Published in final edited form as:
Eur J Neurosci. 2012 April ; 35(7): 991–996. doi:10.1111/j.1460-9568.2011.07982.x.

Model-based learning and the contribution of the orbitofrontal cortex to the model-free world

Michael A. McDannald[1], **Yuji K. Takahashi**[2], **Nina Lopatina**[3], **Brad W. Pietras**[3], **Josh L. Jones**[1], and **Geoffrey Schoenbaum**[1,2]

[1]University of Maryland School of Medicine, 20 Penn St HSF-2, S251 Baltimore, MD, USA

[2]NIDA-IRP, Baltimore, MD, USA

[3]University of Maryland Program in Neuroscience, Baltimore, MD, USA

Abstract

Learning is proposed to occur when there is a discrepancy between reward prediction and reward receipt. At least two separate systems are thought to exist: one in which predictions are proposed to be based on model-free or cached values; and another in which predictions are model-based. A basic neural circuit for model-free reinforcement learning has already been described. In the model-free circuit the ventral striatum (VS) is thought to supply a common-currency reward prediction to midbrain dopamine neurons that compute prediction errors and drive learning. In a model-based system, predictions can include more information about an expected reward, such as its sensory attributes or current, unique value. This detailed prediction allows for both behavioral flexibility and learning driven by changes in sensory features of rewards alone. Recent evidence from animal learning and human imaging suggests that, in addition to model-free information, the VS also signals model-based information. Further, there is evidence that the orbitofrontal cortex (OFC) signals model-based information. Here we review these data and suggest that the OFC provides model-based information to this traditional model-free circuitry and offer possibilities as to how this interaction might occur.

Keywords

model-based; model-free; orbitofrontal cortex; Pavlovian; striatum

Model-free learning

A basic observation from studies of learning is that animals will adjust their behavior to reflect contingencies between events in their environment (Rescorla, 1988). Pavlovian conditioning captures this basic observation. In a standard conditioning experiment a naïve, hungry animal is placed in an experimental chamber in which a neutral cue predicts a palatable food reward. As a result of this predictive relationship, the cue will acquire many properties, including the ability to elicit anticipatory reward behavior. Many theories of learning have been proposed to explain the behavioral change that results from such predictive relationships. One such theory is temporal difference reinforcement learning (TDRL).

Correspondence: Dr G. Schoenbaum, [1]University of Maryland School of Medicine, 20 Penn St HSF-2, S251 Baltimore, MD, USA, schoenbg@schoenbaumlab.org.

In TDRL (Sutton, 1988; Dayan & Sejnowski, 1994), a modern cousin of the Rescorla–Wagner model (Rescorla & Wagner, 1972), reward prediction errors drive cue-reward learning. TDRL is composed of three main components: actor, critic and temporal difference (TD) module. The actor comprises a mapping between environmental cues and actions/behavior. The critic comprises a mapping of environmental cues and their predicted values. The value prediction made by the critic is model-free in the sense that it is made in a common currency, devoid of referents to the specific form and features of the reward that is predicted. Thus, a $20 bill and $20 worth of doughnuts would generate identical value predictions by the model-free critic. The model-free critic feeds this value prediction to the TD module, which compares this information to the ‘actual’ reward value received. A prediction error is then computed. When the model-free critic predicts a value of 0 (which would be the case for a neutral cue) and a palatable food reward is subsequently received (which intrinsically has a value of, e.g. 1), the TD module generates a large, positive prediction error to the reward. This prediction error signal is sent to the actor, strengthening the mapping between that cue and actions/behavior; and to the model-free critic, increasing the reward value predicted. As the model-free critic more accurately predicts reward value, the TD module computes a smaller difference between the predicted and actual reward, resulting in a smaller prediction error. Eventually the reward is so well predicted that little or no error is generated on its receipt, resulting in little or no learning. When the model-free critic predicts a value of 1 (the palatable food reward is completely predicted) and no food is received (a value of 0), however, the TD module generates a large negative prediction error to the reward. A negative prediction error has the opposite effect on the actor, weakening the mapping between cue and actions/behavior, and also on the critic, decreasing the reward value predicted.

