatum, and the parietal and frontal cortices in counting time, the exact nature of this signal is not well understood.

Multiple Predictive Stimuli

So far only the case of just one potentially predictive image has been considered. Now consider the case in which there are multiple predictive stimuli. It turns out that only the temporally simple case (associated with the Rescorla–Wagner rule) has to be treated, since the issues about time t within a trial are orthogonal to those about the multiple stimuli.

The Rescorla–Wagner version of the learning rule (eqn [1]) was designed to address psychological phenomena that arise when there are multiple predictive stimuli. The main idea is to elaborate it to the following:

$$w_i^{n+1} = w_i^n + \alpha_i^n \delta^n x_i^n \quad [8]$$

where there are now multiple stimuli x_i^n, each with its own prediction weight w_i^n, but, critically, still with only one global prediction error δ^n.

The Rescorla–Wagner rule involves the net prediction v^n of juice, defined as the sum of the predictions associated with each stimulus:

$$v^n = x_1^n \times w_1^n + x_2^n \times w_2^n + \ldots \quad [9]$$

If, perhaps as a result of prior learning, one of the stimuli present on a trial already predicts the reward correctly, then the prediction error (of eqn [3]) will be 0 ($\delta^n = 0$), and therefore the synapses associated with any other stimuli which are also present, including those that do not yet predict the reward, will not change. This effect is known in the psychological literature as blocking. The ample behavioral evidence for blocking was taken as an early sign that learning is indeed driven by prediction errors rather than, for instance, simply correlations. There is also some evidence for blocking in the activities of dopamine neurons.

Stimulus multiplicity could also affect the other components of the learning rule. In particular, a number of competing learning rules that have been suggested on the basis of psychological data suggest that there are stimulus-specific predictions (and prediction errors) and invoke stimulus-specific (and changing) learning rates α_i^n. These rules treat the learning rates as associabilities, reflecting a measure of the attention that the learners lavish on the stimuli. The most celebrated of these accounts, due to John Pearce and Geoffrey Hall, suggests that stimuli should be associable to the degree that they have had surprising consequences. In this view, multiple stimuli can interact with one another to produce learning phenomena such as blocking at least partly through competition for these associabilities. This account has attracted many experimental and theoretical studies.

Experimentally, Peter Holland, Michela Gallagher, and their colleagues have collected evidence from lesion, pharmacological, and behavioral studies suggesting that associabilities are realized (at least in rats) through a pathway involving the central nucleus of the amygdala, attentional and control regions of the cortex, and, notably for the topic of this article, various parts of the acetylcholine system. Oddly, there appear to be different anatomical substrates for increases and decreases in stimulus associability.

From a theoretical perspective, the sort of surprise inherent in Pearce and Hall's suggestion can be formulated as a particular form of uncertainty (a fact that, together with Holland and Gallagher's six results and others, led to the suggestion that such uncertainty might be reported by acetylcholine). Normative theories of prediction learning arising in the fields of statistics and engineering actually suggest learning rules rather like eqn [8] and provide a precise characterization of what factors should control α_i^n. The best known and simplest of these theories, called the Kalman filter, suggests that the learning rates act to divide up the impact of the prediction error δ^n on the weights w_i^n for each of the stimuli such that the stimuli whose weights are least well known (i.e., most uncertain) are most susceptible to learning. That is, the more uncertain the prediction made by one stimulus (for instance, because it has been observed only a few times), the more it is deemed responsible for any nonzero prediction error δ^n, and so the more its prediction w_i^n changes. Variants of this theory also link a stimulus's uncertainty to the degree of surprise that has accompanied it, as for associability in the Pearce–Hall rule. Additionally, in noisy environments in which stimuli and rewards are inherently unreliable, the learner should expect substantial prediction errors even if its predictions are as good as can be. This should have the effect of reducing all the learning rates.

