CHAPTER 11

Does the Dopaminergic Error Signal Act Like a Cached-Value Prediction Error?

Melissa J. Sharpe[1,2,3], **Geoffrey Schoenbaum**[1,4,5]

[1]National Institute on Drug Abuse, National Institute of Health, Baltimore, MD, United States; [2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, United States; [3]School of Psychology, UNSW Australia, Sydney, NSW, Australia; [4]Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, MD, United States; [5]Solomon H. Snyder Department of Neuroscience, The John Hopkins University, Baltimore, MD, United States

INTRODUCTION

The finding that dopaminergic neurons in the midbrain signal errors in prediction when an unexpected reward is delivered has transformed the study of behavioral neuroscience. This is because the concept of prediction errors had been the lynch pin of models of reinforcement learning for 50 years before this signal was discovered in the brain (Bush & Mosteller, 1951; Estes, 1950; Rescorla & Wagner, 1972; Sutton & Barto, 1981). Errors in outcome prediction—colloquially referred to as a "surprise" signal—are argued to drive learning about the antecedent stimuli that predict their occurrence. Essentially, the prediction error in these models acts as the teaching signal, which underlies the development of complex relationships between events in our environment (Holland & Rescorla, 1975; Miller & Matzel, 1988; Rescorla, 1973; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). The discovery that this signal actually exists in the brain gave some street credibility to associative models of reinforcement learning.

Yet when this prediction error signal was discovered in the midbrain, it was interpreted in a manner that diverged from the concept of driving real-world associations. Instead, the neural instantiation of this signal was taken to be synonymous with the error contained in the model-free reinforcement algorithm described by Sutton and Barto (1981, 1987, 1998). Here, reinforcement learning consists of the transference of what is termed "cached" value back from the reward to the stimulus, which reliably predicts reward occurrence. Cached value is the quantitative representation of the value presumed to be inherent in the reward. Essentially, an idea of how good a reward is to the subject, divorced from any specific knowledge of the identity or sensory properties of the reward itself. This allows the cue to become endowed with the scalar value, which drives motivated behavior in response to the reward-predictive cue in the future. However, this mechanism does not envision the development of an explicit association between the cue and the reward it predicts.

Goal-Directed Decision Making
ISBN 978-0-12-812098-9, https://doi.org/10.1016/B978-0-12-812098-9.00011-5
© 2018 Elsevier Inc.
All rights reserved.

There is now a host of studies in which phasic activity of dopamine neurons appears to reflect or correlate with errors in cached value. However, in each study that has shown this correlation, it is also possible that these signals reflect errors in the prediction of the specific features of the reward, which are related to value, but exist independent of it. Indeed, dopaminergic error signals have recently been shown to cues that could only acquire an expectation for reward through the deployment of rich associative models of the world (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Sadacca, Jones, & Schoenbaum, 2016). Further, dopamine prediction errors appear to be both sufficient and necessary for acquisition of these more complex model-based associations (Sharpe et al., 2017). Such research suggests that the dopamine prediction error is not synonymous with the model-free reinforcement learning algorithm described by Sutton and Barto (1987). Instead, research implicating the dopamine prediction error in these more complex learning phenomena encourages a step back to the spirit of traditional models of reinforcement learning, which perceive the prediction error as the catalyst for learning about relationships between events in the world.

DOPAMINE NEURONS IN THE MIDBRAIN RESPOND TO REWARD AND REWARD-PAIRED CUES

Perhaps, the first suggestion that dopamine plays a role in reward processing was the finding that rodents will perform an action to receive intracranial stimulation of dopamine neurons in the midbrain (Olds & Milner, 1954; Wise & Rompre, 1989). The propensity to respond was proportional to the frequency of stimulation; rodents responded at a higher frequency for a delivery of a higher frequency of stimulation. Such findings led to the wide-held belief that dopamine functions in the brain to allow natural rewards to possess powerful control over behavior, a concept supported by findings showing that selective dopaminergic lesions of the ventral tegmental area (VTA) reduced subjects desire to pursue natural rewards (Robbins & Everitt, 1982; Spyraki, Fibiger, & Phillips, 1982; Stricker & Zigmond, 1974; Ungerstedt, 1971). The midbrain dopamine system was conceptualized as a system, which registers the receipt of something valuable to a subject, like food or water in a deprived state, which increases the likelihood that a subject will seek out these items in the future and promote survival (Wise, 1987; Wise & Rompre, 1989; Yokel & Wise, 1975).

