Published in final edited form as:

Behav Neurosci. 2017 October ; 131(5): 385–391. doi:10.1037/bne0000211.

Ventral striatum lesions do not affect reinforcement learning with deterministic outcomes on slow time scales

Raquel Vicario-Feliciano, Elisabeth A. Murray, and Bruno B. Averbeck

Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, 20892-4415 USA

Abstract

A large body of work has implicated the ventral striatum (VS) in aspects of reinforcement learning (RL). However, less work has directly examined the effects of lesions in the VS, or other forms of inactivation, on two-armed bandit RL tasks. We have recently found that lesions in the VS in macaque monkeys affect learning with stochastic schedules, but have minimal effects with deterministic schedules. The reasons for this are not currently clear. Because our previous work used short intertrial intervals, one possibility is that the animals were using working memory to bridge stimulus-reward associations from one trial to the next. In the present study, we examined learning of 60 pairs of objects, where the animals received only one trial per day with each pair. The large number of object pairs and the long interval (approximately 24 hours) between trials with a given pair minimized the chances that the animals could use working memory to bridge trials. We found that monkeys with VS lesions were unimpaired relative to controls, which suggests that animals with VS lesions can still learn to select rewarded objects even when they cannot make use of working memory.

Keywords

decision making; reinforcement learning; behavior; dopamine; ventral striatum

Introduction

Reinforcement learning (RL) is the behavioral process of learning to make choices that lead to rewarding outcomes, or to the avoidance of harmful or otherwise aversive outcomes (Averbeck & Costa, 2017). There is a well-developed RL theory literature with models that can account for learning in many contexts (Dayan & Daw, 2008; Mackintosh, 1994; Montague, Dayan, & Sejnowski, 1996; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Sutton & Barto, 1998). A fundamental component of many of these learning algorithms, in one form or another, are reward prediction errors (RPEs). Experimental work (Schultz, 2015) has shown that dopamine neurons code these RPEs. Because there is a large dopamine innervation of the VS, it has been proposed that it integrates RPEs and maintains an ongoing representation of the value of choices (Collins & Frank, 2014; Doya, 2000; Frank, 2005;

Houk, Adamas, & Barto, 1995). This hypothesis has been well supported by imaging data that has shown consistent changes in blood-oxygen-level-dependent (BOLD) signals in the VS that correlate with RPEs (J. O'Doherty et al., 2004; J. P. O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Rutledge, Dean, Caplin, & Glimcher, 2010).

Consistent with the hypothesis that the VS is a key site for the representation of the values of choices, recent work has shown that lesions in the VS can lead to deficits in RL tasks with stochastic reward schedules (Costa, Dal Monte, Lucas, Murray, & Averbeck, 2016). This work has also shown that VS lesions result in smaller deficits than do lesions in the amygdala on this same task. Further, when reward schedules are deterministic, relatively small deficits are seen in monkeys with VS lesions, and the impairment appears to be driven by fast responding in the presence of a speed-accuracy trade-off. One possible explanation for the lack of impairment is that the monkeys with VS lesions were using a win-stay/lose-switch strategy, using prefrontal working memory systems to store the outcome from the last trial, and using that outcome to guide choices. This was possible because each block of trials evaluated learning with a single pair of images and intertrial intervals were on the order of a second. Thus, working memory could bridge the interval from one trial to the next. Consistent with this, it has been shown that working memory can play an important role in RL (Collins & Frank, 2012). Alternatively, after receiving feedback on the outcome of a given trial, monkeys may have used prospective coding of the choice to be made on the next trial in combination with working memory, a phenomenon thought to underlie learning set formation (Murray & Gaffan, 2006). In addition to working memory and prospective coding explanations, the animals may have also been using some form of episodic memory to store outcomes (Gershman & Daw, 2017).

In the present study, we sought to evaluate these possibilities by comparing learning rates of monkeys with VS lesions and unoperated controls on a concurrent object discrimination learning task (Izquierdo & Murray, 2007; Thornton, Rothblat, & Murray, 1997). In this task, monkeys were given 60 trials per day, each with a different pair of objects. One of the objects in each pair was associated with a reward and the other was not. These stimulus-outcome pairings remained fixed across days. Because there were a large number of pairs and 24 hours intervened between trials for each pair, it was improbable that the monkeys could use working memory to solve the task. We show that monkeys with VS lesions are unimpaired relative to unoperated controls in this task. Accordingly, the good learning by monkeys with VS lesions on visual discriminations with deterministic reward schedules does not depend on either working memory or prospective coding.