The different components of a model-free learning mechanism are thought to reside in distinct brain areas. For example, human imaging has found blood oxygen level-dependent (BOLD) signaling in the dorsolateral striatum (DS) consistent with that of the actor (O’Doherty *et al.*, 2004). BOLD signal and primate electrophysiology have found activity in the ventral striatum (VS), consistent with that of a critic (Schultz *et al.*, 1992; O’Doherty *et al.*, 2004). Human imaging (D’Ardenne *et al.*, 2008), primate electrophysiology (Schultz *et al.*, 1997; Fiorillo *et al.*, 2003, 2008) and rodent electrophysiology (Roesch *et al.*, 2007) have all found neural activity consistent with that of TD error signaling in midbrain structures containing dopamine (DA) neurons. Further, fast-scan cyclic voltammetry in rodents, which can measure subsecond DA release, has uncovered DA release patterns in the VS consistent with those predicted by a TD error signaling (Day *et al.*, 2007). The proposed interaction of these brain areas follows exactly as above. When an environmental cue is presented, the VS sends information about the predicted reward value to the midbrain. Midbrain DA neurons compare the predicted value with that actually received, and generate a prediction error. A positive prediction error results in synchronous phasic DA release in both the DS and VS (Day *et al.*, 2007; Roitman *et al.*, 2008) to modulate learning, either directly or due to co-release of other agents (Lapish *et al.*, 2007; Stuber *et al.*, 2010). These results provide a framework for understanding the neural circuits that support model-free learning, which we will return to later.

Limitations of model-free representations

Model-free representations are of obvious benefit. They ensure value is attributed to, and behavior driven by, cues predicting reward. Despite this, model-free representations have significant limitations. Recall that model-free value is represented in a common currency. This means that the predictions used to guide behavior are blind to specific features of rewards. Viewed in this light, model-free representations share some, albeit not all, features with habitual response mechanisms that select actions without concern for the specific reward to be obtained (Dickinson & Balleine, 1994). Indeed, this is why value is described as being ‘cached’ in the cue or action representation. Yet in many or even most circumstances, behavior is guided by a specific representation of the expected reward from which its current, unique value is derived. To make this more concrete, imagine you are on your way to work. You can choose between making it there on time or stopping off to get a doughnut. Clearly this decision would be usefully informed by knowing specific information about the different options: What time is it? Did you pack a lunch? Is your boss out of town? And, most importantly, are the Krispy Kreme doughnuts fresh? Model-free representations do not readily allow decisions to be influenced by such specific information about predicted consequences.

A second limitation of model-free representations is that they do not allow for ‘new learning’ to occur when these specific features of the predicted rewards are changed, so long as general or cached ‘value’ is maintained. Again this reflects the fact that the predictions contain no information about the specific features of the upcoming rewards. Because no specific information concerning the reward is predicted, differences between predicted and actual sensory features of rewards cannot be detected. As a result, no learning can occur. Imagine your employer began to pay you in doughnuts instead of dollars. While this may seem absurd, if you only had a model-free learning system, you would fail to learn from this change, so long as the amount of doughnuts were of exactly equal value to the money expected. Of course this is not what would happen – you would notice the change and learn about the new compensation program! Similarly animals will learn from shifts in the identity of an expected outcome, even when that outcome value is more or less unchanged. Again model-free representations cannot account for such behavior. These observations suggest that, overall, model-free representations alone are insufficient to account for even fairly simple behaviors. Instead we must be able to make predictions about specific features of rewards to account for behavior that is flexible and for learning to occur when there are changes in specific features of predicted rewards. This ability is a core feature of model-based representations.

Model-based representations

Using model-based representations, humans and animals form a sort of ‘cognitive map’ of their environment (Tolman, 1948). This map contains information about how specific environmental events relate to one another. When an animal is confronted with one event, it can use this map to look forward and predict specific features of the upcoming event, such as its timing, probability, sensory aspects and any unique value it may have. Note that is in contrast to a model-free representation that would only predict the common-currency value of an upcoming event. Predicting specific features of upcoming events is a defining feature of a model-based representation. The ability of model-based task representations to predict specific features of upcoming events allows for greater behavioral flexibility. In the aforementioned examples, this would allow you to use relevant information, such as the type of doughnuts or your bosses’ travel plans in deciding whether to stop off on your way to work. Experimentally, this ability is well-captured by reinforcer devaluation (Holland & Straub, 1979). In reinforcer devaluation, a hungry animal learns that a cue predicts a palatable reward, such as sugar. Once the cue–sugar relationship is well-learned, the value of sugar is reduced, either by pairing its consumption with nausea or selective satiation through unlimited sugar consumption. The crucial test comes when the cue is later presented without being reinforced. When this is done there is a spontaneous decrease in cue responding. This decrease could not be observed if animals were only using a model-free representation. To decrease cue responding using a model-free system, the devalued food would have to be encountered and consumed, in order to generate a negative prediction error and modify the cue–behavior mapping of the actor. This would require a number of trials –

all the while the animal would be consuming a devalued food. The spontaneous decrease observed is consistent with the availability and use of a model-based representation. Drawing from their cue–sugar mapping, animals look forward from the cue and predict delivery of the sugar reward. This information, combined with knowledge of sugar's current value, allows cue responding to be spontaneously decreased without need for consuming or even encountering the sugar reward.