Early recording studies in nonhuman primates corroborated a role for dopamine in registering the reinforcing effects of natural rewards (Romo & Schultz, 1990; Schultz, 1986). Such studies showed that dopamine neurons in the VTA respond when subjects receive food in a food-restricted state (Romo & Schultz, 1990; Schultz, 1986). Further, the response of dopamine neurons to receipt of reward correlated with the magnitude of the reward received (Romo & Schultz, 1990; Schultz, 1986; Schultz, Apicella, & Ljungberg, 1993). That is, if a subject received a greater amount of reward, dopamine

neurons increased their response accordingly. And since these early studies, it has been demonstrated that dopamine neurons encode not only the magnitude of reward but also the subjective preference of reward (Fiorillo, Tobler, & Schultz, 2003; Lak, Stauffer, & Schultz, 2014; Stauffer, Lak, & Schultz, 2014). Dopamine neurons showed a preference for a particular reward over another in a manner that reflects the subject's observed choice preference for reward, regardless of caloric content. Such studies demonstrated that dopamine neurons register something about reward in a manner that reflects the subject's desire for that reward.

However, recording activity in dopamine neurons across the course of learning painted a more complex picture of the role of dopamine in reward. Specifically, recordings made across the course of conditioning showed that while dopamine neurons exhibit a phasic increase in activity during reward when nonhuman primates first experience reward (Mirenowicz & Schultz, 1994; Schultz et al., 1993), across time this phasic signal-to-reward receipt waned, appearing to transfer to the predictive cue with learning (Hollerman & Schultz, 1998; Ljungberg, Apicella, & Schultz, 1992; Mirenowicz & Schultz, 1994). Further, this transfer of phasic activity was progressive and occurred over successive pairings of the cue with reward, proportionate to increases in the appetitive response to the reward-paired cue (Mirenowicz & Schultz, 1994). Finally, once the subject learned the relationship between the cue and the reward, the omission of the expected reward elicited a depression in firing of these neurons at the time the reward was expected to occur (Mirenowicz & Schultz, 1994). These recording studies suggested that dopamine does not encode reward per se. Rather, it appeared that dopamine encodes some aspects of the predictive relationship between cues in the environment and rewards.

PREDICTION ERRORS IN MODELS OF REINFORCEMENT LEARNING

At the same time that researchers began to posit that dopamine was important for processing rewards, a parallel literature was being formed, which attempted to understand how natural rewards can function to reinforce responses result in procurement of those rewards. Some of the earliest mathematical models of this nature were developed to explain instrumental conditioning, where subjects learn to perform a particular action, which results in delivery of a particular reward (Bush & Mosteller, 1951; Estes, 1950; Estes & Burke, 1953). These models emerged out of the tradition of Thorndike (1898), in which it was argued that the pleasure derived from reward receipt increased the probability of a response being made in the presence of a particular stimulus. That is, procurement of reward was thought to make an agent more likely to perform a response again due to the reward's reinforcement of the association between the stimulus and the response.

Early mathematical models were simply an attempt to reconcile this idea with the incremental nature of the learning that takes place in instrumental tasks (Bush & Mosteller, 1951;

Estes, 1950; Estes & Burke, 1953; Thorndike, 1898). To do this, they used the standard linear operator, in which learning is governed by the discrepancy between the current probability of making a response and the maximal response probability (Bush & Mosteller, 1951). Here, the maximal response probability is determined by the reinforcer resulting from performance of the response and the effort required to procure that reinforcer. The learning rate parameter acted on the linear operator to govern the proportion of learning occurring on any one trial, allowing for a gradual increase in the probability of a response across the course of learning. Significantly, such models assumed that an increment in one particular response is completely independent of learning another response—even if both lead to the same eventual goal (Bush & Mosteller, 1951). Thus, a reliable correlation between a response and reward in the presence of a particular stimulus was enough to stamp in an association between that particular stimulus and the response, regardless of how many other responses could be made to procure the same reward or if that particular response was causally related to producing a reward. Rather, if a response was produced and was followed by reward, the subject would be more likely to make that particular response again.

Not long thereafter, however, the Kamin (1969) blocking experiments demonstrated that a correlation between events was not sufficient to produce learning. In the prototypical example, Kamin (1969) showed that pairing one stimulus (e.g., cue A) with reward would subsequently block learning about a second stimulus (e.g., cue X) if they were presented in compound (cue AX) with reward. In this example, cue A would have a high probability of producing a response when tested later separately, whereas cue X had a low probability. With these experiments, the notion of "surprise" began to enter the associative conversation. That is, in order for an association between events to form, the subject had to be surprised by the consequence. Learning required an error in the prediction of the reward. And so was born the notion that learning is driven by errors in prediction.

This notion initially took the form of the Rescorla and Wagner (1972) theory. This theory of Pavlovian conditioning deviated from its instrumental predecessors in two important ways. Firstly, the language used to describe the learning changed. Rather than talking in terms of increments in the probability of a response, this theory, instead, discussed learning in terms of associative strength—the strength of associations between events that allow a subject to make predictions about future rewards. Secondly, while the Rescorla-Wagner (1972) model used the same linear operator as that employed by Bush and Mosteller (1951), the update on any one trial according to the Rescorla-Wagner (1972) model was equal to the discrepancy between the maximal associative strength supported by the reward and the associative strength that has already accrued toward all present stimuli. Thus, all present stimuli essentially compete for associative strength with the reward they predict. This allowed the model to explain reports, described by Kamin (1969) and others, in which stimuli conditioned in compound shared the

associative strength derived from reward. All stimuli that are currently present can contribute to the upcoming prediction for reward. Therefore, an error in prediction will only occur if the upcoming reward is not predicted by the summed expectation for reward provided by all present stimuli.

