HHS Public Access

Author manuscript

Curr Opin Neurobiol. Author manuscript; available in PMC 2019 April 01.

Published in final edited form as:

Curr Opin Neurobiol. 2018 April ; 49: 1–7. doi:10.1016/j.conb.2017.10.006.

Model-based predictions for dopamine

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

[1]Princeton Neuroscience Institute & Department of Psychology, Princeton University, Princeton, NJ, 08540

[2]National Institute on Drug Abuse, Baltimore, MD, 21224

[3]School of Psychology, UNSW Australia

Abstract

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

Introduction

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

Are model-free dopamine prediction errors a red herring?

A core tenet of TDRL is that it is 'model-free': learned state values are aggregate, scalar representations of total future expected reward, in some common currency [1,13]. That is,

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

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

Although phasic dopamine signals have predominantly been interpreted as model-free temporal difference prediction errors, a growing number of studies leveraging complex behavioral tasks, alongside novel optogenetic and imaging techniques, are revealing an increasingly detailed picture of dopamine reward prediction errors during learning, and the multiple dimensions of reward prediction on which they are based. Intriguingly, several of these studies have demonstrated a significant degree of heterogeneity in dopaminergic responses during learning, suggesting greater complexity in these signals than previously appreciated. Below we review evidence from these recent studies, asking what is the nature of dopamine signals? Do they reflect an aggregate (scalar) error, or a vector-based signal that includes not only the magnitude of deviation from predictions, but also the identity of the deviation (did I get more food than expected, or water instead of food)? And how might these signals be incorporated into learning algorithms implemented throughout the brain?

Temporal representation and dopamine

One notable property of dopamine prediction errors is that they are temporally precise: if an expected reward is omitted, the phasic decrease in dopamine neuron activity appears just after the time the reward would have occurred [2]. It is this phenomenon that inspired the TDRL algorithm, which models such temporally precise predictions by postulating sequences of time-point states that are triggered by a stimulus (known as the 'complete serial compound,' CSC stimulus representation, or 'tapped delay line'; Figure 1), each of which separately accrues value through experience [6]. However, when a reward is delivered unexpectedly early, dopamine neurons do not display a phasic decrease in activity at the original expected time of reward, as would be implied by the CSC, in which a prediction error updates the value of the current, and not subsequent, timepoint states [18,19]. Reset

mechanisms, in which reward delivery terminates the CSC representation, have been proposed to address this [19], but other challenges suggest that the CSC is perhaps not as viable an explanation for learned timing. Specifically, prediction errors are only slightly enhanced to temporally variable rewards, suggesting that under some conditions reward predictions may have low temporal precision [20], and multiple studies in humans (first inspired by [21]) have shown that a not-fully-predicted reward (or reward omission) affects choice of its related cue on the very next trial, suggesting that the CSC include only a single time-point, which then leaves unexplained how the timing of reward (relative to stimulus onset) is learned.

An alternative is to allow task states to persist for learned durations (formally, a 'semi-Markov' framework), with reward predictions tied to a temporally-evolving belief about the current latent state. Learning values for latent states, rather than cues, incorporates a rich world model, and suggests that prediction error signaling is 'gated' by inference about when one state has transitioned to another [19,22]. Recent work has directly demonstrated that dopamine reward prediction errors are consistent with this framework [23]. Here, when a cue predicted reward delivery with an unknown (but capped) delay, the passage of time since cue onset made reward delivery more likely, eliciting smaller dopamine prediction errors to later rewards. In contrast, when reward delivery was probabilistic, as time passed it became more likely that the trial would not be rewarded, and indeed dopamine responses increased with reward delay. Consistent with this theory, other studies have shown that dopamine activity reflects evolving temporal predictions, suggesting at the very least that inference about the timing of events (for e.g., the hazard rate) influences the computation of dopamine reward prediction errors [20,24–26]. More broadly, optogenetic manipulation of midbrain dopamine activity is sufficient to bidirectionally change judgments on a temporal categorization task [25], directly implicating dopamine signaling in timing processes. It also appears that the generation of prediction errors due to mistimed reward delivery is neurally separable from computing prediction errors due to an unexpected amount of reward, as ventral striatum lesions abolish the former (so a mistimed reward does not elicit a prediction error signal) while leaving prediction errors due to reward magnitude intact. This finding argues against the time-bound representation of value in the CSC representation, suggesting instead a semi-Markov model in which the duration of states and the amount of reward associated with each state are separately learned, and the ventral striatum plays a key role in learning or representing the former, but not necessarily the latter [22].