Similarly, model-based representations also allow learning to be sensitive to changes in the specific features of the expected reward, for example when your employer began to pay you in doughnuts. Expecting money but receiving doughnuts would produce a model-based prediction error signal; no matter if the dollar value of the doughnuts exactly matched your salary. This error signal would in turn drive new learning. This is because money and doughnuts are quite different, physical events. In this way a model-based prediction error signal allows for learning even in situations in which predicted value and actual value are identical. Experimentally, this learning can be isolated using Pavlovian unblocking procedures (Holland, 1984; Rescorla, 1999). In unblocking, animals are first trained that different cues predict a different reward identity or value. Once these associations are established new cues are added and the reward identity or value selectively changed. In a final test responding to the added cues signaling a selective change in reward identity or reward value is assessed. Normal animals show evidence of learning to the cues signaling a change in either identity or value. Learning in value unblocking may be purely driven by a common currency, model-free representation. A small value of reward is predicted, yet a large reward is received. It should be noted that a model-based representation could also support value unblocking, to the extent that a larger reward is a different physical event. Thus, value unblocking, by itself, may not fully distinguish model-based from model-free learning mechanisms. By contrast, learning in identity unblocking requires access to a model-based representation – because the two identities to be learned about are of equivalent value. Predictions of specific reward identity and the resulting model-based prediction errors are sufficient to drive new learning.

At this point we hope to have made clear that both model-free and model-based representations are necessary in order for flexible, goal-directed behavior to be implemented, and for learning to be driven by changes in specific properties of rewards. Next we will discuss emerging evidence that the VS, the critic in the model-free TDRL system, may also function as the model-based critic. We then discuss evidence suggesting that model-based information from the orbito-frontal cortex (OFC) influences processing in the model-free TDRL circuit described above. We end by suggesting possible neural circuits by which model-based information from the OFC might influence information processing in the VS, providing a common circuit for the integration of model-free and model-based systems.

VS, model-free and model-based representations

The neural mechanisms supporting model-based representations are only beginning to be understood. Comparing predictions of model-based and model-free computational models to behavioral and functional magnetic resonance imaging BOLD data for subjects performing similar sequential Markov decision tasks, recent studies have found differential localization of prediction errors (Gläscher *et al.*, 2010; Daw *et al.*, 2011). Gläscher *et al.* (2010) found dissociation, with a model-based prediction error signal in the lateral prefrontal cortex and a model-free prediction error signal in the VS. The model-based prediction error was observed following a transition into an unexpected state, defined by the task structure, and occurred regardless of whether the state was rewarded unexpectedly. Interestingly, Daw *et al.* (2011) demonstrated an integration of neural signatures for model-based and model-free prediction errors in both the medial prefrontal cortex and VS. Notably the Daw *et al.* (2011) paradigm required subjects to constantly update knowledge of the optimal state–action pair and state transitions, thus both model-free and model-based learning were required to make optimal choices. By contrast, in the Gläscher *et al.* (2010) paradigm, subjects were primarily using model-based predictions to guide behavior. This critical difference may explain the dissociation of the two types of signals in the former but not the latter study. Indeed, Daw *et al.* (2011) found that the extent to which ventral striatal BOLD signals associated with model-based computation was correlated with the extent to which subjects' choice behavior was model-based.

These recent human imaging studies implicate the VS in the use of both model-free and model-based representation. A review of the animal behavior literature supports a role for the VS in both as well. The VS contributes to a host of behaviors that to varying degrees are consistent with a role in using model-free representations, including second-order conditioning (Setlow *et al.*, 2002), conditioned place preference (Everitt *et al.*, 1991), general-affective conditioned reinforcement (Ito *et al.*, 2004), and general-affective Pavlovian to instrumental transfer (Corbit *et al.*, 2001; Hall *et al.*, 2001; de Borchgrave *et al.*, 2002; Corbit & Balleine, 2011; Saddoris *et al.*, 2011). In all of these behaviors animals need not represent the sensory aspects of the predicted reward – a representation of value in common currency would suffice. For example, in second-order conditioning (Holland & Rescorla, 1975) animals are first trained that a primary cue predicts reward. Once this is well-learned, a secondary cue now predicts the occurrence of the primary cue. When this is done the secondary cue is found to control considerable reward behavior. Notably, second-order responding is not sensitive to the current value of the reward, a hallmark of model-free representations. This suggests that the secondary cue does not signal any specific features of reward but instead may be signaling a common currency value. The critical involvement of the VS in second-order conditioning strongly suggests it normally contributes to the use of model-free representations.

