frontiers in
NEUROSCIENCE

FOCUSED REVIEW
published: 15 May 2010
doi: 10.3389/neuro.01.006.2010

Expectancies in decision making, reinforcement learning, and ventral striatum

Matthijs A. A. van der Meer* and A. David Redish*

Department of Neuroscience, University of Minnesota, Minneapolis, MN, USA

Decisions can arise in different ways, such as from a gut feeling, doing what worked last time, or planful deliberation. Different decision-making systems are dissociable behaviorally, map onto distinct brain systems, and have different computational demands. For instance, "model-free" decision strategies use prediction errors to estimate scalar action values from previous experience, while "model-based" strategies leverage internal forward models to generate and evaluate potentially rich outcome expectancies. Animal learning studies indicate that expectancies may arise from different sources, including not only forward models but also Pavlovian associations, and the flexibility with which such representations impact behavior may depend on how they are generated. In the light of these considerations, we review the results of van der Meer and Redish (2009a), who found that ventral striatal neurons that respond to reward delivery can also be activated at other points, notably at a decision point where hippocampal forward representations were also observed. These data suggest the possibility that ventral striatal reward representations contribute to model-based expectancies used in deliberative decision making.

Keywords: reward, actor–critic, reinforcement learning, Pavlovian-instrumental transfer, planning

Edited by:
Rui M. Costa,
Instituto Gulbenkian de Ciência,
Portugal

Reviewed by:
Henry H. Yin, Duke University, USA
Shih-Chieh Lin, Duke University
Medical Center, USA

****Correspondence:***

Matthijs A. A. van der Meer
is a Post-doctoral Research Fellow in the Department of Neuroscience at the University of Minnesota. He received his B.Sc. from University College Utrecht, followed by a M.Sc. in Informatics at the University of Edinburgh in 2002. His doctoral thesis work, at the University of Edinburgh's Neuroinformatics Doctoral Training Centre with Drs. Mark van Rossum, Emma Wood, and Paul Dudchenko, was on experimental and computational investigations of head direction cells in the rat. After receiving his Ph.D. in 2007, he joined the Redish lab to study the neural basis of planning at the level of neural ensembles.
mvdm@umn.edu

INTRODUCTION: ANATOMY OF A DECISION

Definitions from different approaches to decision making commonly emphasize that a decision should involve "choice among alternatives" (Glimcher et al., 2008). This rules out the extreme case of a (hypothetical) pure reflex where a given stimulus is always followed by a fixed response, and is more in line with "…the delay, between stimulation and response, that seems so characteristic of thought" (Hebb, 1949). A genuine decision depends on more than external circumstances alone: the chosen response or action can reflect the agent's experience, motivation, goals, and perception of the situation. Thus, theories of decision making, by definition, are concerned with covert processes in the brain; with the representations and computations internal to the decision-maker that give rise to behaviorally observable choice.

A useful simplification in studies of economic decision making has been to focus on "static" decision making (Edwards, 1954), where internal variables are assumed fixed and the decision-maker's response to a variety of different choice menus is observed (for instance, would you rather have one apple or five grapes?). This tradition gave us the concept of value or utility, a common currency that allows comparison of the relative merits of different choices (Bernoulli, 1738; Rangel et al., 2008). In experimental studies of animal learning, the complementary "dynamic" approach has been popular, in which the stimulus or situation is held constant and changes in choice behavior resulting from internal variables, such as learning and motivation, can be studied (Domjan, 1998).

The **reinforcement learning (RL)** framework integrates both of these traditions to form

Reinforcement learning (RL)
A computational framework in which agents learn what actions to take based on reinforcement given by the environment. Provides tools to deal with problems, such as reinforcement being delayed with respect to the actions that lead to it (credit assignment problem) or how to balance taking known good actions with unknown ones that might be better (exploration–exploitation tradeoff).

Actor–critic architecture
A class of RL algorithm with two distinct but interacting components. The "actor" decides what actions to take, and the "critic" evaluates how well each action turned out by computing a prediction error. Several studies report a mapping of these components onto distinct structures in the brain.

State space
In order to learn what action to take in a given situation, an agent must be able to detect what situation or state it is in. In RL, the set of all states is known as the state space, which may include location within an environment or the presence of a discriminative stimulus.

Expectancy
A representation of a particular future event or outcome, such as that of food following a predictive (Pavlovian) stimulus or an outcome generated by a **forward model**.

***Correspondence:**

***A. David Redish** is Associate Professor in the Department of Neuroscience at the University of Minnesota. He received his undergraduate degree in writing and computer science from Johns Hopkins in 1991 and his Ph.D. in Computer Science from Carnegie Mellon University in 1997, where he was a student member of the Center for the Neural Basis of Cognition, under the supervision of Dr. David Touretzky. He worked as a Post-doctoral Research Fellow with Drs. Bruce McNaughton and Carol Barnes at the University of Arizona from 1997 to 2000. He has been at the University of Minnesota since 2000, where his lab studies decision-making, particularly issues of covert cognition in rats and failures of decision-making systems in humans.*
redish@umn.edu

an explicit computational account of not only how an agent might choose among alternatives based on a set of internal variables, but also how those variables are learned and modified from experience. The RL framework covers a range of models and methods, but most share common elements exemplified by the basic temporal-difference (TD) algorithm (Sutton and Barto, 1998). Briefly, TD-RL algorithms, such as the **actor–critic** variant, operate on a set of distinct situations or states (such as being in a particular location, or the presentation of a tone stimulus; this set is known as the **state space**), in which one or more actions are available (such as "go left"). Actions can change the state the agent is in and may lead to rewards, conceptualized as scalars in a common reward currency; the agent has to learn from experience which actions lead to the most reward. It does this by updating the expected value of actions based on how much better or worse than expected those actions turn out to be: that is, it relies on a TD prediction error. A single static decision consists of the actor choosing an action based on the learned or "cached" values of the available actions (perhaps it picks the one with the highest value). From the observed outcome, the critic computes a prediction error by comparing the expected value with the value of the new state plus any rewards received. If the prediction error is non-zero, the critic updates its own state value, and the actor's action value is updated in parallel. Thus, by learning a value function over states, the critic allows the actor to learn action values that maximize reward.