In general, it is often implicitly assumed that states correspond directly to percepts of cues in the environment [27,28]. However, apart from the challenges that timing poses to such an account, even straightforward neural representations of the environment are an interpretation of the external reality through, at minimum, a relevance filter [29,30]. It is therefore natural to extend TDRL models by allowing expected value to be calculated with respect to inferred states that capture the learned structure of a task [17,31–34]. The mapping between observations (such as cues and rewards) and underlying task states may be probabilistic (as in 'partially observable environments') or ambiguous (for example in the case of conflicting or mixed cues) [19,35–37], making state inference itself a non-trivial process. However, it is important to keep in mind that both model-free and model-based values can be learned/computed for states that do not correspond directly to observable cues—prediction errors

based on inferred states are not, in of themselves, a departure from model-free TDRL, since at the time the errors are generated, they may still be based on cached values attached to the hidden states through direct experience.

Not all dopaminergic predictions are learned through direct experience

Indeed, a central aspect of TDRL that makes it model free is that, in the algorithm, values for state are learned (and cached) through direct experience with the state. Recent work suggests, however, that phasic dopamine may reflect values that have been learned indirectly. Of particular relevance is a sensory preconditioning experiment showing that reward predictions that are ascribed to a cue solely through its relationship to another neutral cue are reflected in dopamine neuron firing. Here, two neutral cues (A and B) were first presented in sequence multiple times (A→B), and then one of the cues, B, was paired with food in a separate training session. Behaviorally, this later training is known to endow cue A with reward-predictive value. Importantly, the authors showed that after B→food training, the presentation of cue A elicited a phasic increase in dopamine, which was correlated with activity elicited by presentation of cue B. This suggests that the expectation that A would lead to reward, presumably computed through model-based forward simulation of A→B and B→food, was available to dopamine neurons [38].

Notably, TDRL has no mechanism by which value can transfer between predictive cues retrospectively. Attempts have been made to explain these results by enhancing TDRL to operate not only on the current state, but on states that are inferred to be related to the current state—a departure from pure model-free reinforcement learning—as in ‘mediated learning,’ [39,40] or the Kalman TD model [32,41]. These explanations suggest that during the pairing of B with food, a neural representation of A is activated by association to B, and therefore also associated with the food. However, if the orbitofrontal cortex—an area associated with model-based computing of values—is inactivated at test, responding to A is abolished, while responding to B is intact [42]. Given that OFC has been repeatedly shown to be unnecessary for conditioned responding to cues directly paired with reward (for example, cue B in this experiment), this result strongly suggests that the value of A is computed in OFC at the time of the test and not during the B→food training. That dopamine prediction errors may reflect this computed-on-the-fly value is also consistent with accumulating evidence from fMRI showing that prediction error signals include model-based information and that model-based decisions are sensitive to striatal dopamine [43–45].

We note that even if model-based values are used to compute prediction errors, the error itself may still influence only model-free learning, for instance of a behavioral policy [46]. Indeed, it is possible that at test A invokes a model-based representation of the inferred B, the cached value of which is available to dopamine neurons. Under this scenario, the prediction error signaled to A arises from the cached value of B not A [47]. It is also important to note that adding inferred states and access to model-based values does not (yet) require that dopamine convey a prediction error signal that is used for learning the model itself. However, optogenetic silencing in a related task shows that dopamine transients are in fact required for the initial formation of associations between cues A and B, even though no

rewards were present, and therefore learning in that phase could not have been driven by scalar prediction errors [48].