In line with the expanded role for the VS, findings from animal behavior also implicate the VS in the use of model-based representations. Behaviors reflecting model-based representations would be those in which specific information about upcoming rewards is necessary to guide behavior. Reinforcer devaluation, which is affected by VS lesions, falls into this category (Lex & Hauber, 2010; Singh *et al.*, 2010). Pavlovian to instrumental transfer based on the specific features of rewards would also require a model-based representation. In this task, three cues are first trained to predict three different kinds of reward. Next, two different responses are trained to produce two of these rewards. In a final test each of the three cues are presented while animals are responding on one of the two responses in extinction. Specific transfer is found when the cue facilitates responding for the action leading to the same reward, relative to the response leading to the different reward. The specificity of transfer could only be observed if animals had formed specific predictions to both the cue and response in initial learning. This effect is dependent on the shell subregion of the VS (Corbit *et al.*, 2001; Corbit & Balleine, 2011). Thus, there is evidence the VS contributes to behaviors potentially tapping into both model-free and model-based representations.

Critical contributions of the VS to model-free and model-based learning

Recently we have also examined the role of the VS in learning driven by model-free and model-based representations (McDannald *et al.*, 2011). This work utilized unblocking procedures (Holland, 1984; Rescorla, 1999) to demonstrate that violations of either predicted reward identity (model-based) or predicted reward value (model-free) are sufficient to drive new learning. As described earlier, three cues were first trained to predict different quantities and flavors of reward. Once these reward predictions had been well-formed, new cues were added in compound with each of the originally trained cues and properties of the reward were changed. In the identity condition the quantity of reward was held constant but the flavor was changed, selectively violating the expected reward identity. Because the two flavors were equally preferred, and thus any value difference was minimal or non-existent, it is difficult for a model-free mechanism to support learning in this situation. In the value condition the flavor of the reward was held constant but the quantity was increased, selectively violating the predicted reward value. Because a prediction of common currency value would suffice, a model-free mechanism would support learning in this situation. These were contrasted with a ‘blocked’ condition in which the flavor and quantity of reward were held constant, no prediction was violated. Normal animals demonstrated learning to the added cues signaling changes in either reward identity or value, suggesting that normal animals use both model-based and model-free learning processes. VS-lesioned animals failed to show either identity or value unblocking, suggesting a failure to employ either model-free or model-based learning. This finding supports the view that the VS is necessary for using model-free and model-based representations to drive new learning.

OFC as model-based critic

In addition to an extensive literature on the VS, there is a great deal known about the function of the OFC. Unlike the evidence implicating the VS, there is little evidence that the OFC is necessary for behaviors requiring only general or cached information about value. For example, the OFC is not necessary for general conditioned reinforcement (Burke *et al.*, 2008), tracking value during discrimination learning (Walton *et al.*, 2010), or cue-potentiated feeding (McDannald *et al.*, 2005). However, OFC function is necessary for a host of behaviors requiring specific information of rewards and their unique value (Delamater, 2007). One such example is differential outcome expectancy (Trapold & Overmier, 1972). In this task animals must discriminate specific cue–response–outcome chains. For example, when one cue is present only a left response is reinforced, but when another cue is present only a right response is reinforced. Normally this discrimination is difficult, but if the rewards produced by the two chains differ, learning proceeds much more rapidly. This facilitation is thought to result from the use of outcome-specific expectancies. Interestingly, OFC lesions abolish this facilitation (McDannald *et al.*, 2005; Ramirez & Savage, 2007). Similarly the OFC has been shown to be necessary for outcome-specific Pavlovian to instrumental transfer (Ostlund & Balleine, 2007), outcome-specific conditioned reinforcement (Burke *et al.*, 2008) and reinforcer devaluation (Gallagher *et al.*, 1999; Pickens *et al.*, 2003, 2005; Izquierdo *et al.*, 2004). These findings are each consistent with a critical role for the OFC in using model-based representations to guide behavior.

However, the role of the OFC is not limited to guiding behavior, it is also critical when this information is necessary to drive learning. This is evident in the performance of OFC-lesioned animals in the same unblocking task above (Burke *et al.*, 2008; McDannald *et al.*, 2011). Lesions had no effect on initial conditioning or compound training. However, lesioned animals were significantly impaired in identity unblocking. When the quantity of reward was held constant but the specific features of the predicted reward changed, learning did not occur. These findings are consistent with the OFC being necessary for the use of model-based representations. Interestingly, OFC-lesioned animals showed normal value unblocking, demonstrating an intact ability to learn when there was a greater than predicted quantity of reward. Thus, while VS was generally important for unblocking, OFC appears to be specifically involved in unblocking that requires model-based information.