While this model did not elaborate on the specific nature of representations of the events being associated, focusing instead on the conditions for learning (e.g., *when* learning will take place rather than *what* is learned), it had important consequences for how associative theorists have since conceptualized learning. Specifically, the emphasis on changing the strength of associations changed the conversation in associative learning theory from one about strength of the response to one about the underlying associative framework underlying the likelihood of a response. This tradition has continued with the elaboration of empirically derived accounts, designed to explain more complex behaviors, which share in spirit the idea that errors in prediction drive changes in our understanding of the causal structure of our environment (Balleine & Dickinson, 1991; Colwill & Rescorla, 1985; Holland & Rescorla, 1975; Miller & Matzel, 1988; Rescorla, 1973; Wagner, Spear, & Miller, 1981).

This brings us to the model currently applied to interpret the dopaminergic prediction error, developed by Sutton and Barto (1981, 1987). While this reinforcement learning model arose out of the field of machine learning, it was influenced by the work described above. Specifically, this model attempted to bridge the concepts derived from the associative learning literature described above with that observed in the field of neurobiology. Essentially, Sutton and Barto (1981) wanted to apply the concepts of associative learning to Hebbian synaptic plasticity—the biological principal that if neuron A repeatedly evokes firing of another neuron B, then the firing of neuron A will be more efficacious in inducing activity in neuron B in the future (Hebb, 1949, 2005). Synaptic plasticity was essentially taken as a neurobiological instantiation of associative learning. In the earliest version of this model, Sutton and Barto (1981) argued that a reward acts to strengthen the ability of the stimulus to elicit a response, going back to the Thorndike (1898) tradition. According to this model, the response is the same response usually elicited by the reward, which comes to be controlled by the predictive stimulus as the associative strength between the stimulus and response increases. The rules for how the reward acts to strengthen the weights between the stimulus and response were adopted from the Rescorla–Wagner model (1972), where learning was driven by the discrepancy between the expected reward predicted by all stimuli present and the actual reward received. In essence, the prediction error acted to bias the response to obtain the maximum amount of reward. Critically, while the Sutton and Barto (1981) model borrowed some aspects of the Rescorla-Wagner (1972) model, in the form of the use of an error incorporating a summed prediction, the content of what could be learned narrowed sharply from the budding notions of a cognitive framework of relationships between events. In the place of this idea, the Sutton and Barto (1981) model

substituted a somewhat Thorndikian conception of stimulus—response associations governing learning and behavior.

Later versions of the Sutton and Barto (1981) model moved even further away from the concept of associative strength between stimuli and responses and instead argued that a reward-predictive cue actually acquires the value inherent in the reward, through a process whereby the value inherent in the reward backpropagated to the antecedent cue (Sutton & Barto, 1987, 1998). Referred to as temporal difference reinforcement learning (TDRL), this iteration of the model conceptualized a stimulus (or, more generally, a "state") as segregated into multiple consecutive time steps. Each time step of a stimulus was associated with its own scalar value estimate that allows a subject to track the expectation of future sum of rewards through time. This development allowed the model to make accurate value predictions from stimulus onset despite the delay until reward delivery, as reward value backpropagates to the initial time step. Further, it estimated when the reward will occur as time steps of the stimulus that are closer to reward will acquire greater value. These value estimates can also be used to choose appropriate actions, where an action can be elected on the basis of the associated state value. Thus, TDRL was a time-derivative model of the original Sutton and Barto (1981), where it was argued that the value inherent in the reward propagates back to the state, which preceded reward delivery to subsequently influence a choice between future actions (Sutton & Barto, 1987, 1998). Importantly, the prediction error was divorced from the system that arbitrates between action choices. Rather, it just acted to endow the reward-predictive cue with the value inherent in the reward, which drove selection of the response associated with that stimulus.

THEORETICAL INTERPRETATIONS OF THE DOPAMINE PREDICTION ERROR

As noted above, Schultz, Dayan, and Montague (1997) hypothesized that dopamine signals acted as the prediction error postulated by the model-free reinforcement learning algorithm described by Sutton and Barto (1981, 1987, 1998). Their proposal was prompted by the close correspondence between several specific features of the model and the pattern of firing in these neurons. Most importantly, the observation that the dopamine response to reward transferred back to onset of the reward-predictive cue seemed to fulfill critical predictions of the Sutton and Barto model (1987), in which the value inherent in reward transferred back to the cue, which predicts its occurrence. Thus, phasic dopamine activity was taken to be the value signal hypothesized by the Sutton and Barto model (1981, 1987), which drives backpropagation of reward value to the stimulus (Schultz et al., 1997). Effectively phasic dopaminergic activity was argued to provide the learning signal that allows the state preceding reward delivery to acquire the value approaching the sum of future rewards, subsequently serving to bias an

organism toward a choice of an action, which would result in the procurement of maximal reward.