Multiple dimensions of prediction in dopamine responses

Another fundamental property of TDRL is that it learns aggregate, scalar predictions of the sum of future rewards predicated on occupying the current state—a 'common currency' value that sums over apples, oranges, sex and sleep. As alluded to above, and complicating the mapping between dopamine and TDRL even further, it appears that dopamine neurons respond to deviations from predictions in dimensions other than scalar value [49]. In particular, prediction errors have been recorded for an unexpected change in the flavor of reward pellets, even though there was no change in their subjective value [50]. Such "state prediction errors," that is, prediction errors due to an unexpected state ("I got chocolate milk rather than vanilla"), suggest that the identity of the outcome is a component of reward prediction in dopamine circuits, at odds with the model-free framework that explicitly ignores specific identities and compares values in common currency. Information about outcome identity may reflect inputs from the orbitofrontal cortex [51] which track multiple specific features of outcomes beyond reward amount [52,53].

Model-based learning with dopamine prediction errors

All told, current findings suggest that dopamine neurons have access to model-based representations of expected rewards that reflect learned properties beyond a scalar representation of value (Figure 1, right). However, the convergence of TDRL to a useful value representation stems from the alignment between the computational goal of the agent (to maximize total reward through value-guided action) and the single dimension along which reward predictions are represented (i.e. scalar value). Unless used judiciously, a generalized prediction error signal [54] that responds to *any* mismatch along multiple dimensions of an outcome (e.g., the color of a reward, or the oddly shaped plate it was served in) might erroneously perturb value representations upon which choices are putatively based, biasing the animal away from the normative goal (for example, towards preferring low-quality food served in ever-changing plates, rather than high-quality food served in more mundane dinnerware). Such biases have indeed been identified in the influence of novelty and information on both dopamine reward prediction errors and value-guided choice [55,56], but it is unclear how widespread they are.

Indeed, to be truly useful for learning a world model, 'model-based prediction errors' must be computed for every aspect of the model in parallel—a multidimensional (i.e., vector) prediction error that signals not only that there is a mismatch between expectation and reality, but exactly what dimension of prediction was misaligned [34,57,58]. Do dopamine neurons signal such model-based prediction errors? If so, ideally, these would be broadcast in parallel so that the correct component of the model might be updated via its respective prediction error [19,22] (Figure 1, right). This would allow a segregation of learning across different dimensions of reward prediction such as value, state identity, or time, supported by separable neural populations. Such segregation might account for the distinct pattern of prediction-error signaling in dopamine terminals across striatal subregions [59,60], and

might be a more prominent feature of dopamine activity than previously detected, in part due to a sampling bias whereby experiments investigating dopamine signaling have almost exclusively manipulated reward value, not other state dimensions.

Moreover, because much of what we know about dopamine activity is derived from the analysis of activity of individual neurons or localized dopamine release or from techniques that average these signals over large populations, we may be missing more complex spatiotemporal and network interactions that can only be uncovered by treating these neurons as ensembles with unique input and output relationships. For example, target regions that receive, and learn from, dopamine prediction error signals might locally separate the incoming signal into distinct components, allowing the relevant dimensions of prediction to be flexibly decoded, depending on the current task and internal goals. For example, cholinergic signaling in the striatum is known to powerfully modulate dopamine release [61,62], implying local circuit control over the influence of dopamine signals according to the current state of the task [63,64]. However, exactly how a truly multiplexed prediction error could be separated into its orthogonal components is not trivial, to say the least.

So what is the role of dopamine in learning?

One thing that these recent studies make clear is that a better understanding of the computational role of dopamine entails a broader consideration of what it means for a reinforcement learning algorithm to be ‘model-based’ [34]. Model-based prediction in RL has been most strongly identified with the use of models for forward planning, enabling values to be computed on the fly (as opposed to cached) in order to flexibly support goal-directed behavior [65]. But models may also be exploited to enable learning over hidden states, for example in algorithms that combine inference with TDRL [36,66]. Indeed, the necessity to represent states through time, either by a CSC or other, more complex state representation [67,68], can be thought of as a model of the past—and now unobservable—state of the environment. Overall, the dopaminergic signatures of model-based prediction we have highlighted draw attention to the question of what is being learned *about*—while a relatively straightforward stimulus representation may be evident to an experimenter, such a representation may not form the basis of learning for a behaving animal in more complex tasks [66].