Interaction of model-free and model-based learning systems

The results described above suggest that model-based information in the prefrontal cortex, particularly the OFC, can impact model-free learning in the VS and downstream midbrain regions. Such a functional interaction is consistent with anatomical studies that demonstrate strong, unilateral projections from the OFC to the VS (Brog *et al.*, 1993; Haber *et al.*, 1995). Notably, the results above leave open a number of interpretations for how model-based information from the OFC might be integrated both within VS and downstream in the TD module (Fig. 1, all circuits). Distinct populations of neurons within the VS may function as model-free and model-based critics (Fig. 1, circuits 1–3), or information from both sources may be integrated within a single population of neurons (Fig. 1, circuit 4). While model-based information is likely to be supplied to the VS by the OFC, model-free information is likely either a product of the VS itself or is supplied by another brain region. Likewise, model-free and model-based information may be output separately from the VS, with model-free information being sent to dopaminergic midbrain neurons serving as the model-free module and a yet-to-be-identified structure serving as the model-based module (Fig. 1, circuit 1) or, alternatively, DA midbrain neurons may contain separate neuronal populations that serve as the model-based and model-free TD modules (Fig. 1, circuit 2) or may contain only a single TD module, responsible for calculating model-based and model-free prediction errors (Fig. 1, circuits 3 and 4).

While these ideas are speculative, there is currently some evidence supporting a strong integration between these two systems. The finding of BOLD signal consistent with both model-based and model-free prediction errors in the VS suggests that dopaminergic input to this region may reflect both types of information. Further neural correlates with action sequences, inferred values and impending actions evident in recent DA recording studies could also reflect model-based input (Morris *et al.*, 2006; Bromberg-Martin *et al.*, 2010; Xin & Costa, 2010).

Consistent with this, we have recently found that the OFC provides a critical source of information about predicted rewards used by dopaminergic error signaling mechanisms. Neural activity was recorded from putative DA neurons in the ventral tegmental area (VTA) in animals with ipsilateral sham or neurotoxic lesions of the OFC (Takahashi *et al.*, 2011). Recordings were made in a simple odor-guided choice task in which different odor cues indicated that a sucrose reward was available in one of two nearby fluid wells. During recording, we manipulated the timing or size of reward across blocks of trials to induce discrepancies between predicted and actual rewards. Firing in DA neurons in sham animals was greater for an unpredicted reward and declined with learning. After learning, these same neurons also suppressed firing upon omission of a predicted reward. Ipsilateral OFC lesions did not affect animals' performance on the task and also did not change phasic firing of DA neurons to reward. However, DA neurons in OFC-lesioned animals failed to reduce firing to reward with learning, and also failed to suppress firing on reward omission after learning. While this study does not specify the type of prediction error induced (model-free vs. model-based), these results suggest that output from the OFC regarding predicted rewards contributes to error signaling by DA neurons in the VTA.

Conclusion

Here we have reviewed evidence that the VS contributes to both model-free and model-based learning systems, while the OFC appears to selectively contribute to model-based learning. Future studies will provide a more detailed account of the OFC–VS interactions that give rise to model-based learning and how this is integrated with the existing model-free system. Determining whether similar or different VS and DA neural populations process model-free and model-based information will be critical in understanding how these two kinds of information are used to guide behavior and drive learning. Further, model-free and model-based learning mechanisms appear to be differentially affected by drugs of abuse – impairing model-based (Schoenbaum *et al.*, 2004; Schoenbaum & Setlow, 2005) but enhancing model-free representations (Wyvell & Berridge, 2001; Saddoris *et al.*, 2011). Studies of behavioral neuroscience have greatly benefited from theoretical distinctions as acquisition vs. expression (Lazaro-Munoz *et al.*, 2010), appetitive vs. aversive (Balleine & Killcross, 2006), motivational vs. cognitive (Holland & Gallagher, 2004), etc. We feel as though the distinction between model-free and model-based representations is equally important. Describing the neural circuits that give rise to model-free and model-based learning will further our understanding of how these systems support adaptive behavior/learning, and will facilitate the development of new therapies to ameliorate their dysfunction in addiction and other disorders.

Acknowledgments

This study was supported by R01DA015718 to G.S.

Abbreviations

BOLD	blood oxygen level-dependent
DA	dopamine
OFC	orbitofrontal cortex
TD	temporal difference
TDRL	temporal difference reinforcement learning
VS	ventral striatum
VTA	ventral tegmental area

References

Balleine BW, Killcross S. Parallel incentive processing: an integrated view of amygdala function. Trends Neurosci. 2006; 29:272–279. [PubMed: 16545468]

de Borchgrave R, Rawlins JN, Dickinson A, Balleine BW. Effects of cytotoxic nucleus accumbens lesions on instrumental conditioning in rats. Exp. Brain Res. 2002; 144:50–68. [PubMed: 11976759]

Brog JS, Salyapongse A, Deutch AY, Zahm DS. The patterns of afferent innervation of the core and shell in the "accumbens" part of the rat ventral striatum: immunohistochemical detection of retrogradely transported fluoro-gold. J. Comp. Neurol. 1993; 338:255–278. [PubMed: 8308171]