In the dynamic (learning) sense, such TD-RL algorithms are very flexible in that they can learn solutions to a variety of complex tasks. However, a key limitation is their dependence on cached action values to make a decision, which means there is no information available about the consequences of actions. This limitation renders decisions inflexible with respect to changing goals and motivations (Dayan, 2002; Daw et al., 2005; Niv et al., 2006). Furthermore, because such cached action values are based only on actual rewards received in the past, they cannot support latent learning, are not available in novel situations, and are only reliable if the world does not change too rapidly relative to the speed of learning. The first limitation is illustrated, for instance, by experiments that involve a motivational shift (Krieckhaus and Wolf, 1968; Dickinson and Balleine, 1994). In an illustrative setup (Dickinson and Dawson, 1987), there is a training phase where action A (left lever) leads to water reward, and action B (right lever) to food reward, calibrated such that both actions are chosen approximately equally. From experience, the agent learns action values for action A and B. Next, the agent is made thirsty and returned to the testing chamber where actions A and B are available but do not lead to reward. All the agent has to go on is its previously learned, cached values for A and B, thus expressing no preference between them[1]. However, what can be observed experimentally is that animals now prefer the left lever (which previously led to water) indicating that they can adjust their choice depending on motivational state (Dickinson and Dawson, 1987)[2]. In contrast, the model is limited by its previously learned values that do not take the motivational shift into account[3]. Furthermore, there are other experimental results which are difficult to explain if decisions are based on cached values that do not include sensory properties of the outcome, such as the differential outcomes effect (Urcuioli, 2005), "causal reasoning" (Blaisdell et al., 2006), shortcut behavior (Tolman, 1948) and specific Pavlovian-instrumental transfer (discussed in detail below).

Such considerations motivated the notion that animals have knowledge about the consequences of their actions, and that they can use such knowledge, or **expectancies**, to make informed decisions (Tolman, 1932; Bolles, 1972; Balleine and Dickinson, 1998). An expectancy can be loosely defined as a representation of an outcome before it occurs; as we discuss in the final section, they may be generated in different ways including action–outcome as well as stimulus–stimulus (Pavlovian) associations. In the context of a motivational shift, an expectancy-based decision mechanism is thought to require two components: genera-

[1]An alternate scenario is that the motivational shift causes the agent to be in a new state. However, in this case, it will not have any cached values at all, so again no preference would be predicted.

[2]For clarity, we have ignored the important but complex issue of under precisely what conditions animals respond immediately, as opposed to only after further experience, to motivational shifts and reinforcer revaluation procedures (see, e.g., Dickinson and Balleine 1994 for details). For now, we merely wish to point out that, under some conditions, they do.

[3]One might imagine a variety of subtle modifications that would enable an actor–critic model to choose appropriately following motivational shifts. For instance, an agent who actually experiences both hungry and thirsty states during training could learn separate cached values for each, such that it would be sensitive to motivational shifts by calling up the relevant set of values. While the learning of multiple value functions would work for this specific experimental situation, it seems unlikely to generalize to different implementations of the procedure (such as pairing a specific outcome with illness; Garcia et al. 1970).

Forward model
In the RL domain, a model of the world that allows an agent to make predictions about the outcomes of its actions (forward in time or "lookahead"). For instance, knowing that pressing a certain lever leads to a "water" outcome or being able to plan a detour if the usual route is blocked, require forward models.

Dynamic evaluation lookahead
An evaluation of a future outcome that takes the agent's current motivational state into account. A two-step process that requires prediction, then evaluation, of the outcome, mapping the prediction onto a value usable for decision making.

Model-free versus model-based RL
Model-free RL maintains a set of values for available actions indicating how successful each action was in the past. It has no concept of the actual outcome (such as food or water) of that action. In contrast, model-based RL takes advantage of such world knowledge, such that a choice which leads to water might be preferred when thirsty.

Decoding
Mapping neural activity to what it represents, such as in reconstructing the identity of a stimulus from spike train data or estimating the location of an animal based on the activity of place cells.

tion of action–outcomes, and evaluation of such outcomes which takes current motivational state and goals into account. Put simply, the rat presses the lever because it predicts a food outcome, and it currently wants the food. This approach is sometimes referred to as "**model-based**" because it relies on a **forward model** of the environment to generate outcomes; in principle, this mechanism needs not be restricted to simply predicting the outcome of a lever press, but could include mental simulation or planning over extended and varied state spaces, such as spatial maps or Tower of London puzzles (Newell and Simon, 1972; Shallice, 1982; Gilbert and Wilson, 2007). While a model of the environment is a necessary component of this approach, it is only half of the solution[4] and a dynamic outcome evaluation step is also required. Thus, we will refer to it here as **dynamic evaluation lookahead** to emphasize the importance of the evaluation step; basic TD-RL, which relies on cached values in the absence of a forward model and dynamic evaluation, we term **"model-free"** (Daw et al., 2005).