This interpretation of phasic dopamine activity as a cached-value error signal has been remarkably influential across the decades since its proposal. The general idea has permeated how the field of neuroscience views not only dopamine function but also the functions of a host of brain regions that interact with the midbrain and are known to be involved in associative learning. This pervasiveness is despite our growing understanding of the complexity of learning. We now know unequivocally that organisms respond to stimuli in the environment as a consequence of relationships between events either in addition or in spite of any sort of static or cached value that may have accrued toward them in prior training (Balleine & Dickinson, 1991, 1998; Colwill & Rescorla, 1985; Delamater, 1996; Dickinson & Balleine, 1994; Holland & Rescorla, 1975; Killcross & Coutureau, 2003; Rescorla, 1973; Rescorla & Wagner, 1972). Indeed, humans and other animals are capable of developing rich models of the associative relationships between events in the world that are used in the absence of direct experience to influence ongoing and future behavior (Brogden, 1939; Colwill & Rescorla, 1985; Tolman, 1948). Yet the interpretation that the dopamine prediction error functions only to endow reward-predictive cues with a scalar value precludes the involvement of dopamine in the development of these more complex models of the environment. So, could dopamine also be involved in more complex forms of associative learning that transcend cached value or is it restricted to the model-free reinforcement learning algorithm described by Sutton and Barto (1981, 1987, 1998)?

The hypothesis that phasic dopamine acts only as a cached-value prediction error (Sutton & Barto, 1981, 1998) makes three notable predictions about when changes in phasic dopaminergic activity should be seen and what sorts of learning this phasic activity can support. Firstly, this theory predicts that stimulation or inhibition of dopamine neurons should act as a value signal to produce increments and decrements in responding toward reward-paired cues. Secondly, such manipulations of dopamine activity should not produce learning about the relationships between events of the world outside of a scalar expectation of value. Finally, phasic activity of dopamine neurons should not be evident in response to valueless changes in reward or to cues which have come to predict reward indirectly. We will now discuss these predictions in light of several recent studies that we believe provide particularly strong tests of their validity.

PREDICTION ONE: PHASIC STIMULATION OR INHIBITION OF DOPAMINE NEURONS SHOULD SUBSTITUTE AS A CACHED-VALUE PREDICTION ERROR TO DRIVE LEARNING

According to Schultz et al. (1997), an increase in phasic activity of dopamine neurons should serve to increase the value attributed toward a reward-paired cue, whereas a phasic

decrease should reduce the value attributed to a cue. The advent of optogenetics has afforded us the temporal specificity to manipulate putative dopamine neurons in manner that allows us to causally test this hypothesis (Deisseroth, 2011; Deisseroth et al., 2006). Ideally, such experiments should arrange the learning materials so that all that is lacking is contingency—or an error in reward prediction—and then attempt to restore that error by manipulating the dopamine neurons precisely when it would be expected to occur. For example, Steinberg et al. (2013) used an optogenetic approach to mimic a positive reward prediction error during the blocking task first described by Kamin (1969). Specifically, rats were first presented with an auditory stimulus (e.g., A), which predicted reward. Following this training, a novel visual cue (e.g., X) was presented in compound with A and followed by delivery of the same reward. Rats in the control group failed to learn about cue X, presumably because cue X was blocked by prior training with cue A and reward. However, stimulation of dopaminergic neurons in the VTA during reward delivery after presentation of the AX compound unblocked learning about cue X. This was evident in an increase in rats' responses to the food port when cue X was presented alone in extinction. These data are consistent with the value hypothesis. Specifically, the artificial prediction error could be construed as attaching excess value to cue X despite the predictability of the reward, allowing cue X to become associated with the particular response being made at the time to enter the food port during presentation of the conditioned cue AX.

However, this study does not rule out a simple alternative, which is that the dopamine signal is increasing the salience of the preceding cue, which would also be expected to cause learning. If the dopamine signal acts in this manner, then inhibiting it would result in less learning. If, on the other hand, it acts as a cached-value error signal then phasic inhibition of dopamine should cause extinction learning, essentially decreasing the value attributed to a cue (Schultz et al., 1997). To test this question, Chang et al. (2016) optogenetically produced a brief negative error in VTA dopamine neurons during an overexpectation task. Overexpectation usually involves first pairing two cues (e.g., A and X) individually with reward (e.g., three food pellets). Then, these two cues are paired in compound with the same magnitude of reward. Usually, learning to cue X will decrease as the reward is now "overexpected" by the sum of the expectations elicited by cue A and X (e.g., six food pellets). However, in a modified version of this task Chang et al. (2016) delivered the expected reward (e.g., six pellets) during the compound phase, in order to eliminate the normal negative prediction error and prevent extinction. In half of the rats, dopamine activity in the VTA was inhibited during delivery of the final three pellets in the compound stage. Chang et al. (2016) found that brief inhibition of dopamine neurons during pellet delivery in the compound phase of this modified overexpectation task restored the normal extinction learning to cue X. That is, responding to X decreased with introduction of a brief inhibition of dopamine neurons during reward receipt. Again, these results are consistent with the hypothesis that VTA DA acts as a

bidirectional value signal described in the model-free reinforcement learning algorithm postulated by Sutton and Barto (1981, 1987; Schultz et al., 1997). Specifically, that phasic inhibition of dopamine can act to decrease the value attributed to a cue and therefore reduce the response associated with that cue state.