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. J. Neurophysiol. 2010; 104:1068–1076. [PubMed: 20538770]

Burke KA, Franz TM, Miller DN, Schoenbaum G. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. Nature. 2008; 454:340–344. [PubMed: 18563088]

Corbit LH, Balleine BW. The general and outcome-specific forms of pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. J. Neurosci. 2011; 31:11786–11794. [PubMed: 21849539]

Corbit LH, Muir JL, Balleine BW. The role of the nucleus accumbens in instrumental conditioning: evidence of a functional dissociation between accumbens core and shell. J. Neurosci. 2001; 21:3251–3260. [PubMed: 11312310]

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science. 2008; 319:1264–1267. [PubMed: 18309087]

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]

Day JJ, Roitman MF, Wightman RM, Carelli RM. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat. Neurosci. 2007; 10:1020–1028. [PubMed: 17603481]

Dayan P, Sejnowski TJ. Td(Lambda) converges with probability-1. Mach. Learn. 1994; 14:295–301.

Delamater AR. The role of the orbitofrontal cortex in sensory-specific encoding of associations in pavlovian and instrumental conditioning. Ann. NY Acad. Sci. 2007; 1121:152–173. [PubMed: 17872387]

Dickinson A, Balleine BW. Motivational control of goal-directed action. Anim. Learn. Behav. 1994; 22:1–18.

Everitt BJ, Morris KA, O'Brien A, Robbins TW. The basolateral amygdala-ventral striatal system and conditioned place preference: further evidence of limbic-striatal interactions underlying reward-related processes. Neuroscience. 1991; 42:1–18. [PubMed: 1830641]

Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science. 2003; 299:1898–1902. [PubMed: 12649484]

Fiorillo CD, Newsome WT, Schultz W. The temporal precision of reward prediction in dopamine neurons. Nat. Neurosci. 2008; 11:966–973.

Gallagher M, McMahan RW, Schoenbaum G. Orbitofrontal cortex and representation of incentive value in associative learning. J. Neurosci. 1999; 19:6610–6614. [PubMed: 10414988]

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]

Haber SN, Kunishio K, Mizobuchi M, Lynd-Balta E. The orbital and medial prefrontal circuit through the primate basal ganglia. J. Neurosci. 1995; 15:4851–4867. [PubMed: 7623116]

Hall J, Parkinson JA, Connor TM, Dickinson A, Everitt BJ. Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour. Eur. J. Neurosci. 2001; 13:1984–1992. [PubMed: 11403692]

Holland PC. Unblocking in Pavlovian appetitive conditioning. J. Exp. Psychol. Anim. Behav. Process. 1984; 10:476–497. [PubMed: 6491608]

Holland PC, Gallagher M. Amygdala-frontal interactions and reward expectancy. Curr. Opin. Neurobiol. 2004; 14:148–155. [PubMed: 15082318]

Holland PC, Rescorla RA. The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning. J. Exp. Psychol. Anim. Behav. Process. 1975; 1:355–363. [PubMed: 1202141]

Holland PC, Straub JJ. Differential effects of two ways of devaluing the unconditioned stimulus after Pavlovian appetitive conditioning. J. Exp. Psychol. Anim. Behav. Process. 1979; 5:65–78. [PubMed: 528879]

Ito R, Robbins TW, Everitt BJ. Differential control over cocaine-seeking behavior by nucleus accumbens core and shell. Nat. Neurosci. 2004; 7:389–397. [PubMed: 15034590]

Izquierdo A, Suda RK, Murray EA. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J. Neurosci. 2004; 24:7540–7548. [PubMed: 15329401]

Lapish CC, Kroener S, Durstewitz D, Lavin A, Seamans JK. The ability of the mesocortical dopamine system to operate in distinct temporal modes. Psychopharmacology. 2007; 191:609–625. [PubMed: 17086392]

Lazaro-Munoz G, LeDoux JE, Cain CK. Sidman instrumental avoidance initially depends on lateral and basal amygdala and is constrained by central amygdala-mediated Pavlovian processes. Biol. Psychiatry. 2010; 67:1120–1127. [PubMed: 20110085]

Lex B, Hauber W. The role of nucleus accumbens dopamine in outcome encoding in instrumental and Pavlovian conditioning. Neurobiol. Learn. Mem. 2010; 93:283–290. [PubMed: 19931626]

McDannald MA, Saddoris MP, Gallagher M, Holland PC. Lesions of orbitofrontal cortex impair rats' differential outcome expectancy learning but not conditioned stimulus-potentiated feeding. J. Neurosci. 2005; 25:4626–4632. [PubMed: 15872110]