POTENTIAL NEURAL CORRELATES OF DYNAMIC EVALUATION LOOKAHEAD

The fact that humans and animals respond appropriately to motivational shifts and other tasks thought to require outcome representations implies the presence of a controller such as dynamic evaluation lookahead. However, it appears a model-free controller is also used in some conditions. Which one is in control of behavior can depend on factors such as the amount of training and the reinforcement schedule. For instance, with extended training behavior can become "habitual", or resistant to reinforcer devaluation, which tends to be effective during early learning (Adams and Dickinson 1981; Daw et al. 2005, but see Colwill and Rescorla 1985). In devaluation in lever pressing tasks, as well as in other procedures, behavior that in principle requires only action values appears to depend on the dorsolateral striatum (Packard and McGaugh, 1996; Yin et al., 2004). In contrast, as might be expected from the variety of world knowledge required for model-based methods, model-based control appears to be more domain-specific. For instance, the ability to plan a route to a particular place requires the hippocampus (Morris et al., 1982; Redish, 1999), while sensitivity to devaluation relies on a limbic network that includes the basolateral amygdala, orbitofrontal cortex, and possibly ventral striatum (Corbit et al. 2001; Pickens et al. 2005; Johnson et al. 2009b, but see de Borchgrave et al. 2002). We focus here on recent results aimed at elucidating the neural basis of model-based decision making.

Recall that dynamic evaluation lookahead requires both the generation and evaluation of potential choice outcomes, implying the existence of neural representations spatio-temporally dissociated from current stimuli (Johnson et al., 2009a). Johnson and Redish (2007) recently identified a possible neural correlate of the internal generation of potential choice outcomes. Recording from ensembles of hippocampal neurons, it was found that while the ensemble usually represented locations close to the animal's actual location (as would be expected from "place cells"), during pauses at the final choice point of the Multiple-T task (**Figure 1A**), the **decoded** location could be observed to sweep down one arm of the maze, then the other, before the rat made a decision (**Figure 1B,C**). Further analyses revealed that on average, the decoded representation was more forward of the animal than backward (implying that it is not a general degeneration of the representation into randomness), tended to represent one choice or the other rather than simultaneously, and tended to be more forward early during sessions (when rats were still uncertain about the correct choice) compared to late (when performance was stable). While the precise relationship of such hippocampal "sweeps" to individual actions or decisions is presently unknown, the manner in which they occur (during pauses at the choice point, during early but not late learning) suggests an involvement in decision making. Consistent with a role in dynamic evaluation lookahead, the hippocampus is required for behaviors requiring route planning in rats (Redish, 1999), and is implicated in imagination, self-projection, and constructive memory in humans (Buckner and Carroll, 2007; Hassabis et al., 2007). If hippocampal sweeps are the neural correlate of the generation of possibilities in dynamic evaluation lookahead, where is the evaluation?

Following the dynamic evaluation lookahead model, any behavioral impact of sweeps (generation of possibilities) would depend on an assignment of a value signal (evaluation). The hippocampal formation sends a functional projection to the ventral striatum (Groenewegen et al., 1987; Ito et al., 2008) and hippocampal network activity can modulate ventral striatal firing (Lansink et al., 2009). Thus, van der Meer and

[4]Indeed, a half that has also been used separately: see for instance Dyna-Q (Sutton, 1990) which also uses a transition model to generate action outcomes, but these are simply evaluated based on cached values without taking the agent's motivational state or goals into account.

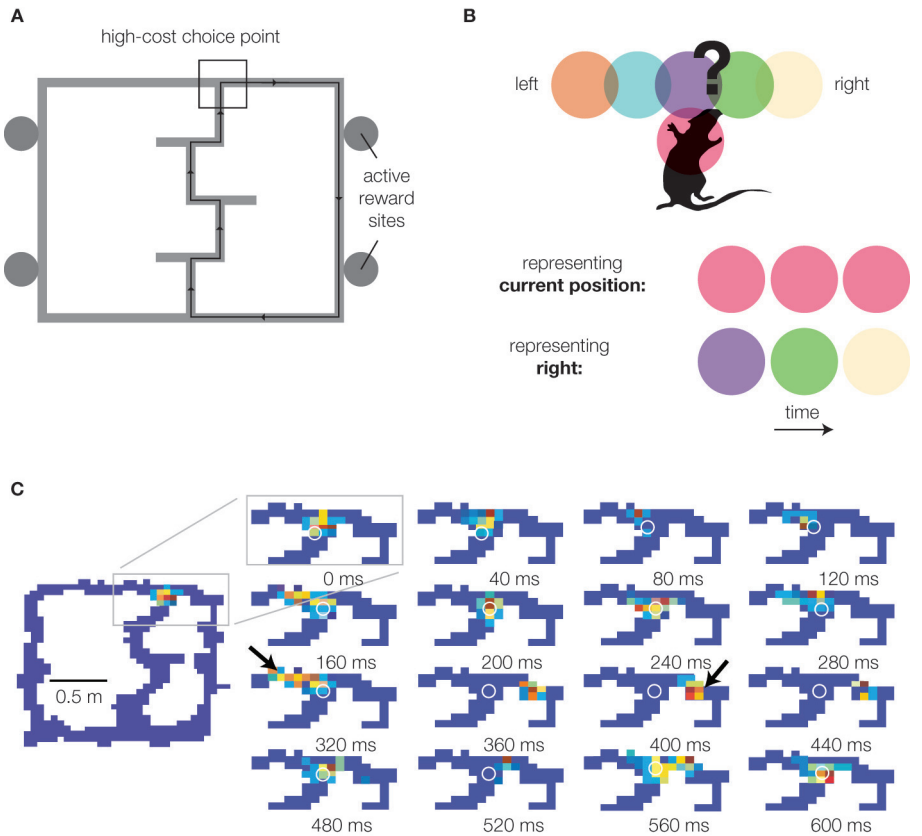

Figure 1 | Representation of forward possibilities at the choice point of the Multiple-T maze. **(A)** The Multiple-T maze. Rats are trained to run laps on an elevated track for food reward. Only one side (right in this example) is rewarded in any given session, but which side is varied between sessions, such that rats start out uncertain about the correct choice. Over the first 10 laps, choice performance increases rapidly, coincident with a tendency to pause at the final choice point (van der Meer and Redish, 2009a). Over the course of a session, rats continue to refine their path, indicating learning beyond choice (Schmitzer-Torbert and Redish, 2002). **(B)** Decoding methods schematic. Neurons in the rat hippocampus tend to be active in specific places on the track: five such "place fields" (colored circles) around the choice point [black box in **(A)**] are shown. By observing which cells are active at any given time, we can infer what location is being represented. If the rat is simply representing its current location, the red cell will be active. In contrast, when the rat pauses at the choice point, activity from the purple, green, and yellow cells might be observed in sequence. This indicates the rat is representing a location distant from its current location (the right maze arm in this example) – a key component of planning. **(C)** Sequence of place representations decoded from actual neurons as the rat (location indicated by the white o) pauses at the final choice point. Red indicates high probability, blue indicates low probability. Note how even though the rat (o) stays stationary, the decoded probability sweeps down the left arm of the maze, then the right (arrows). Data from Johnson and Redish (2007).