Although the Kalman filter certainly captures something of the flavor of the Pearce–Hall learning rule, its detailed predictions and putative cholinergic substrates have yet to be fully tested, either behaviorally or neurally. Norepinephrine also seems to exert rather general effects on the overall speed of learning, particularly in circumstances in which there are gross changes to an environment. One theoretical account of this suggests that norepinephrine reports a different form of uncertainty, that associated with the whole context. This is coupled to learning since contextual change typically induces a need for gross revision or fresh acquisition of predictions. There are two final issues associated with multiple stimuli. First, the task of combining the predictions of reward made by multiple stimuli can be considered an example of the more general statistical problem of combining multiple sources of evidence. There is ample psychological data suggesting that, as in most statistical accounts, this sort of combination is typically not additive (as in eqn [9]) but rather is competitive, with the predictions made by different stimuli being averaged together and the prediction of the most reliable (least uncertain) predictor being weighted most heavily. How important this is for prediction learning, and indeed what its neural substrate is, are presently unclear.

Second, the notion that each stimulus has its own independent, binary input feature x_i is too simple, particularly if the stimulus representations associated with prediction learning are those in refined areas such as prefrontal cortex. Rather, sophisticated cortical (and perhaps hippocampal) mechanisms will play a critical role in creating stimulus representations that are appropriately tailored to sensory experience.

Action Learning

So far, this article has considered prediction learning but not the selection of actions based on such predictions. Although the prediction of upcoming reward can in itself directly trigger conditioned behavioral responses, such as approach, or the salivation of Pavlov's dogs, these represent a fairly limited portion of the behavioral repertoire. Animals can also learn to take arbitrary actions (such as pressing a lever) in order to obtain rewards that are contingent on such actions. This learning is known as instrumental conditioning, and its neural substrates have actually been more extensively investigated than those of Pavlovian conditioning. This area has revealed some important complexities; for instance, in rats, an instrumentally conditioned behavior such as a lever press can apparently arise through either of at least two neurally and behaviorally dissociable pathways. One of these involves at least the prelimbic prefrontal cortex and the dorsomedial striatum, may be somewhat independent of dopaminergic neuromodulation, and produces actions that can behaviorally be shown to be sensitive to the particular reward (such as food) expected. Such actions are therefore known as 'goal-directed.' The other pathway involves the dorsolateral striatum and the infralimbic prefrontal cortex, is strongly dependent on dopamine, and produces so-called 'habitual' actions. These are behaviorally characterized by their insensitivity to the particular reward expected; for instance, a rat may habitually press a food delivery lever even if it is not hungry.

There is both behavioral and neural evidence that the habitual action system is based on prediction learning of almost exactly the sort described above. This is consonant with a much larger body of evidence that dopamine is involved in action choice; besides habitual instrumental action, the neuromodulator is associated with addictive drugs, with self-stimulation experiments in which animals will work to receive electrical stimulation in some brain areas, and with motor pathologies such as Parkinson's disease. One particular view of action learning has recently received strong support from data on the activity of dopamine neurons in a task which involves monkeys' making choices between alternatives associated with different probabilities of reward. Under this scheme, the predictions should be not of the reward that will accrue after time t (written as $v^n(t)$ in eqn [4]), but rather of those that will accrue after time t if the learner chooses a particular action (say, a (t)) at time t. These so-called Q-values (written as $Q(t,a(t))$ were originally suggested by Christopher Watkins. If such predictions were known perfectly, action selection could ensue simply by choosing the action whose value is greatest, that is, the one predictive of the most reward. For a number of reasons, it can be prudent to choose somewhat randomly, assigning a greater chance to actions with greater Q-values:

$$P(a(t)=a)=\frac{e^{\beta Q(t,a)}}{\sum_b e^{\beta Q(t,b)}} \qquad [10]$$

Here, $\beta > 0$ controls how fierce the competition is; the larger β, the more likely the action with the largest Q-value will be chosen. One reason to employ such a probabilistic rule is to ensure exploration of unfamiliar alternatives and to avoid behavioral rigidity. It has been suggested that norepinephrine regulates exploration by controlling β, coupling the long-term success of a behavioral policy to reduced exploration and long-term failure to enhanced exploration. There is also starting to be some evidence (from functional imaging in humans) about the involvement of neural structures such as the frontal pole and the intraparietal sulcus in regulating exploration on a trial-by-trial basis.