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J. Neurosci. 2011; 31:2700–2705. [PubMed: 21325538]

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. 2006; 9:1057–1063. [PubMed: 16862149]

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science. 2004; 304:452–454. [PubMed: 15087550]

Ostlund SB, Balleine BW. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. J. Neurosci. 2007; 27:4819–4825. [PubMed: 17475789]

Pickens CL, Saddoris MP, Setlow B, Gallagher M, Holland PC, Schoenbaum G. Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. J. Neurosci. 2003; 23:11078–11084. [PubMed: 14657165]

Pickens CL, Saddoris MP, Gallagher M, Holland PC. Orbitofrontal lesions impair use of cue-outcome associations in a devaluation task. Behav. Neurosci. 2005; 119:317–322. [PubMed: 15727536]

Ramirez DR, Savage LM. Differential involvement of the basolateral amygdala, orbitofrontal cortex, and nucleus accumbens core in the acquisition and use of reward expectancies. Behav. Neurosci. 2007; 121:896–906. [PubMed: 17907822]

Rescorla RA. Pavlovian conditioning. It's not what you think it is. Am. Psychol. 1988; 43:151–160. [PubMed: 3364852]

Rescorla RA. Learning about qualitatively different outcomes during a blocking procedure. Anim. Learn. Behav. 1999; 27:140–151.

Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH.; Prokasy, WF., editors. Classical Conditioning II: Current Research and Theory. New York: Appleton Century Crofts; 1972. p. 64-99.

Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. Neurosci. 2007; 10:1615–1624. [PubMed: 18026098]

Roitman MF, Wheeler RA, Wightman RM, Carelli RM. Real-time chemical responses in the nucleus accumbens differentiate rewarding and aversive stimuli. Nat. Neurosci. 2008; 11:1376–1377. [PubMed: 18978779]

Saddoris MP, Stamatakis A, Carelli RM. Neural correlates of Pavlovian-to-instrumental transfer in the nucleus accumbens shell are selectively potentiated following cocaine self-administration. Eur. J. Neurosci. 2011; 33:2274–2287. [PubMed: 21507084]

Schoenbaum G, Setlow B. Cocaine makes actions insensitive to outcomes but not extinction: implications for altered orbitofrontal-amygdalar function. Cereb. Cortex. 2005; 15:1162–1169. [PubMed: 15563719]

Schoenbaum G, Saddoris MP, Ramus SJ, Shaham Y, Setlow B. Cocaine-experienced rats exhibit learning deficits in a task sensitive to orbitofrontal cortex lesions. Eur. J. Neurosci. 2004; 19:1997–2002. [PubMed: 15078575]

Schultz W, Apicella P, Scarnati E, Ljungberg T. Neuronal activity in monkey ventral striatum related to the expectation of reward. J. Neurosci. 1992; 12:4595–4610. [PubMed: 1464759]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]

Setlow B, Holland PC, Gallagher M. Disconnection of the basolateral amygdala complex and nucleus accumbens impairs appetitive pavlovian second-order conditioned responses. Behav. Neurosci. 2002; 116:267–275. [PubMed: 11996312]

Singh T, McDannald MA, Haney RZ, Cerri DH, Schoenbaum G. Nucleus accumbens core and shell are necessary for reinforcer devaluation effects on pavlovian conditioned responding. Front. Integr. Neurosci. 2010; 4:126. [PubMed: 21088698]

Stuber GD, Hnasko TS, Britt JP, Edwards RH, Bonci A. Dopaminergic terminals in the nucleus accumbens but not the dorsal striatum corelease glutamate. J. Neurosci. 2010; 30:8229–8233. [PubMed: 20554874]

Sutton RS. Learning to predict by the method of temporal difference. Mach. Learn. 1988; 3:9–44.

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. Nat. Neurosci. 2011; 14:1590–1597. [PubMed: 22037501]

Tolman E. Cognitive maps in rats and men. Psychol. Rev. 1948; 55:189–208. [PubMed: 18870876]

Trapold, MA.; Overmier, JB. The second learning process in isntrumental training. In: Black, A.; Prokasy, WF., editors. Classical Conditioning II. New York: Appleton-Century-Crofts; 1972. p. 427-452.