PREDICTION TWO: WHAT IS LEARNED OR STAMPED IN BY THE PHASIC DOPAMINE SIGNAL SHOULD BE RELATED TO GENERIC OR CACHED VALUE

Experiments showing that optogenetic inhibition or stimulation can drive increases or decreases in responding to reward-predictive cues are consistent with the idea that this signal constitutes a scalar value, which increases or decreases the value attributed to a reward-paired cue. However, in both the studies described above (Chang et al., 2016; Steinberg et al., 2013), the learning induced by manipulating the firing of the dopamine neurons could consist of either general value or the formation of a more detailed associations between the cue and reward in the case of unblocking, and the cue and reward omission in the case of extinction. The former would constitute a learning mechanism consistent with that described in the model-free reinforcement algorithm postulated by Sutton and Barto (1981, 1998), where the latter is a more complex association between events that transcends the backpropagation of value to the reward-predictive cue.

To test whether dopamine transients are sufficient for associative learning beyond value, we used sensory preconditioning. Sensory preconditioning normally entails first pairing two neutral cues together in close succession such that an association forms between them (e.g., C→X). The development of this association can be revealed if one of those cues is later paired with reward. Specifically, if cue X is paired with reward, both cue C and X will elicit an expectation for reward when presented individually under extinction conditions. As cue C has never been directly paired with reward, it can only acquire an association with reward via its association with cue X, which allows it to enter into an association with the reward. This is supported by studies that have shown that cue C will not support conditioned reinforcement—rats will press a lever for cue X but not for cue C (Sharpe, Batchelor, & Schoenbaum, 2017). This suggests that cue C does not have any value independent of the food that it predicts, and so rats will not exert effort to obtain presentations of that stimulus alone. Thus, the sensory preconditioning procedure is well suited to an investigation of whether phasic dopamine may also support the development of rich associations between events in a manner that transcends cached value.

Using a modified version of the sensory preconditioning procedure, we aimed to assess whether optogenetic stimulation of dopamine neurons in the VTA could support associative learning beyond the transfer of cached value (Sharpe et al., 2017).

To do this, we first reduced the likelihood that rats would form an association between the two neutral cues C and X by pairing cue A with X (A→X). Subsequently, cues A and C were presented in compound prior to presentation of cue X (AC→X). Cue X was later paired with reward. In controls, we found that learning about cue C was blocked in this design, similarly to the original blocking studies shown with cues and rewards (Kamin, 1969). However, in our experimental group, stimulation of phasic dopamine at the beginning of X following the AC compound (i.e., AC→X trials) unblocked learning about cue C. These rats entered the magazine when cue C was presented in the final probe test as though they expected delivery of the reward. This suggests that triggering the dopamine neurons to fire at the start of X served to facilitate the formation of an association of a relationship between C and X, which allowed cue C to enter into a direct relationship with reward paired with X. This was confirmed by subsequent tests, which revealed that responding to C was sensitive to devaluation of the reward, showing rats responded to C because they desired the particular food reward. We also do not believe our results can be accounted for by salience, since there was no change in the rate of learning about X during conditioning, after the dopamine stimulation, nor was there any increase in learning about A (which would be evident in our design as stronger blocking of cue D; Sharpe et al., 2017). Thus, overall, these data suggest that the dopamine prediction error is capable of supporting the development of more complex associations than that envisioned by model-free reinforcement learning algorithms (Schultz et al., 1997; Sutton & Barto, 1981, 1998).

PREDICTION THREE: PHASIC CHANGES IN DOPAMINE SHOULD NOT REFLECT INFORMATION ABOUT CUE–REWARD RELATIONSHIPS THAT DOES NOT REFLECT DIRECT EXPERIENCE

If dopamine transients signal the prediction error is that contained in the model-free reinforcement algorithm described by Sutton and Barto (1981, 1987), then phasic dopaminergic activity should not reflect associations that have been inferred from prior associative relationships or a change in current state of the environment. This is because the error contained in the model-free reinforcement learning algorithm only receives predictions based on the value that backpropagates from the reward to the cue after the cue and reward have been paired in close succession (Sutton & Barto, 1981, 1998). This cannot happen if no direct association has been experienced. While the findings from Sharpe et al. (2017) suggest that dopamine can support the acquisition of complex associations between events (rewarding or otherwise), this does not require that the content of the information encoded in the prediction error signal itself go beyond errors in cached value. That is, stimulation or inhibition of dopamine could be allowing

other neural structures to form more complex associations about the relationship between events, yet phasic activity in dopaminergic neurons may be ignorant of these associations under normal circumstances, operating only in response to cached-value errors.