Materials and Methods

Subjects

We tested 15 rhesus macaques (*Macaca mulatta*; 12 males and 3 females) that weighed between 4.6–9.7 kg. All monkeys were experimentally naïve at the start of the experiment. Three of the monkeys (male) had bilateral excitotoxic VS lesions. Lesions were made with injections of quinolinic acid placed stereotactically, with predetermined coordinates from MRI scans, into the VS. These were the same operated monkeys studied by Costa et al (2016). The data from the VS lesion animals were compared to data from 11 unoperated

control monkeys. One male healthy control was excluded from analysis because his behavior deteriorated in the middle of training. Therefore, removing him increased the likelihood of observing a deficit in the VS lesioned animals relative to the unoperated controls. All monkeys were fed a controlled diet (approximately 14 biscuits per day) of primate chow (no. 5038; PMI Feeds, St. Louis, MO) to ensure motivation to complete the task while still maintaining optimal body weight (more than 85% of their body weight before the beginning of training). All monkeys had full access to water during the course of the experiment, and the primate chow was supplemented with fruit. All experimental procedures were performed in accordance with the Guide for the Care and Use of Laboratory Animals and were approved by the National Institute of Mental Health Animal Care and Use Committee.

Pretraining and Food Preference Test

Monkeys were familiarized with the Wisconsin General Testing Apparatus (WGTA). Before starting the main task, monkeys were allowed to take food freely from a test tray measuring 19.2 cm × 72.7 cm × 1.9 cm (width × length × height, respectively). After monkeys were familiarized with taking food rewards from one of three wells in the tray, one of the wells was covered with a plaque. After monkeys learned to displace the plaques presented singly over each well, to obtain food reward, the plaques were replaced with three “junk” objects dedicated to the pretraining phase. Similar to the plaques, if the monkey displaced the single object covering the well, it was allowed to take the food reward hidden underneath. Prior to starting the main task, all monkeys also completed a food preference test, using six different foods. During the test, they were presented with each possible pair of two food choices, one in each well, for a total of ten times per session. The position (left or right) of the food was always counterbalanced by session, with each session having 30 trials and 10-s intertrial intervals. Of the possible food choices, monkeys received half a peanut, half a fruit snack, or an M&M.

Main task

Monkeys were tested on a 60-pair concurrent object discrimination learning task (Fig. 1), administered in a WGTA in a dark room. During the task, only the testing compartment was illuminated with two incandescent 60W light bulbs. For each of the 60 pairs, one object was rewarded ($S_{1+} \dots S_{60+}$) and the other was not ($S_{1-} \dots S_{60-}$). In each trial, the rewarded object was pseudorandomly presented on the left or right. If the monkey chose the rewarded object (S_{n+}), it was allowed to take the food reward hidden underneath. Otherwise, the trial was terminated without correction. The order of the pairs was fixed across days, with no repetition of pairs of objects in a given session. Objects assigned positive or negative values were also kept constant across all training sessions. Half the positive objects (S_{+}) were rewarded with one type of food, and half with another type of food; the two foods were chosen to be approximately equally palatable, as judged by the food preference test. Subjects completed sessions until they averaged 90% accuracy across five consecutive sessions (i.e., 270 or more correct choices from 300 trials).

Reinforcement Learning Model

We used a Rescorla-Wagner reinforcement learning model to estimate the learning rates (α) and inverse temperature (β). Value updates were given by:

$$v_{i,j}(k+1) = v_{i,j}(k) + \alpha (R(k) - v_{i,j}(k))$$

where $v_{i,j}$ was the value estimate for option i ($i \in \{1, 2\}$) of pair j ($j \in \{1, \dots, 60\}$), in testing session k . The variable R was the reward feedback for the current choice (0 or 1), and α was the learning rate parameter. We used a logistic function to generate choice probability estimates using the values ($v_{i,j}$) of the objects, i , in each pair, j , in all of the sessions, k , completed by each of the monkeys:

$$d_{1,j}(k) = (1 + e^{\beta(v_{2,j}(k) - v_{1,j}(k))})^{-1}, d_{2,j}(k) = 1 - d_{1,j}(k)$$

The likelihood of the data given the model was therefore given by:

$$f(D|\beta, \alpha) = \prod_{k,j} [d_{1,j}(k)c(k) + d_{2,j}(k)(1-c(k))]$$

For every pair in each session, $c(k)$ had a value of 1 if option 1 was chosen and a value of 0 if option 2 was chosen. We used standard function optimization methods to maximize the values of the log of the likelihood.