Walton ME, Behrens TEJ, Buckley MJ, Rudebeck PH, Rushworth MFS. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron. 2010; 65:927–939. [PubMed: 20346766]

Wyvell CL, Berridge KC. Incentive sensitization by previous amphetamine exposure: increased cue-triggered "wanting" for sucrose reward. J. Neurosci. 2001; 21:7831–7840. [PubMed: 11567074]

Xin J, Costa RM. Start / stop signals emerge in nigrostriatal circuits during sequence learning. Nature. 2010; 466:457–462. [PubMed: 20651684]

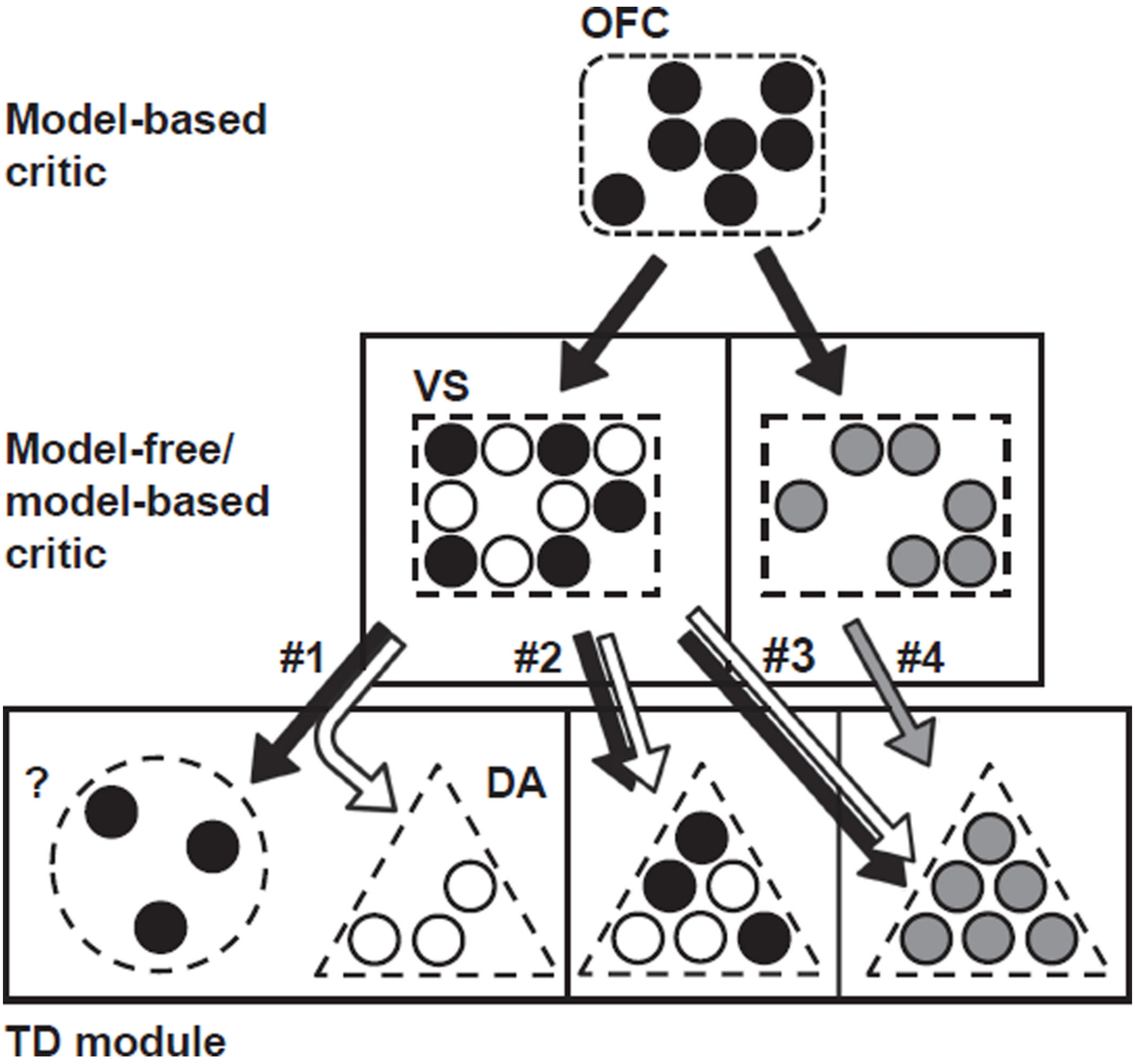

Fig. 1.

Possible neural circuits for model-free and model-based learning. Each row corresponds to a different component of the model/brain area of the circuit: model-based critic; model-based/ model-free critic; and temporal difference (TD) module. The final input of each of the four proposed neural circuits to the TD module is labeled 1–4. The orbitofrontal cortex (OFC) is represented by a dashed, rounded rectangle; the ventral striatum (VS) by a dashed, squared rectangle; midbrain dopamine neurons (DA) by a dashed, triangle, and a yet-to-be-identified brain area by a dashed, circle. The color of the circles within each of these brain areas represents the kind of information processed: black – model-based; white – model-free; and gray – model-free/model-based. The arrows represent the kind of information sent, respecting the same color distinctions as above. For the sake of simplicity, arrows indicating information sent from the TD module to the critics have been omitted.