The suggestion that dopamine signals a multidimensional model-based prediction-error signal departs considerably from the claim (and supporting evidence) that all dopamine neurons broadcast a single, scalar quantity across vast areas of the brain. But, it is hard to see how lumping together all model-based prediction errors into one aggregate signal would be useful for downstream learning, unless we modify what we think the prediction error does downstream. One possibility is that the dopamine prediction-error signal enhances learning in target areas indiscriminately, without signaling the direction of learning—similar to a salience signal, in the service of learning rather than action—and information about what exact prediction was violated is available from other sources. Indeed, sensory and associative areas that have a detailed representation of the current state (including all cue and reward properties deemed relevant to the task) may be in the best position to know exactly in

what ways this state is unexpected. Unfortunately, this re-envisioning of the role of phasic dopamine signals would not explain why some prediction errors, namely those to reward omission, are signaled by pauses in firing. Multiplexing of model-free scalar prediction errors and model-based multidimensional prediction errors may be the answer – but only future experiments directly testing for the existence of several of these errors at once, will tell. In any case, what is becoming clear is that phasic dopamine signals, until recently a beacon of computationally-interpretable brain activity, may not be as simple as we once hoped they were.

Acknowledgments

This work was funded by grant R01DA042065 from the National Institute on Drug Abuse (AJL, YN), grant W911NF-14-1-0101 from the Army Research Office (YN, MJS), an NHMRC CJ Martin fellowship (MJS), and the Intramural Research Program at the National Institute on Drug Abuse (ZIA-DA000587) (MJS, GS). The opinions expressed in this article are the authors' own and do not reflect the view of the NIH/DHHS. The authors have no conflicts of interest to report.

References

1. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience. 1996; 16:1936–1947. [PubMed: 8774460]
2. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]
3. Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience. 2007; 10:1615–1624. [PubMed: 18026098]
4. Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 2015; 525:243–246. [PubMed: 26322583]
5. Niv Y, Schoenbaum G. Dialogues on prediction errors. Trends in Cognitive Sciences. 2008; 12:265–272. [PubMed: 18567531]
6. Sutton, RS., Barto, AG. Reinforcement learning: An introduction. Vol. 1. MIT press; Cambridge: 1998.
7. Reynolds JN, Hyland BI, Wickens JR. A cellular mechanism of reward-related learning. Nature. 2001; 413:67–70. [PubMed: 11544526]
8. Yagishita S, Hayashi-Takagi A, Ellis-Davies GC, Urakubo H, Ishii S, Kasai H. A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science. 2014; 345:1616–1620. [PubMed: 25258080]
9. Hollerman JR, Schultz W. Dopamine neurons report an error in the temporal prediction of reward during learning. Nature neuroscience. 1998; 1:304–309. [PubMed: 10195164]
10. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nature Neuroscience. 2006; 9:1057–1063. [PubMed: 16862149]
11. Tobler PN, Dickinson A, Schultz W. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. Journal of Neuroscience. 2003; 23:10402–10410. [PubMed: 14614099]
12. Pan W-X, Schmidt R, Wickens JR, Hyland BI. Tripartite Mechanism of Extinction Suggested by Dopamine Neuron Activity and Temporal Difference Model. The Journal of Neuroscience. 2008; 28:9619. [PubMed: 18815248]
13. Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005; 47:129–141. [PubMed: 15996553]
14. Kobayashi S, Schultz W. Influence of reward delays on responses of dopamine neurons. Journal of Neuroscience. 2008; 28:7837–7846. [PubMed: 18667616]

15. Lak A, Stauffer WR, Schultz W. Dopamine prediction error responses integrate subjective value from different reward dimensions. Proceedings of the National Academy of Sciences. 2014; 111:2343–2348.

16. Dayan P, Berridge KC. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. Cognitive, Affective, & Behavioral Neuroscience. 2014; 14:473–492.

17. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 2005; 8:1704–1711. [PubMed: 16286932]

18. Dayan P, Niv Y. Reinforcement learning: the good, the bad and the ugly. Current Opinion in Neurobiology. 2008; 18:185–196. [PubMed: 18708140]

19. Daw ND, Courville AC, Touretzky DS. Representation and timing in theories of the dopamine system. Neural Computation. 2006; 18:1637–1677. [PubMed: 16764517]

20. Fiorillo CD, Newsome WT, Schultz W. The temporal precision of reward prediction in dopamine neurons. Nature neuroscience. 2008; 11:966–973. [PubMed: 18660807]

21. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron. 2003; 38:329–337. [PubMed: 12718865]

22. Takahashi Yuji K, Langdon Angela J, Niv Y, Schoenbaum G. Temporal Specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. Neuron. 2016; 91:182–193. [PubMed: 27292535]

23**. Starkweather CK, Babayan BM, Uchida N, Gershman SJ. Dopamine reward prediction errors reflect hidden-state inference across time. Nature Neuroscience. 2017; 20:581–589. The authors elegantly demonstrate that the profile of dopamine prediction error signals to reward delivered following a predictive cue after a variable delay depends critically on the learned structure of a task, in this case, whether a reward follows the cue on 100% or 90% of trials. They show how the distinct patterns of prediction errors cannot be accounted for by model-free TDRL with a CSC stimulus representation, but instead are consistent with a TDRL algorithm that learns over a belief regarding the current state of the task. [PubMed: 28263301]

24. Nomoto K, Schultz W, Watanabe T, Sakagami M. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. Journal of neuroscience. 2010; 30:10692–10702. [PubMed: 20702700]

25**. Soares S, Atallah BV, Paton JJ. Midbrain dopamine neurons control judgment of time. Science. 2016; 354:1273–1277. Using fiber photometry in mice, the authors demonstrate that dopamine responses to a cue marking the end of a variable interval during a temporal discrimination task signal a prediction error that reflects both the average expected reward and an evolving belief about the likely timing of the cue (i.e. the hazard function). Optogenetic activation or inhibition of dopamine activity was sufficient to systematically bias discrimination performance, consistent with a role for dopamine in the judgement of time. [PubMed: 27940870]

26. Pasquereau B, Turner RS. Dopamine neurons encode errors in predicting movement trigger occurrence. Journal of Neurophysiology. 2015; 113:1110. [PubMed: 25411459]

27. Nakahara H. Multiplexing signals in reinforcement learning with internal models and dopamine. Current Opinion in Neurobiology. 2014; 25:123–129. [PubMed: 24463329]

28. Nakahara H, Hikosaka O. Learning to represent reward structure: A key to adapting to complex environments. Neuroscience Research. 2012; 74:177–183. [PubMed: 23069349]

29. Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, Wilson RC. Reinforcement learning in multidimensional environments relies on attention mechanisms. The Journal of Neuroscience. 2015; 35:8145–8157. [PubMed: 26019331]

30. Leong YC, Radulescu A, Daniel R, DeWoskin V, Niv Y. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron. 2017; 93:451–463. [PubMed: 28103483]

31. Daw ND, Dayan P. The algorithmic anatomy of model-based evaluation. Philosophical Transactions of the Royal Society B. Biological Sciences. 2014:369.

32. Gershman SJ. A unifying probabilistic view of associative learning. PLoS Computational Biology. 2015; 11:e1004567. [PubMed: 26535896]

33. Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O. Dopamine neurons can represent context-dependent prediction error. Neuron. 2004; 41:269–280. [PubMed: 14741107]
34. Samejima K, Doya K. Multiple Representations of Belief States and Action Values in Corticobasal Ganglia Loops. Annals of the New York Academy of Sciences. 2007; 1104:213–228. [PubMed: 17435124]
35. Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artificial Intelligence. 1998; 101:99–134.
36. Rao RP. Decision making under uncertainty: a neural model based on partially observable markov decision processes. Frontiers in Computational Neuroscience. 2010; 4:146. [PubMed: 21152255]
37. Gershman SJ, Niv Y. Learning latent structure: carving nature at its joints. Current Opinion in Neurobiology. 2010; 20:251–256. [PubMed: 20227271]
38. Sadacca B, Jones JL, Schoenbaum G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. eLife. 2016; 5:e13665. [PubMed: 26949249]
39. Holland PC. Acquisition of representation-mediated conditioned food aversions. Learning and Motivation. 1981; 12:1–18.
40. Wimmer GE, Shohamy D. Preference by association: how memory mechanisms in the hippocampus bias decisions. Science. 2012; 338:270–273. [PubMed: 23066083]
41. Gershman SJ. Dopamine, Inference, and Uncertainty. bioRxiv. 2017
42. Jones JL, Esber GR, McDannald MA, Gruber AJ, Hernandez A, Mirenzi A, Schoenbaum G. Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values. Science. 2012; 338:953. [PubMed: 23162000]
43*. Doll BB, Simon DA, Daw ND. The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology. 2012; 22:1075–1081. The authors provide a comprehensive review of the neural and behavioral signatures of model-based reinforcement learning in humans, focusing primarily on findings from fMRI studies. They highlight the puzzling finding that brain regions typically associated with cached-value learning also display signatures of model-based computation. [PubMed: 22959354]
44**. Sharp ME, Foerde K, Daw ND, Shohamy D. Dopamine selectively remediates 'model-based' reward learning: a computational approach. Brain. 2015; 139:355–364. In a novel task combining sensory preconditioning and blocking, the authors demonstrate that the optogenetic activation of VTA dopamine neurons during the pairing of neutral cues is sufficient for the formation of model-based associations between these cues. Further, the authors show that suppression of dopamine neuron activity during the pairing of the neutral cues in the preconditioning phase reduces responding based on the cue-cue relationship at test, demonstrating that dopamine transients are also necessary for the acquisition of model-based associations between cues. [PubMed: 26685155]
45. Deserno L, Huys QJM, Boehme R, Buchert R, Heinze H-J, Grace AA, Dolan RJ, Heinz A, Schlagenhauf F. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proceedings of the National Academy of Sciences. 2015; 112:1595–1600.
46. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69:1204–1215. [PubMed: 21435563]
47. Doll BB, Daw ND. Prediction Error: The expanding role of dopamine. eLife. 2016; 5:e15963. [PubMed: 27099987]
48. Sharpe MJ, Chang CY, Liu MA, Batchelor HM, Mueller LE, Jones JL, Niv Y, Schoenbaum G. Dopamine transients are sufficient and necessary for acquisition of model-based associations. Nature Neuroscience. 2017
49. Bromberg-Martin ES, Matsumoto M, Hikosaka O. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron. 2010; 68:815–834. [PubMed: 21144997]
50. Takahashi YK, Batchelor HM, Liu B, Khanna A, Morales M, Schoenbaum G. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. 2017
51. Takahashi YK, Roesch MR, Wilson RC, Toreson K, O'Donnell P, Niv Y, Schoenbaum G. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nature Neuroscience. 2011; 14:1590–1597. [PubMed: 22037501]

52. Lopatina N, Sadacca BF, McDannald MA, Styer CV, Peterson JF, Cheer JF, Schoenbaum G. Ensembles in medial and lateral orbitofrontal cortex construct cognitive maps emphasizing different features of the behavioral landscape. Behavioral Neuroscience. 2017; 131:201–212. [PubMed: 28541078]

53. Rudebeck PH, Murray EA. The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. Neuron. 2014; 84:1143–1156. [PubMed: 25521376]

54. Hiroyuki N. Multiplexing signals in reinforcement learning with internal models and dopamine. Current opinion in neurobiology. 2014; 25:123–129. [PubMed: 24463329]

55. Bromberg-Martin ES, Hikosaka O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron. 2009; 63:119–126. [PubMed: 19607797]

56. Madan CR, Ludvig EA, Spetch ML. Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. Psychonomic Bulletin & Review. 2014; 21:629–636. [PubMed: 24189991]

57. Doya K, Samejima K, Katagiri K-i, Kawato M. Multiple Model-Based Reinforcement Learning. Neural Computation. 2002; 14:1347–1369. [PubMed: 12020450]

58. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010; 66:585–595. [PubMed: 20510862]

59*. Menegas W, Babayan BM, Uchida N, Watabe-Uchida M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. eLife. 2017; 6:e21886. The authors record axonal dopamine activity in four different regions of the striatum using fiber photometry during classical conditioning, and demonstrate that the pattern of dopamine transients to novel stimuli depends on axonal target. In particular, while dopamine responses in ventral striatum show initial excitation only to rewards and acquire a response to the cue through learning, dopamine responses in the tail of the striatum show strong initial excitation to novel cues that attenuates with learning. These results suggest an anatomical separation between dopamine neurons that signal classic reward prediction errors and those that signal novelty, or errors in sensory prediction. [PubMed: 28054919]