Results

Animals were run on a concurrent learning task. Sixty pairs of objects were presented for choice each day, with one pair shown on each trial (Fig. 1). One of the objects in each pair was always associated with reward and the other was not. Because there were many object pairs and they were presented at 24-hour intervals, learning required the monkeys to integrate the rewarding outcome across days. Animals with VS lesions completed between 15–28 sessions. Unoperated controls completed between 10–37 sessions. Under these conditions, it seems unlikely that they would have been able to use an explicit memory system to remember the outcomes associated with each pair on previous days. In addition, their learning was relatively slow, perhaps also indicative of an integrative learning mechanism.

We first evaluated the accuracy with which each monkey chose the correct object, from each of the 60 pairs, across days. We found that the overall fraction of correct choices did not differ across sessions between groups (Fig. 2A; $t(65)=0.941$). Because the fraction correct does not take into account learning rates, we next used an RL model to analyze the behavior of both groups. Implementing an RL model allowed us to determine, for each monkey, the learning rate (α) and the choice consistency, or inverse temperature (β). In this model, the learning rate characterized the rate at which performance approached asymptote, and the choice consistency characterized the asymptotic performance. We found no differences between groups in learning rate, α , (Fig. 2B; $t(14)=0.320$) or inverse temperature, β (Fig. 2C; $t(14)=0.899$).

Next, we evaluated the overall fraction of correct choices between groups for the first ten sessions, as the first ten sessions included all of the animals. This allowed us to more accurately study the behavior of the animals before they started to reach criterion. We found that the overall fraction of correct choices did not differ across sessions between groups (Fig. 3A; $t(20)=0.756$). Again, when we implemented the RL model, we found no differences between the two groups in learning rate, α , (Fig. 3B; $t(14)=0.321$) or inverse temperature, β (Fig. 3C; $t(14)=0.849$).

Discussion

We examined the effects of VS lesions on a concurrent object discrimination learning task. We used 60 pairs of objects, and the animals only received one trial per day with each pair. One object in each pair was always rewarded and the other was never rewarded. The large number of objects and long intervals between trials made it unlikely that animals could use neural systems that subserve working memory or prospective coding to solve the task. In support of this contention, tasks with relatively short delays between trials with a given pair of objects lead to development of discrimination learning set, reversal learning set and prospective coding, whereas tasks like the present one with long intertrial intervals do not (Browning, Easton, & Gaffan, 2007; Murray & Gaffan, 2006; Wilson & Gaffan, 2008). When we assessed overall accuracy and parameters estimated using a reinforcement learning model, we found no differences between animals with VS lesions and unoperated controls in the task. These results are consistent with our previous work that showed minimal deficits on a 2-armed bandit reinforcement learning task with deterministic rewards, following VS lesions (Costa et al., 2016). However, in the previous study, only one pair of objects was learned in each block, and the intertrial interval was only a few seconds, so the animals could have solved the task using either working memory or prospective coding. The present results do not preclude the possibility that performance of the VS lesion animals in the previous task was partially mediated by working memory. In fact, given the short intertrial intervals and the repetition of the same stimuli, some contribution of working memory seems likely. However, the present results do show that learning to select rewarding stimuli in bandit tasks, under some conditions, can be accomplished without a VS and without working memory to bridge delays.

There are at least two related hypotheses about the role of the VS in behavior. One suggests that the VS integrates RPEs to generate a value estimate for choices (Collins & Frank, 2014; Frank, 2005; Houk et al., 1995). Functional imaging studies have shown that these RPEs correlate with the BOLD signal in the VS in Pavlovian and instrumental tasks (J. O'Doherty et al., 2004; J. P. O'Doherty et al., 2003; Rutledge et al., 2010). Several early studies used Pavlovian paradigms, in which a cue predicted delivery of juice to subjects, to try to understand these RPEs (Berns, McClure, Pagnoni, & Montague, 2001; McClure, Berns, & Montague, 2003; Pagnoni, Zink, Montague, & Berns, 2002). They found positive and negative temporal reward prediction error signals in the VS, when the time of juice delivery was varied. However, these signals were violations of expectations after learning, and acquisition of associations was not examined. Additional studies have also shown aversive prediction error responses in second order pain conditioning tasks, after acquisition of associations (Seymour et al., 2004).