Assessing whether the dopamine prediction error has access to information about the relationship between events requires examining how dopamine neurons or dopamine release changes in response to errors that reflect such associative information. There are now a growing number of studies that do this (Aitken, Greenfield, & Wassum, 2016; Bromberg-Martin & Hikosaka, 2009; Nakahara, Itoh, Kawagoe, Takikawa, & Hikosaka, 2004; Papageorgiou, Baudonnat, Cucca, & Walton, 2016; Sadacca et al., 2016; Takahashi et al., 2011). For example, dopamine activity to reward-paired cues changes depending on the physiological state of the subject (Aitken et al., 2016; Papageorgiou et al., 2016). In one study, Papageorgiou et al. (2016) monitored dopamine release using fast scan voltammetry in the nucleus accumbens (NaCC), as rats were performing an instrumental learning task. Here, rats had a choice of pressing one of two levers for one of two rewards (R1 → O1 or R2 → O2). On some of the trials, rats were presented with one lever option (forced trials; R1 or R2) while on others they could make a choice between pressing either one of the two levers (choice trials; R1 and R2). Prior to test sessions, rats were given free access to one of the rewards (e.g., devaluing O1). Subsequently, rats exhibited a preference for the lever associated with the nondevalued reward they had not had access to prior to the session (R2 → O2). Papageorgiou et al. (2016) found that dopamine release to the reward-paired cues (i.e., the insertion of the lever into the behavioral chamber) was modulated by outcome devaluation prior to the rats experiencing the lever producing the now devalued outcome. That is, the dopamine response to lever presentation on forced trials reflected the new value of the devalued reward before it had been experienced with the lever-press response. Further, the dopaminergic response to presentation of the other lever was increased, showing an increased preference for the nondevalued option. This demonstrates that dopamine responses to reward-paired cues can update in response to the current physiological state of the subject without the subject directly experiencing the association between the cue and now devalued reward. These data are at odds with an interpretation of the dopamine signal as the model-free reinforcement learning algorithm described by Sutton and Barto (1981, 1998), since the cue and the devalued reward have never been paired, and so the new value of the reward cannot be attributed to the cue.

The data from Papageorgiou et al. (2016) beg the question of whether the phasic dopamine signal might also reflect information about an entirely new association developed in the absence of experience. In line with this possibility, Sadacca et al. (2016) showed that phasic activity of dopamine neurons can reflect associations between cues and rewards that have been inferred from prior knowledge of associative relationships in the experimental context. Specifically, Sadacca et al. (2016) recorded the activity of

putative dopamine neurons in the VTA during sensory preconditioning. In this study, rats were first presented with two neutral cues in close temporal succession (A → B). Following this training, one of these cues was paired with reward (B → US). During conditioning, putative dopamine neurons exhibited the expected reward prediction error correlates, firing to reward early in conditioning, and transferring this response back to the cue later in learning. After conditioning, in the probe test in which both cues A and B were presented in the absence of reward, putative dopamine neurons continued to exhibit increased firing to B − the cue paired with reward − while also now firing to A, the cue paired with B in the preconditioning phase. Further, dopamine neuron firing to A and B was correlated, suggesting that the information signaled in response to A was the same as what was signaled in response to B. The simplest interpretation of these data is that dopamine neurons in the VTA signal reward prediction errors similarly whether they are based on directly experienced associations or whether they require inference. Again, this is not accommodated by a theory which argues that the dopamine signal reflects value that has backpropagated from the reward to a cue from their pairing (a notion reinforced by data showing a preconditioned cue does not acquire general value during the preconditioning procedure; This is the Sharpe, Batchelor, and Schoenbaum, 2017, eLife paper again. Rather, these data suggest that dopamine neurons may make more general predictions about the nature of upcoming rewards, garnered from associative model of the world and based on past experience.

CONCLUSIONS

In this chapter, we have discussed data that are problematic for the hypothesis that phasic dopamine signals encode a scalar cached-value signal, which allows a state preceding reward to acquire the value inherent in the reward and motivate behavior. Specifically, optogenetic stimulation or inhibition of has been found not only to increase or decrease responding to a cue preceding reward but also to facilitate the acquisition of associations between two neutral stimuli. Further, changes in phasic dopaminergic activity have been observed in response to a change in the physiological state of the subject despite the subject never experiencing pairing of the cue with the outcome that has been devalued in that state. Finally, the phasic dopaminergic response has also been seen in response to cues that have come to predict reward via prior knowledge of associative relationships in the experimental context, without being directly paired with reward. Such data challenge the conception that transient changes in dopamine carry the cached-value prediction error contained in model-free reinforcement learning algorithms (Sutton & Barto, 1981, 1998). Specifically, such evidence is outside the realm of a theory which argues that a cue only acquires a dopaminergic response via the backpropagation of value. Value cannot transfer back to a cue, which has not been paired with something valuable, and a value signal cannot facilitate the acquisition of associations between neutral stimuli.