Redish (2009a) hypothesized that ventral striatum might play an evaluative role that connects sweeps (possible actions) to behavioral choice (actions). As a first step toward testing this idea, van der Meer and Redish (2009a) recorded from ventral striatal neurons on the same Multiple-T task on which hippocampal sweeps had been observed (Johnson and Redish, 2007). The approach taken was to first isolate cells apparently involved in encoding reward receipt or value (as defined by a significant response to actual reward receipt: food pellets delivered following arrival at the correct maze arm) and then to ask if these neurons were also active at other points on the track. If so, this would indicate potential participation in covert outcome representations. Indeed, the first observation of van der Meer and Redish (2009a) is that ventral striatal neurons, which responded to reward delivery, often fired a small number of spikes at other locations on the track (**Figure 2A**). Based on the Johnson and Redish (2007) finding of sweeps at the choice point, the *a priori* predic-

Covert representation
Neural activity that is not directly attributable to external stimulation or resulting behavior, such as the consideration of possibilities during deliberation or the mental rotation of images.

tion was that if these non-local reward spikes are related to sweeps, they should occur preferentially at the choice point. Although the effect was subtle, this is what was found: compared to non-reward responsive cells, reward cells had a higher firing rate specifically at the choice point (**Figure 2B**). This implies that at the choice point, animals have access to internally generated reward expectancies, which could allow them to modify their actions in the absence of immediate reward.

Next, van der Meer and Redish (2009a) examined the time course of the reward activity at the choice point. Both behavioral evidence and the time course of sweeps suggest a change in strategy on this Multiple-T task (**Figure 1A**), where, initially, behavior is under deliberative, dynamic evaluation lookahead control, but later it is less so. Consistent with this idea, late during sessions, when rats no longer paused at the final choice point, there was no longer any difference between reward and non-reward cell firing at this choice point. It was also found that when the rat deviated from its normal path in an error, representation of reward was increased before turning around. Thus, the **covert representation** of reward effect cannot be easily explained by reward predictive cue–responses, because the effect is specific to choice points, while other places (closer to the reward sites) are more predictive of reward, and because it is present early, but not late, in a constant environment. Instead, this effect suggests ventral striatum may be involved in the evaluation of internally generated possibilities during decision making. We explore this idea in the following section.

VENTRAL STRIATUM AS THE EVALUATOR IN DYNAMIC EVALUATION LOOKAHEAD

Actor–critic models have been especially relevant to neuroscience because of the experimentally observed mapping of its internal variables and processes onto dissociable brain areas. In particular, a common suggestion is that the dorsolateral striatum implements something like the actor, while the ventral tegmental area (VTA) and the ventral striatum work together to implement something like the critic (Houk et al., 1995; O'Doherty et al., 2004). While fMRI studies