The recent evidence from the activity of dopamine neurons during choice between several possibly rewarding actions favors a particular temporal-difference rule for learning the Q-values. Just as the prediction error $\delta^n(t)$ in eqn [7] is based on the difference between two successive predictions $v^n(t+1)$ and $v^n(t)$, this new prediction error is based on the difference between two successive predictions, $Q^n(t+1,a(t+1))$ and $Q^n(t, a(t))$, depending on the actions that are actually executed at times t and $t+1$:

$$\delta^n(t)=r^n(t)+Q^n(t+1,a(t+1))-Q^n(t,a(t)) \qquad [11]$$

This prediction error can then be used at the heart of a learning rule just like that in eqn [1], with the prediction $Q^n(t,a(t))$ being updated on the basis of this prediction error.

Appetitive and Aversive Predictions

This article has discussed appetitive predictions and the role of dopamine. However, in many circumstances, it is actually even more important to predict aversive outcomes. Further, animals can certainly learn to execute particular actions to avoid punishments and can titrate their actions to balance the benefits and costs of rewards and punishments associated with particular actions. The difference between appetitive and aversive predictions and outcomes is completely transparent to the normative theories, which happily consider them to live at positive and negative segments of a single dimension of value. However, there is a huge body of psychological and neural data showing that animals treat them very differently. There are also many open controversies in the literature, including a very active debate about the role of dopamine in aversively motivated behaviors. Aversive predictions play a key role in psychiatric conditions such as anxiety disorders (whose basis may be nonfactual predictions of threat) and paranoid schizophrenia.

These various data have yet to be encompassed in the sort of integrated computational, psychological, and neurobiological account outlined here for appetitive prediction learning and action choice. A popular psychological idea is that separate, but interacting, opponent systems are involved in training appetitive and aversive predictions. One possibility, suggested by William Deakin and Frederico Graeff and still somewhat speculative, is that one part of the serotonin system acts as the aversive opponent to dopamine. The data supporting this idea mainly originate in studies of the behavior of animals under pharmacological manipulation of serotonergic and dopaminergic function. Full tests of this hypothesis using electrophysiological recordings of serotonergic neurons are only just beginning.

Immediate Effects of Neuromodulators

This article has emphasized the effects of neuromodulators on learning. However, a more traditional view of neuromodulation, established most firmly in studies of central pattern generation in invertebrates, is that it regulates connectivity and excitability over the short term, allowing multiple functional networks to share a single anatomical substrate. Even in the context of the sorts of conditioning functions discussed here, there is indeed evidence that the neuromodulators exert short-term effects not mediated through learning. A few, in particular, have been most extensively investigated, including the impact of dopamine on working memory and on the vigor of responding, and the role of acetylcholine and norepinephrine in regulating different sources of information associated with making predictions.

There is quite some evidence that dopamine, like other neuromodulators, can immediately influence the excitability of neurons, often characterized by the gain of their input–output functions. If the neurons are connected in competitive networks (for instance, subserving the choice between multiple possible actions), then changing the gain (a bit like changing β in eqn [10]) can change the dynamics and stability of the competition. The effect of this on action choice has been studied in some detail by Jonathan Cohen and his colleagues. In prefrontal cortex, a similar effect of dopamine has been suggested as being important for gating the short-term storage of information (such as the fact that an image predicting reward was recently presented). As shown, this sort of storage plays a critical role in making appropriate predictions and therefore also in choosing appropriate actions.