So where to now? One theory that warrants consideration is that put forward by Nakahara (2014). Nakahara (2014) argues that dopamine prediction errors can be influenced by more than the expectation elicited from the current state. That is, a prediction error does not need to be calculated on the basis of current sensory information. Rather, a prediction error can be calculated on the basis of hidden states derived from prior experience. Further, the dopamine prediction error in this model can also be used to update these internal models of the environment. However, while the calculation of prediction errors can utilize information garnered from internal models of the world to generate a prediction about upcoming rewards, the signal itself still reflects the discrepancy between the value of the reward expected and that received. Thus, while Nakahara (2014) extends what the dopamine prediction error can use to make predictions, the prediction itself is still one of value which similarly serves to update the expected value of future rewards.

An alternative proposal is that dopamine transients reflect errors in event prediction more generally and that they are also involved in supporting learning about future events whether those events are the delivery of a particular reward, presentation of a neutral stimulus, or even absence of some stimuli or some other events. This would constitute a return to thinking about the prediction error in associative theory as driving real-world associations between events, as described in earlier theories of associative learning (Colwill & Rescorla, 1985; Holland & Rescorla, 1975; Miller & Matzel, 1988; Rescorla, 1973; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972; Wagner et al., 1981) but somewhat abandoned by the world of neuroscience with the advent of TDRL (Sutton & Barto, 1981, 1987, 1998). Data from our lab already show that dopamine transients are both sufficient and also necessary for learning associations between neutral cues that inherently have no value (Sharpe et al., 2017). This is one prediction of the account described above. Below, we consider several more that might be examined in the future to support this hypothesis.

FUTURE DIRECTIONS

Conceptualizing the dopamine prediction error as a signal that detects a discrepancy between expected and actual events makes some testable predictions about when phasic activity in dopamine neurons should be observed. While Sadacca et al. (2016) and Papageorgiou et al. (2016) showed that dopamine activity to reward-paired cues can change as a result of knowledge not acquired through direct experience, in each case dopaminergic activity still signaled an upcoming prediction that could be construed as being about reward value. However, the alternative proposal made here suggests that changes in phasic dopaminergic activity would also be seen as a result of other changes in the predicted event that do not constitute a shift in value. For example, an increase in dopaminergic signaling should occur in response to a change in the identity of a reward. That is, if a cue previously paired with a particular reward was unexpectedly presented

with a different reward that was equally valuable, we would expect to see a prediction error in dopaminergic neurons. While a few studies have looked at errors in response to change value of different rewards and have claimed not to see evidence of such a signal, these studies do not examine how dopamine signals change when identity is altered independent of value or reward preference (Lak et al., 2014; Stauffer et al., 2014). Thus, while their results show that the value error is similar across different identity rewards, they do not address whether changes in identity evoke error signals. It is also worth noting that it may be necessary to move beyond single unit correlates to appreciate more subtle error signaling functions of the dopamine system. As in other brain areas, meaningful signals may be carried in the pattern of finding across an ensemble of dopamine neurons, which may not be evident in individual "grandmother" neurons. In this regard, value errors may be a particularly amazing example of a more general population function. In any event, whether in individual units or ensemble responses, such a finding would represent strong evidence that the dopamine prediction error accesses information about the content of what is expected, independent of its meaning with regard to value.

Further, future research may also search for the presence of a dopaminergic error signal when a more general associative relationship between neutral stimuli is violated even in the absence of rewards. It is well-established that dopamine neurons in the midbrain fire when a novel stimulus is first presented unexpectedly, (Schultz, 1998). While this has been interpreted in the literature as a "novelty bonus" (Kakade & Dayan, 2002), it is also possible that this is an error signal in response to the appearance of an unexpected stimulus. It would be valuable to assess in an appropriately controlled environment whether these dopamine signals are seen when the contingency between neutral stimuli is manipulated such that expectation about upcoming stimuli is violated. Such research would support the hypothesis that the dopamine prediction error may reflect a more general signal for detecting the discrepancy between actual and expected events. Experiments like these would be useful since positive findings would open up new possibilities for how this biological signal may support associative learning in these and other contexts.

REFERENCES

Aitken, T. J., Greenfield, V. Y., & Wassum, K. M. (2016). Nucleus accumbens core dopamine signaling tracks the need-based motivational value of food-paired cues. *Journal of Neurochemistry, 136*(5), 1026–1036.

Balleine, B., & Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology, 43*(3), 279–296.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37*(4), 407–419.

Blundell, P., Hall, G., & Killcross, S. (2003). Preserved sensitivity to outcome value after lesions of the basolateral amygdala. *Journal of Neuroscience, 23*(20), 7702–7709.