Further studies have examined BOLD responses during the acquisition of Pavlovian associations between cues and pleasant tasting juice, combined with temporal-difference RL models (J. P. O'Doherty et al., 2003). These studies have found correlations with appetitive prediction errors at the time of the cue (CS+) and at the time of the reward (US) in the VS (specifically the ventral putamen). In contrast to earlier studies, which used only Pavlovian paradigms (i.e., the subjects did not choose among options), prediction error signals have also been studied in a choice task, and compared to those of a matched Pavlovian task (J. O'Doherty et al., 2004). The results from this choice study provide evidence in favor of the dorsal striatum coding prediction errors specifically in choice tasks, and the VS coding prediction errors in both choice and Pavlovian tasks, supporting the hypothesis that the dorsal striatum is consistent with an actor that learns action values, and the ventral striatum is consistent with a critic that learns state values (Sutton & Barto, 1998).

The finding that the VS codes appetitive RPEs has been replicated across several additional papers that have examined other aspects of learning, including the explore-exploit trade-off (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), risk based decision making (Niv, Edlund, Dayan, & O'Doherty, 2012), and formal approaches that try to rigorously show that the striatal BOLD signal is coding RPEs (Rutledge et al., 2010). However, across the studies, the BOLD signal in the VS sometimes has been monotonically related to positive RPEs, and sometimes monotonically related to negative RPEs. Therefore, the VS BOLD signal does not consistently encode positive RPEs monotonically.

In another line of research, the VS has been hypothesized to interface between stimulus-outcome associations formed in the amygdala and the motor system (Belin, Jonkman, Dickinson, Robbins, & Everitt, 2009; Cardinal, Parkinson, Hall, & Everitt, 2002; Floresco, 2015; Shiflett & Balleine, 2010). Several paradigms have shown that, following the formation of a Pavlovian association between a neutral cue (CS+) and a reward (US), presentation of the Pavlovian cue can affect ongoing motor behavior. For example, in Pavlovian instrumental transfer (PIT), animals more frequently press a lever associated with a specific reward outcome in the presence of a CS+ that had previously been associated with that outcome. Lesions in the VS reduce the enhanced responding in the presence of the CS+ (Hall, Parkinson, Connor, Dickinson, & Everitt, 2001), as do lesions that disconnect the amygdala and VS (Shiflett & Balleine, 2010). Similarly, in conditioned reinforcement, animals work to obtain a CS+ previously associated with reward, and the CS+ can support new learning. Lesions of the basolateral amygdala (Parkinson et al., 2001) and of the VS (McDannald, Setlow, & Holland, 2013) disrupt the processes underlying conditioned reinforcement. In addition, injection of amphetamines into the VS potentiates conditioned reinforcement (Burns, Robbins, & Everitt, 1993). At the same time, rats with VS lesions learn responses for primary rewards. These and related findings have led to the idea that the VS is essential, specifically for acquiring or expressing the incentive properties of stimuli that have been associated with reward.

In addition to the Pavlovian effects on instrumental behavior, previous work has also compared the effects of NAc core vs. shell lesions on deterministic and stochastic reversal learning, where the rats had to learn which of two levers was more rewarding (Dalton, Phillips, & Floresco, 2014; Floresco, Ghods-Sharifi, Vexelman, & Magyar, 2006). These

studies found that NAc shell lesions affected probabilistic reversal learning, whereas NAc core lesions affected switching of response strategies, and not vice versa. Additional work (Saddoris, Cacciapaglia, Wightman, & Carelli, 2015) has also shown that core dopamine tracks prediction error, while shell dopamine tracks motivationally salient stimuli, and dopamine release in the NAc is specific to sign tracking rats that learn to approach reward predicting cues (Flagel et al., 2011). However, no studies have identified functional differences between these structures in monkeys (Meredith 1996; Friedman 2002). Nonetheless, because structural separations have been shown between them, our lesions covered both.

In our previous work, we found that lesions in the VS decreased reaction times to saccade to a peripheral target (Costa et al., 2016). This may be consistent with the previous results in rodents that showed that the VS mediated the impact of a CS+ on the rate of behavioral responding (Hall et al., 2001). However, we also found that our animals exhibited a speed accuracy trade-off, such that picking the more rewarding option happened more frequently when responses were slower, up to about 220 ms. This was true across both the lesioned and control animals. Because the monkeys with VS lesions were faster to respond, their behavioral deficit in the deterministic condition was driven by responding before their internal decision mechanisms had settled on the correct response (Frank, Samanta, Moustafa, & Sherman, 2007; Lee, Seo, Dal Monte, & Averbeck, 2015). When behavioral performance was compared between groups, using subsets of trials in each group that resulted in matched reaction time distributions, there was no longer a performance deficit in the VS lesioned group. There were larger deficits when stochastic reward schedules were used (e.g., when one option was rewarded 70% of the time and the other 30% of the time), and these could not be accounted for by matching reaction times.