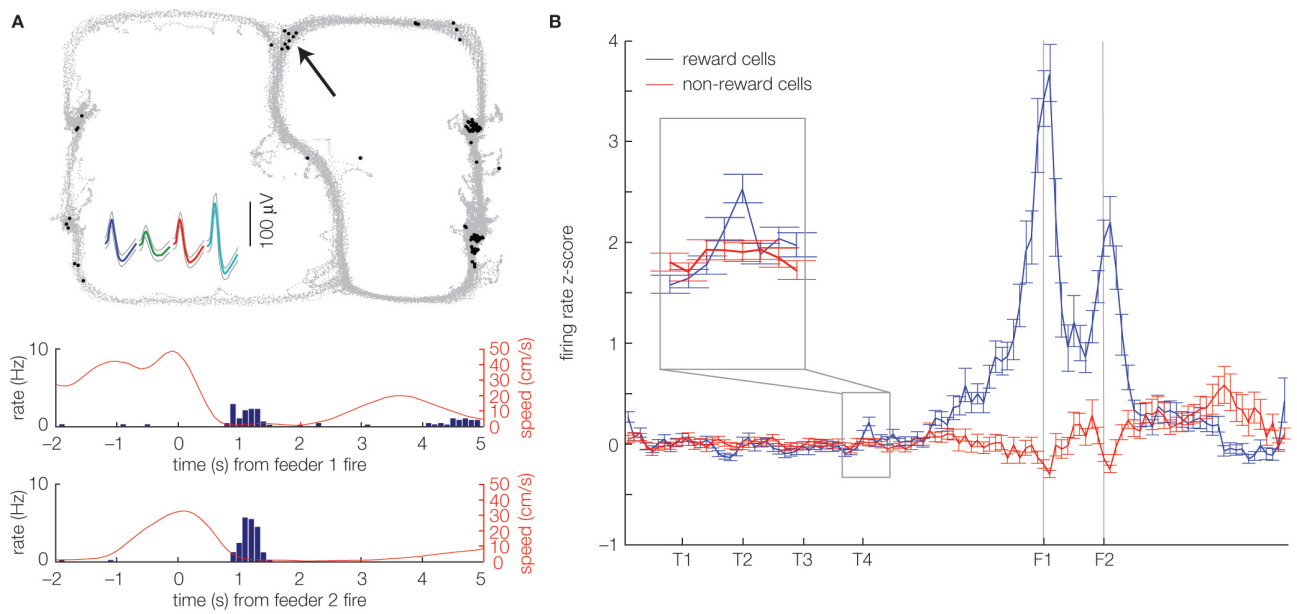

Figure 2 | Ventral striatal neurons show covert expectation of reward at a maze choice point. (A) Example of a reward-responsive neuron in ventral striatum that also fires spikes at other points on the maze, notably at the choice point (black arrow). The top panel shows the rat's path on the track (gray dots), with the black dots indicating the location of the rat when a spike was fired. This neuron responded to reward delivery at the two feeder sites on the right side of the track, as indicated by the transient increase in firing rate at the time of reward delivery (peri-event time histograms, bottom panels). Data taken from a variation on the Multiple-T task published by van der Meer and Redish (2009a). **(B)** Averaged over all cells, reward-responsive (blue), but not non-reward-responsive neurons (red) show a slight but significant increase in firing at the final choice point (T4) during early laps (1–10). Replot of the data in Figure 5, van der Meer and Redish (2009a), obtained by normalizing each cell's firing rate against the distribution of firing rates over the sequence of turns (from S to past T4) in laps 1–10; the original figure published by van der Meer and Redish (2009a) normalized against the firing rate distribution over the same segment of the track, but from all laps. The covert representation of reward effect at T4 was robust against this choice of normalization method.

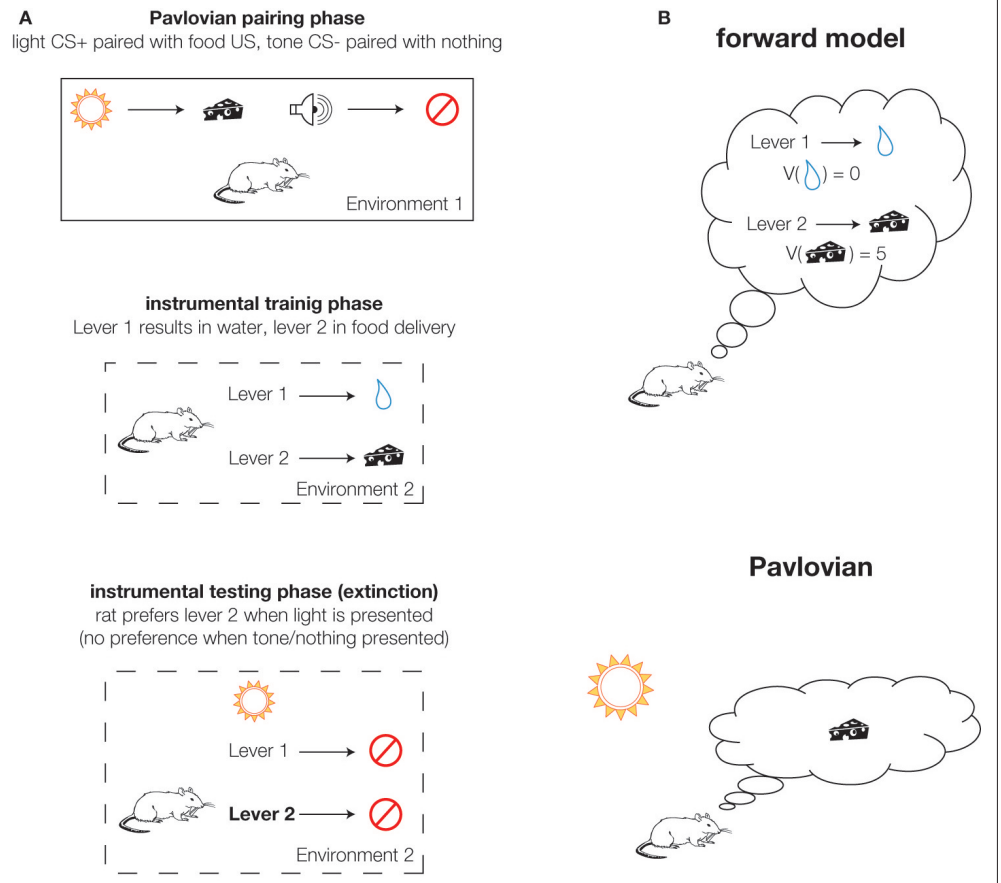

Figure 3 | Expectancies generated by specific Pavlovian-instrumental transfer (PIT) and an internal forward model. (A) Schematic representation of a canonical specific PIT experiment (after Kruse et al. 1983). The animal is exposed to Pavlovian pairing of a light (conditioned stimulus or CS+) preceding food delivery (the unconditioned stimulus or US). Then, in a different environment, the animal learns to press lever 1 to obtain water and lever 2 to obtain food (the same food as in the pairing phase). Ideally these are calibrated such that the animal presses both equally. A specific PIT effect is obtained if, during the critical testing phase, the effect of presenting the light CS+ is to bias the animal's choice towards lever 2 (which previously resulted in food, as predicted by the light). Because this effect is specific to the food, it requires the animal to have an expectancy of the food when pressing lever 2. **(B)** Illustration of the distinction between outcome expectancies generated by an internal forward model (top) and presentation of the light CS+ (bottom). In the forward model case, the animal predicts the outcomes of the different available actions [which are then thought to be available for dynamic evaluation (V)]. In the Pavlovian case, the food outcome is activated by the learned association with the light cue, in the absence of a forward model.

have reliably found value signals in the human ventral striatum (e.g., Preuschoff et al. 2006), the ventral striatum–critic connection has been less frequently made in recording studies (but see Cromwell and Schultz 2003; Takahashi et al. 2008). However, there are reports of ventral striatal firing patterns which are potentially consistent with a critic role. For instance, some ventral striatal neurons respond to actual reward receipt, as well as to cues that predict them (Williams et al., 1993; Setlow et al., 2003; Roitman et al., 2005); this dual encoding of actual and predicted rewards is an important computational requirement of the critic. Also, neurons which ramp up activity at the time or location of reward receipt are commonly found (Schultz et al. 1992; Lavoie and Mizumori 1994; Miyazaki et al. 1998; Khamassi et al. 2008, see **Figure 3** in van der Meer and Redish 2009a), matching what would be expected of a critic state value function.