Dopamine also appears to control the vigor or the force of behavioral responses. For instance, the overt effects of dopaminergic agonists such as amphetamine are hyperactivity; antagonists produce lethargy and affect the propensity to choose actions that have high costs. These effects also appear to be immediate, in that learning is not required for them to be exhibited. The reinforcement learning theories so far described do not encompass such phenomena, since they are designed to address choice among a discrete set of candidate actions and lack any analogue of a response vigor. A recent extension of the standard reinforcement learning picture accommodates free-operant tasks, in which subjects choose the latency of each of their responses, balancing the costs and benefits of vigorous responding. The model argues that tonic levels of dopamine report the average rate of reward and act as a form of opportunity cost which determines the optimal response vigor. The greater the opportunity cost, the greater the expectation of reward per unit time, and so the faster the participant must act to achieve this. This also implicates tonic dopamine in mediating the effects of motivation on response vigor. For instance, satiety, reducing the value of each reward, would decrease the net average rate of reward and thus result in lower tonic dopamine levels and more-sluggish responding.

Finally, if acetylcholine reports on the learner's uncertainty about aspects of its environment, then from a statistical viewpoint, it should influence not only learning but also the interpretation of sensory input – in statistical terms, inference. In particular, if uncertainty is high, then 'top-down' information about prior beliefs and expectations should be less trusted than 'bottom-up' information from sensation. This involves just the sort of regulation of connectivity that is implied by invertebrate neuromodulation. There is indeed substantial evidence that acetylcholine can boost the impact of bottom-up information compared with top-down information. Concomitantly, if norepinephrine reports on contextual change, then it should also have the effect of regulating the use of top-down information.

Summary

Neuromodulators are involved in some of the most basic adaptive functions of an organism, notably how it behaves and learns to behave appropriately in light of rewards, punishments, and other information. Dopamine and, putatively, serotonin appear to report prediction errors for future reward and punishment, to influence the acquisition of predictions and appropriate habitual actions, and to effect motivationally appropriate responding. Acetylcholine and norepinephrine seem to report on forms of uncertainty to control the allocation of learning among different possible predictors, the overall speed of learning, and the way that information from sensation is integrated with information based on past experience in an environment.

Accounts adapted from statistics, operations research, computer science, and engineering offer a normative underpinning for both the behavioral data and at least some key aspects of the neural substrates.

See also: Acetylcholine Neurotransmission in CNS; Basal Ganglia: Acetylcholine Interactions and Behavior; Basal Ganglia: Habit; Conditioning: Theories; Delayed Reinforcement: Neuroscience; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Neuroendocrine

Control of Energy Balance (Central Circuits/Mechanisms); Neuromodulation; Neutrotransmission and Neuromodulation: Acetylcholine; Operant Conditioning of Reflexes; Prediction Errors in Neural Processing: Imaging in Humans; Procedural Learning: Classical Conditioning; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Sleep and Sleep States: Cytokines and Neuromodulation.

Further Reading

Aston-Jones G and Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Reviews of Neuroscience* 28: 403–450.

Braver TS, Barch DM, and Cohen JD (1999) Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry* 46: 312–328.

Daw ND, Kakade S, and Dayan P (2002) Opponent interactions between serotonin and dopamine. *Neural Networks* 15: 603–616.

Dayan P, Kakade S, and Montague PR (2000) Learning and selective attention. *Nature Neuroscience* 3: 1218–1223.

Deakin JFW and Graeff FG (1991) 5-HT and mechanisms of defence. *Journal of Psychopharmacology* 5: 305–316.

Dickinson A (1980) *Contemporary Animal Learning Theory.* Cambridge, UK: Cambridge University Press.

Doya K (2000) Meta-learning, neuromodulation and emotion. In: Hatano G, Okada N, and Tanabe H (eds.) *Affective Minds*, pp. 101–104. Amsterdam: Elsevier Science.

Holland PC and Gallagher M (1999) Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences* 3: 65–73.

Pearce JM and Hall G (1980) A model for Pavlovian learning: Variation in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review* 87: 532–552.

Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Sutton RS and Barto AG (1998) *Reinforcement Learning.* Cambridge, MA: MIT Press.

Yu AJ and Dayan P (2005) Uncertainty, neuromodulation, and attention. *Neuron* 46: 481–492.