Brogden, W. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology, 25*(4), 323.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron, 63*(1), 119–126.
Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58*(5), 313.
Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., & Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nature Neuroscience, 19*(1), 111–116.
Colwill, R. M., & Rescorla, R. A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes, 11*(1), 120.
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.
Deisseroth, K. (2011). Optogenetics. *Nature Methods, 8*(1), 26–29.
Deisseroth, K., Feng, G., Majewska, A. K., Miesenböck, G., Ting, A., & Schnitzer, M. J. (2006). Next-generation optical technologies for illuminating genetically targeted brain circuits. *Journal of Neuroscience, 26*.
Delamater, A. R. (1996). Effects of several extinction treatments upon the integrity of Pavlovian stimulus-outcome associations. *Learning & Behavior, 24*(4), 437–449.
Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior, 22*(1), 1–18.
Estes, W. K. (1950). Effects of competing reactions on the conditioning curve for bar pressing. *Journal of Experimental Psychology, 40*(2), 200.
Estes, W. K., & Burke, C. J. (1953). A theory of stimulus variability in learning. *Psychological Review, 60*(4), 276.
Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*(5614), 1898–1902.
Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach*. John Wiley & Sons.
Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press.
Holland, P. C., & Rescorla, R. A. (1975). The effect of two ways of devaluing the unconditioned stimulus after first-and second-order appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 1*(4), 355.
Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1*(4), 304–309.
Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks, 15*(4), 549–559.
Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. *Punishment and Aversive Behavior*, 279–296.
Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex, 13*(4), 400–408.
Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences of the United States of America, 111*(6), 2343–2348.
Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology, 67*(1), 145–163.
Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. *Psychology of Learning and Motivation, 22*, 51–92.
Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology, 72*(2), 1024–1027.
Nakahara, H. (2014). Multiplexing signals in reinforcement learning with internal models and dopamine. *Current Opinion in Neurobiology, 25*, 123–129.
Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41*(2), 269–280.
Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology, 47*(6), 419.
Papageorgiou, G. K., Baudonnat, M., Cucca, F., & Walton, M. E. (2016). Mesolimbic dopamine encodes prediction errors in a state-dependent manner. *Cell Reports, 15*(2), 221–228.

Rescorla, R. A. (1973). Effects of US habituation following conditioning. *Journal of Comparative and Physiological Psychology, 82*(1), 137.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2*, 64–99.
Robbins, T., & Everitt, B. (1982). Functional studies of the central catecholamines. *International Review of Neurobiology, 23*, 303–365.
Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology, 63*(3), 592–606.
Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife, 5*, e13665.
Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *Journal of Neurophysiology, 56*(5), 1439–1461.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*(1), 1–27.
Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience, 13*(3), 900–913.
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.
Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., … Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience, 20*.
Sharpe, M. J., Batchelor, H. M., & Schoenbaum, G. (2017). Preconditioned cues have no value. *Elife, 6*.
Spyraki, C., Fibiger, H. C., & Phillips, A. G. (1982). Dopaminergic substrates of amphetamine-induced place preference conditioning. *Brain Research, 253*(1), 185–193.
Stauffer, W. R., Lak, A., & Schultz, W. (2014). Dopamine reward prediction error responses reflect marginal utility. *Current Biology, 24*(21), 2491–2500.
Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience, 16*(7), 966–973.
Stricker, E. M., & Zigmond, M. J. (1974). Effects on homeostasis of intraventricular injections of 6-hydroxydopamine in rats. *Journal of Comparative and Physiological Psychology, 86*(6), 973.
Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*(2), 135–170.
Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Paper presented at the proceedings of the ninth annual conference of the cognitive science society*.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT Press Cambridge.
Takahashi, Y. K., Roesch, M. R., Wilson, R. C., Toreson, K., O'donnell, P., Niv, Y., & Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature Neuroscience, 14*(12), 1590–1597.
Thorndike, E. L. (1898). Review of animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*.
Tolman, E. C. (1948). *Cognitive maps in rats and men*. American Psychological Association.
Ungerstedt, U. (1971). Adipsia and aphagia after 6-hydroxydopamine induced degeneration of the nigro-striatal dopamine system. *Acta Physiologica Scandinavica, 82*(S367), 95–122.
Wagner, A., & Rescorla, R. (1972). Inhibition in Pavlovian conditioning: Application of a theory. *Inhibition and Learning*, 301–336.
Wagner, A. R., Spear, N., & Miller, R. (1981). SOP: A model of automatic memory processing in animal behavior. *Information Processing in Animals: Memory Mechanisms, 85*, 5–47.
Wise, R. A. (1987). The role of reward pathways in the development of drug dependence. *Pharmacology & Therapeutics, 35*(1–2), 227–263.
Wise, R. A., & Rompre, P.-P. (1989). Brain dopamine and reward. *Annual Review of Psychology, 40*(1), 191–225.
Yokel, R. A., & Wise, R. A. (1975). Increased lever pressing for amphetamine after pimozide in rats: Implications for a dopamine theory of reward. *Science, 187*(4176), 547–549.