In the strict actor–critic formulation, the critic only serves to train the actor; it is not required

[5] Non-specific PIT is also observed and refers to a general change in response across available actions (Estes, 1948).

for a single static decision. This is consistent with a rat lesion study that found performance on a well-trained cued choice task was less affected by ventral striatal inactivation during choice, as compared to inactivation during training (Atallah et al., 2007). However, extensive evidence also suggests that ventral striatum is more directly involved in decision making. In particular, as reviewed in Cardinal et al. (2002), ventral striatum is thought to support the behavioral impact of motivationally relevant cues in effects such as autoshaping, conditioned reinforcement, and Pavlovian-instrumental transfer (PIT; Kruse et al. 1983; Colwill and Rescorla 1988; Corbit and Janak 2007; Talmi et al. 2008). For instance, in specific PIT (**Figure 3A**), a Pavlovian association is triggered by the presentation of the conditioned stimulus (CS, e.g., a tone) which has previously only been experienced in a different context than that where the choice is made. This association results in an expectancy containing certain properties of the unconditioned stimulus (US, e.g., food reward) which are sufficient to bias the subject's choice toward actions that result in that US. For instance, given a choice between food and water, presentation of a Pavlovian cue that (in a different context) was paired with food will tend to bias the subject toward choosing food rather than water. Because this effect is reinforcer-specific[5], there must be an expectancy involved that contains outcome-specific properties, as in dynamic evaluation lookahead. However, in specific PIT, this expectancy is not generated by an internal forward model as the outcome of a particular action, but rather by Pavlovian association (**Figure 3B**).

As ventral striatum appears to be required for specific PIT (Corbit et al., 2001; Cardinal et al., 2002), this implies not only that ventral striatum can influence individual decisions, but also that it can do so through an outcome-specific expectancy biasing the subject toward a particular action. Note the similarity between this process and dynamic evaluation lookahead, where an internally generated representation of a particular outcome is involved in choice. Given that ventral striatal afferents, such as the hippocampus, can represent potential outcomes, we propose that ventral striatum evaluates such internally generated expectancies. In the actor–critic algorithm, the critic reports the value of cues or states that "actually occur"; the critic would also be well equipped to report values for "internally generated" cues or states, such as those resulting from model-based lookahead or Pavlovian associations. This is reminiscent of the idea that ventral striatum "mediates the motivational impact of reward-predictive cues" (Cardinal et al., 2002; Schoenbaum and Setlow, 2003) and congruent with an action-biasing role "from motivation to action" (Mogenson et al., 1980), but maintains a similar computational role across model-free and dynamic evaluation lookahead control and across experimental paradigms, by including not just the evaluation of actual outcomes but also that of imagined or potential outcomes. Such an extended role can reconcile the suggestion that ventral striatum serves as the critic in an implementation of a model-free RL algorithm with evidence for its more direct involvement in decision making as demonstrated by effects, such as PIT.

A specific prediction of this extended role for ventral striatum is that there should be value-related neural activation during expectancy-based decisions, such as dynamic evaluation lookahead and specific PIT. The data of van der Meer and Redish (2009a), as well as those of others (German and Fields, 2007) are consistent with this proposal. German and Fields (2007) found that in a morphine-conditioned place preference task in a three-chamber environment, ventral striatal neurons that were selectively active in one of the chambers tended to be transiently active just before the rats initiated a journey toward that particular chamber. However, it is not known (in either study) whether these representations encode only a scalar value representation (good, bad) or reflect a specific outcome (such as food or water); value manipulations could address this issue. Although the time course of reward cell firing at the choice point reported by van der Meer and Redish (2009a) suggests a possible relationship with the behavioral strategy used (dynamic evaluation lookahead versus model-free cached values), it would be useful to verify this with a behavioral intervention, such as devaluation. Finally, the temporal relationship between this putative ventral striatal evaluation signal and outcome signals elsewhere is not known. For instance, the spatio-temporal distribution of the non-local reward cell activity in ventral striatum matched that of hippocampal "sweeps"; whether these effects coincide on the millisecond time scale of cognition is still an open question. Interestingly, there is evidence that hippocampal activity can selectively impact reward-related neurons in ventral striatum (Lansink et al., 2008). A possible mechanism for organizing relevant inputs to ventral striatum could be provided by gamma

[6]Recall that in specific PIT, presentation of, e.g., a light CS that predicts a food CS, biases the animal toward taking the action that leads to food. Holland (2004) showed that this effect was not diminished by devaluing the food.

oscillations mediated by fast-spiking interneurons (Berke, 2009; van der Meer and Redish, 2009b); consistent with this idea, van der Meer and Redish (2009b) found that ~80 Hz gamma oscillations, which are prominent in ventral striatal afferents including the hippocampus and frontal cortices, were increased specifically at the final choice point during early learning.