60*. Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, Davidson TJ, Daw ND, Witten IB. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nature Neuroscience. 2016; 19:845–854. The authors record from VTA dopamine axon terminals in the ventral striatum (VS) and dorsomedial striatum (DMS) of mice performing an instrumental reversal-learning task, and find distinct patterns of activity in the two striatal subregions. While both populations reflected a reward prediction error signal, VS dopamine terminals preferentially respond to rewards and reward-predicting cues, and DMS dopamine terminals preferentially respond to contralateral choice, in line with the behavioral specificity of the target subregion. [PubMed: 27110917]

61. Cragg SJ. Meaningful silences: how dopamine listens to the ACh pause. Trends in Neurosciences. 2006; 29:125–131. [PubMed: 16443285]

62. Threlfell S, Lalic T, Platt NJ, Jennings KA, Deisseroth K, Cragg SJ. Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron. 2012; 75:58–64. [PubMed: 22794260]

63. Bradfield LA, Bertran-Gonzalez J, Chieng B, Balleine BW. The thalamostriatal pathway and cholinergic control of goal-directed action: interlacing new with existing learning in the striatum. Neuron. 2013; 79:153–166. [PubMed: 23770257]

64. Stalnaker TA, Berg B, Aujla N, Schoenbaum G. Cholinergic interneurons use orbitofrontal input to track beliefs about current state. Journal of Neuroscience. 2016; 36:6242–6257. [PubMed: 27277802]

65. Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998; 37:407–419. [PubMed: 9704982]

66. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Current Biology. 2017; 27:821–832. [PubMed: 28285994]

67. Gershman SJ, Moustafa AA, Ludvig EA. Time representation in reinforcement learning models of the basal ganglia. Frontiers in Computational Neuroscience. 2014:7. [PubMed: 24478689]

68. Ludvig EA, Sutton RS, Kehoe EJ. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural computation. 2008; 20:3034–3054. [PubMed: 18624657]

Highlights

- Recent work shows that dopamine reward prediction error signals reflect model-based information.
- These model-based predictions rely on complex internal representations of multiple dimensions of the expected outcome, including reward identity, delay, variability.
- We review recent work establishing the role of dopamine in model-based learning, with a focus on computational implications for how dopamine signals influence learning in the brain.

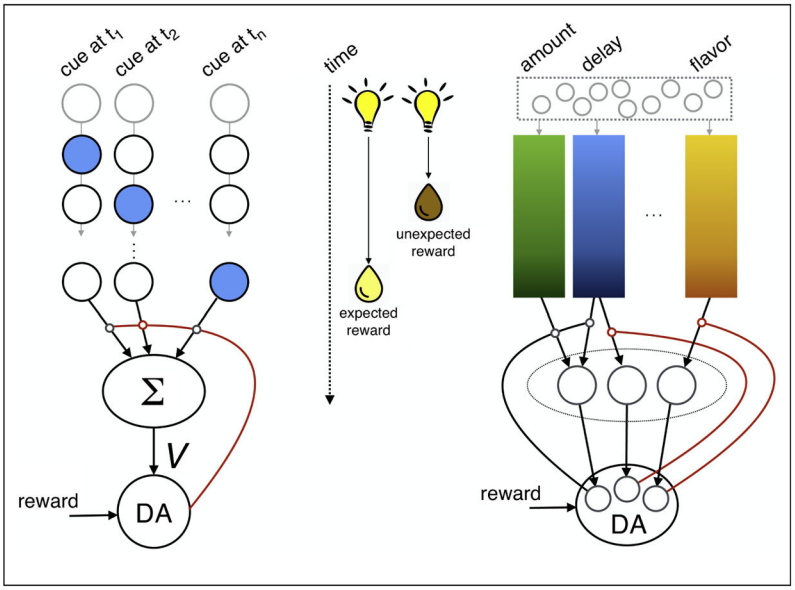

Figure 1.

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