There is, however, an intriguing challenge to the role of ventral striatum as the evaluator in dynamic evaluation lookahead: the way in which expectancies can influence choice behavior may depend on the way in which they are generated. In particular, behavior under the influence of specific PIT effects is not sensitive to devaluation of the US[6], even though the procedure itself produces choice behavior requiring a representation of that US (Holland, 2004). This result suggests that while specific PIT and dynamic evaluation lookahead both depend on the generation of a specific outcome expectancy, the existence of such an expectancy alone is not sufficient for dynamic evaluation in decision making. It raises the question of how the different impacts of internally generated versus cued outcome expectancies are implemented on the neural level. In experimental settings used to identify outcome representations with recording techniques, different ways of generating expectancies can be difficult to distinguish because of the presence of reward-predictive cues (e.g., Colwill and Rescorla 1988; Schoenbaum et al. 1998). To the extent that the static spatial setting of the Multiple-T maze contains reward-predictive cues, they are not specific or maximally predictive at the choice point, such that the representations of reward at the choice point reported by van der Meer and Redish (2009a) are unlikely to result from Pavlovian associations, but instead are likely to reflect internally generated expectancies. However, little is known about the mechanism by which expectancies become linked to particular actions; two recent reports finding action-specific value representations in ventral striatum (Ito and Doya, 2009; Roesch et al., 2009) can provide a basis for investigating this issue.

In summary, the results obtained by van der Meer and Redish (2009a) show that ventral striatal representations of reward can be activated not just by the delivery of actual reward, but also during decision making. The spatio-temporal specificity of this effect suggests that covert representation of reward in ventral striatum may contribute to internally generated, dynamic evaluation lookahead. A role for ventral striatum as evaluating, or translating to action, the motivational relevance of internally generated expectancies is a natural extension of its commonly proposed role as critic. Future work may address the content of its neural representations during procedures that seem to generate expectancies with different properties, such as reinforcer devaluation and PIT, as well as its relationship to individual choices and other outcome-specific signals in the brain.

ACKNOWLEDGMENTS

We thank Bruce Overmier and Adam Steiner for their comments on an earlier version of the manuscript, and Kenji Doya, Yael Niv, Geoffrey Schoenbaum, and Eric Zilli for discussion.

REFERENCES

Adams, C. D., and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. B*, 2, 109–121.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., and O'Reilly, R. C. (2007). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat. Neurosci.* 10, 126–131.

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

Berke, J. D. (2009). Fast oscillations in cortical-striatal networks switch frequency following rewarding events and stimulant drugs. *Eur. J. Neurosci.* 30, 848–859.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Acad. Sci. Imp. Petr.* 5, 175–192.

Blaisdell, A. P., Sawa, K., Leising, K. J., and Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, 311, 1020–1022.

Bolles, R. C. (1972). Reinforcement, expectancy, and learning. *Psychol. Rev.* 77, 32–48.

Buckner, R. L., and Carroll, D. C. (2007). Self-projection and the brain. *Trends Cogn. Sci.* 11, 49–57.

Cardinal, R. N., Parkinson, J. A., Hall, J., and Everitt, B. J. (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci. Biobehav. Rev.* 26, 321–352.

Colwill, R., and Rescorla, R. (1985). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *J. Exp. Psychol.* 1, 520–536.

Colwill, R., and Rescorla, R. (1988). Associations between the discriminative stimulus and the reinforcer in instrumental learning. *J. Exp. Psychol. Anim. Behav. Process.* 14, 155–164.

Corbit, L. H., and Janak, P. H. (2007). Inactivation of the lateral but not medial dorsal striatum eliminates the excitatory impact of pavlovian stimuli on instrumental responding. *J. Neurosci.* 27, 13977–13981.

Corbit, L. H., Muir, J. L., and Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. *J. Neurosci.* 21, 3251–3260.

Cromwell, H. C., and Schultz, W. (2003). Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *J. Neurophysiol.* 89, 2823–2838.

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

Dayan, P. (2002). Motivated reinforcement learning. In Advances in Neural Information Processing Systems, Vol. 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds (Cambridge, MA, MIT Press), pp. 11–18.

de Borchgrave, R., Rawlins, J. N. P., Dickinson, A., and Balleine, B. W. (2002). Effects of cytotoxic nucleus accumbens lesions on instrumental conditioning in rats. *Exp. Brain Res.* 144, 50–68.

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

Dickinson, A., and Dawson, G. (1987). Pavlovian processes in the motivational control of instrumental performance. *Q. J. Exp. Psychol. B*, 39, 201–213.

Domjan, M. (1998). The Principles of Learning and Behavior, 4th edn. USA, Brooks/Cole.

Edwards, W. (1954). The theory of decision making. *Psychol. Bull.* 51, 380–417.

Estes, W. (1948). Discriminative conditioning. II. Effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *J. Exp. Psychol.* 38, 173–177.

Garcia, J., Kovner, R., and Green, K. (1970). Cue properties vs. palatability of flavors in avoidance learning. *Psychon. Sci.* 20, 3–314.

German, P. W., and Fields, H. L. (2007). Rat nucleus accumbens neurons persistently encode locations associated with morphine reward. *J. Neurophysiol.* 97, 2094–2106.

Gilbert, D. T., and Wilson, T. D. (2007). Prospection: experiencing the future. *Science*, 317, 1351–1354.

Glimcher, P. W., Camerer, C., Poldrack, R. A., and Fehr, E., eds (2008). Neuroeconomics: Decision Making and the Brain. New York, Academic Press.

Groenewegen, H. J., Vermeulen-Van der Zee, E., te Kortschot, A., and Witter, M. P. (1987). Organization of the projections from the subiculum to the ventral striatum in the rat. a study using anterograde transport of phaseolus vulgaris leucoagglutinin. *Neuroscience*, 23, 103–120.

Hassabis, D., Kumaran, D., Vann, S. D., and Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl. Acad. Sci. USA*, 104, 1726–1731.

Hebb, D. O. (1949). The Organization of Behavior. New York, Wiley.

Holland, P. C. (2004). Relations between pavlovian-instrumental transfer and reinforcer devaluation. *J. Exp. Psychol. Anim. Behav. Proc.* 30, 104–117.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Models of Information Processing in the Basal Ganglia, J. C. Houk, J. L. Davis and D. G. Beiser, eds (Cambridge, MA, MIT Press), pp. 249–270.

Ito, M., and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* 29, 9861–9874.

Ito, R., Robbins, T. W., Pennartz, C. M., and Everitt, B. J. (2008). Functional interaction between the hippocampus and nucleus accumbens shell is necessary for the acquisition of appetitive spatial context conditioning. *J. Neurosci.* 28, 6950–6959.

Johnson, A., and Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* 27, 12176–12189.

Johnson, A., Fenton, A. A., Kentros, C., and Redish, A. D. (2009a). Looking for cognition in the structure within the noise. *Trends Cogn. Sci.* 13, 55–64.

Johnson, A. W., Gallagher, M., and Holland, P. C. (2009b). The basolateral amygdala is critical to the expression of pavlovian and instrumental outcome-specific reinforcer devaluation effects. *J. Neurosci.* 29, 696–704.

Khamassi, M., Mulder, A. B., Tabuchi, E., Douchamps, V., and Wiener, S. I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. *Eur. J. Neurosci.* 28, 1849–1866.

Krieckhaus, E., and Wolf, G. (1968). Acquisition of sodium by rats: interaction of innate mechanisms and latent learning. *J. Comp. Physiol. Psychol.* 65, 197–201.

Kruse, J., Overmier, J., Konz, W., and Rokke, E. (1983). Pavlovian conditioned stimulus effects upon instrumental choice behavior are reinforcer specific. *Learn. Motiv.* 14, 165–181.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., Joosten, R. N. J. M. A., McNaughton, B. L., and Pennartz, C. M. A. (2008). Preferential reactivation of motivationally relevant information in the ventral striatum. *J. Neurosci.* 28, 6372–6382.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., and Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* 7, e1000173. doi: 10.1371/journal.pbio.1000173.

Lavoie, A. M., and Mizumori, S. J. (1994). Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. *Brain Res.* 638, 157–168.

Miyazaki, K., Mogi, E., Araki, N., and Matsumoto, G. (1998). Reward-quality dependent anticipation in rat nucleus accumbens. *Neuroreport*, 9, 3943–3948.

Mogenson, G. J., Jones, D. L., and Yim, C. Y. (1980). From motivation to action: functional interface between the limbic system and the motor system. *Prog. Neurobiol.* 14, 69–97.

Morris, R. G. M., Garrud, P., Rawlins, J. N. P., and O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, 297, 681–683.

Newell, A., and Simon, H. A. (1972). Human Problem Solving. Englewood Cliffs, NJ, Prentice-Hall.

Niv, Y., Joel, D., and Dayan, P. (2006). A normative perspective on motivation. *Trends Cogn. Sci.* 10, 375–381.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304, 452–454.

Packard, M. G., and McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol. Learn. Mem.* 65, 65–72.

Pickens, C. L., Saddoris, M. P., Gallagher, M., and Holland, P. C. (2005). Orbitofrontal lesions impair use of cue-outcome associations in a devaluation task. *Behav. Neurosci.* 119, 317–322.

Preuschoff, K., Bossaerts, P., and Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, 51, 381–390.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* 9, 545–556.

Redish, A. D. (1999). Beyond the Cognitive Map. Cambridge, MA, MIT Press.

Roesch, M. R., Singh, T., Brown, P. L., Mullins, S. E., and Schoenbaum, G. (2009). Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *J. Neurosci.* 29, 13365–13376.

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

Schmitzer-Torbert, N. C., and Redish, A. D. (2002). Development of path stereotypy in a single day in rats on a multiple-T maze. *Arch. Ital. Biol.* 140, 295–301.

Schoenbaum, G., and Setlow, B. (2003). Lesions of nucleus accumbens disrupt learning about aversive outcomes. *J. Neurosci.* 23, 9833–9841.

Schoenbaum, G., Chiba, A. A., and Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nat. Neurosci.* 1, 155–159.

Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.* 12, 4595–4610.

Setlow, B., Schoenbaum, G., and Gallagher, M. (2003). Neural encoding in ventral striatum during olfactory discrimination learning. *Neuron*, 38, 625–636.

Shallice, T. (1982). Specific impairments of planning. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 298, 199–209.

Sutton, R. (1990). First results with Dyna, an integrated architecture for learning, planning and reacting. In Neural Networks for Control, W. T. Miller, R. S. Sutton, and P. J. Werbos, eds (Cambridge, MA, MIT Press), pp. 179–189.

Sutton, R., and Barto, A. (1998). Reinforcement Learning: An Introduction. Cambridge, MA, MIT Press.

Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008). Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Front. Neurosci.* 2, 86–99. doi:10.3389/neuro.01.014.2008.

Talmi, D., Seymour, B., Dayan, P., and Dolan, R. J. (2008). Human pavlovian-instrumental transfer. *J. Neurosci.* 28, 360–368.

Tolman, E. C. (1932). Purposive Behavior in Animals and Men. New York, Appleton-Century-Crofts.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208.

Urcuioli, P. J. (2005). Behavioral and associative effects of differential outcomes in discrimination learning. *Learn. Behav.* 33, 1–21.

van der Meer, M. A. A., and Redish, A. D. (2009a). Covert expectation-of-reward in rat ventral striatum at decision points. *Front. Integr. Neurosci.* 3, 1. doi: 10.3389/neuro.07.001.2009.

van der Meer, M. A. A., and Redish, A. D. (2009b). Low and high gamma oscillations in rat ventral striatum have distinct relationships to behavior, reward, and spiking activity on a learned spatial decision task. *Front. Integr. Neurosci.* 3, 9. doi: 10.3389/neuro.07.009.2009.

Williams, G., Rolls, E., Leonard, C., and Stern, C. (1993). Neuronal responses in the ventral striatum of the behaving macaque. *Behav. Brain Res.* 55, 243–252.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* 19, 181–189.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 October 2009; paper pending published: 24 October 2009; accepted: 10 November 2009; published: 15 April 2010.

*Citation: Front. Neurosci. (2010) **4**, 1: 29–37. doi: 10.3389/neuro.01.006.2010*