In addition to reinforcement learning and working memory, episodic memory could also contribute to learning in our tasks (Gershman & Daw, 2017; Lengyel & Dayan, 2007). It has been proposed that, when state spaces are large or continuous, and little data is available, neither model based nor model free learning mechanisms can effectively mediate behavior. In this situation, episodic memory can provide explicit recall of action sequences and their consequences. Previous studies have found, however, that medial-temporal lobe lesions that remove the hippocampus, or the hippocampus together with the adjacent entorhinal, perirhinal, and parahippocampal cortex, regions often thought to be important for episodic memory, do not affect learning in the concurrent object discrimination learning task used in the present study (Chudasama, Wright, & Murray, 2008; Malamut, Saunders, & Mishkin, 1984). Furthermore, bandit tasks have complex state spaces but relatively simple sufficient statistics (Averbeck, 2015). Episodic memory, however, could contribute to some of the performance in this task by facilitating the memory of a few of the choice-outcome events.

The present study contributes to delineating the role of the VS in reinforcement learning. The VS appears to play a minimal role in learning when object choices are deterministically rewarded, as in the current study. However, when rewards are delivered probabilistically, VS lesions have a larger effect on learning, as shown in our previous work (Costa et al., 2016). One possible explanation for this finding is that, in intact monkeys, conditioned reinforcement processes reinforce chosen cues in stochastic paradigms in trials where

rewards (USs) are not delivered. If so, and if VS lesions disrupt the conditioned reinforcement process that normally bridges unrewarded trials in learning paradigms where rewards are delivered stochastically, then monkeys with VS lesions would be impaired. Whether the specific plasticity that underlies probabilistic learning is happening in the VS, or the learning is happening elsewhere (e.g., the amygdala) is not yet clear. It is possible that the VS is inheriting much of its value representation from the amygdala, and that the role of the VS is to interface this value representation with a unique set of areas with which the VS interacts (Averbeck & Costa, 2017). Further work will be required to test these hypotheses.

Acknowledgments

We thank Dawn Anuszkiewicz-Lundgren for behavioral training and testing.

References

- Averbeck BB. Theory of choice in bandit, information sampling and foraging tasks. *PLoS computational biology*. 2015; 11(3):e1004164.doi: 10.1371/journal.pcbi.1004164 [PubMed: 25815510]
- Averbeck BB, Costa VD. Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience*. 2017; 20(4):505–512. DOI: 10.1038/nn.4506 [PubMed: 28352111]
- Belin D, Jonkman S, Dickinson A, Robbins TW, Everitt BJ. Parallel and interactive learning processes within the basal ganglia: relevance for the understanding of addiction. *Behavioural brain research*. 2009; 199(1):89–102. DOI: 10.1016/j.bbr.2008.09.027 [PubMed: 18950658]
- Berns GS, McClure SM, Pagnoni G, Montague PR. Predictability modulates human brain response to reward. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2001; 21(8):2793–2798. [PubMed: 11306631]
- Browning PG, Easton A, Gaffan D. Frontal-temporal disconnection abolishes object discrimination learning set in macaque monkeys. *Cereb Cortex*. 2007; 17(4):859–864. DOI: 10.1093/cercor/bhk039 [PubMed: 16707734]
- Burns LH, Robbins TW, Everitt BJ. Differential effects of excitotoxic lesions of the basolateral amygdala, ventral subiculum and medial prefrontal cortex on responding with conditioned reinforcement and locomotor activity potentiated by intra-accumbens infusions of D-amphetamine. *Behavioural brain research*. 1993; 55(2):167–183. [PubMed: 8357526]
- Cardinal RN, Parkinson JA, Hall J, Everitt BJ. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and biobehavioral reviews*. 2002; 26(3):321–352. [PubMed: 12034134]
- Chudasama Y, Wright KS, Murray EA. Hippocampal lesions in rhesus monkeys disrupt emotional responses but not reinforcer devaluation effects. *Biological psychiatry*. 2008; 63(11):1084–1091. DOI: 10.1016/j.biopsych.2007.11.012 [PubMed: 18191111]
- Collins AG, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European journal of neuroscience*. 2012; 35(7):1024–1035. DOI: 10.1111/j.1460-9568.2011.07980.x [PubMed: 22487033]
- Collins AG, Frank MJ. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*. 2014; 121(3):337–366. DOI: 10.1037/a0037015 [PubMed: 25090423]
- Costa VD, Dal Monte O, Lucas DR, Murray EA, Averbeck BB. Amygdala and Ventral Striatum Make Distinct Contributions to Reinforcement Learning. *Neuron*. 2016; 92(2):505–517. DOI: 10.1016/j.neuron.2016.09.025 [PubMed: 27720488]
- Dalton GL, Phillips AG, Floresco SB. Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2014; 34(13):4618–4626. DOI: 10.1523/JNEUROSCI.5058-13.2014 [PubMed: 24672007]

- Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. *Nature*. 2006; 441(7095):876–879. [PubMed: 16778890]
- Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. *Cognitive, affective & behavioral neuroscience*. 2008; 8(4):429–453. DOI: 10.3758/CABN.8.4.429
- Doya K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current opinion in neurobiology*. 2000; 10(6):732–739. [PubMed: 11240282]
- Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, ... Akil H. A selective role for dopamine in stimulus-reward learning. *Nature*. 2011; 469(7328):53–57. DOI: 10.1038/nature09588 [PubMed: 21150898]
- Floresco SB. The nucleus accumbens: an interface between cognition, emotion, and action. *Annual Review of Psychology*. 2015; 66:25–52. DOI: 10.1146/annurev-psych-010213-115159
- Floresco SB, Ghods-Sharifi S, Vexelman C, Magyar O. Dissociable roles for the nucleus accumbens core and shell in regulating set shifting. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2006; 26(9):2449–2457. DOI: 10.1523/JNEUROSCI.4431-05.2006 [PubMed: 16510723]
- Frank MJ. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*. 2005; 17(1):51–72. DOI: 10.1162/0898929052880093 [PubMed: 15701239]
- Frank MJ, Samanta J, Moustafa AA, Sherman SJ. Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*. 2007; 318(5854):1309–1312. [PubMed: 17962524]
- Gershman SJ, Daw ND. Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*. 2017; 68:101–128. DOI: 10.1146/annurev-psych-122414-033625
- 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. *The European journal of neuroscience*. 2001; 13(10):1984–1992. [PubMed: 11403692]
- Houk, JC., Adamas, JL., Barto, AG. A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In: Houk, JC., Davis, JL., Beiser, DG., editors. *Models of information processing in the basal ganglia*. Cambridge, MA: MIT Press; 1995. p. 249–274.
- Izquierdo A, Murray EA. Selective bilateral amygdala lesions in rhesus monkeys fail to disrupt object reversal learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2007; 27(5):1054–1062. DOI: 10.1523/JNEUROSCI.3616-06.2007 [PubMed: 17267559]
- Lee E, Seo M, Dal Monte O, Averbeck BB. Injection of a Dopamine Type 2 Receptor Antagonist into the Dorsal Striatum Disrupts Choices Driven by Previous Outcomes, But Not Perceptual Inference. *Journal of Neuroscience*. 2015
- Lengyel M, Dayan P. Hippocampal contributions to control: the third way. *Advances in Neural Information Processing Systems*. 2007; 20:889–896.
- Mackintosh, NJ., editor. *Animal learning and cognition*. 1. New York: Academic Press; 1994.
- Malamut BL, Saunders RC, Mishkin M. Monkeys with combined amygdalo-hippocampal lesions succeed in object discrimination learning despite 24-hour intertrial intervals. *Behavioral neuroscience*. 1984; 98(5):759–769. [PubMed: 6487412]
- McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. *Neuron*. 2003; 38(2):339–346. [PubMed: 12718866]
- McDannald MA, Setlow B, Holland PC. Effects of ventral striatal lesions on first- and second-order appetitive conditioning. *The European journal of neuroscience*. 2013; 38(4):2589–2599. DOI: 10.1111/ejn.12255 [PubMed: 23691939]
- Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci*. 1996; 16(5):1936–1947. [PubMed: 8774460]
- Murray EA, Gaffan D. Prospective memory in the formation of learning sets by rhesus monkeys (*Macaca mulatta*). *Journal of experimental psychology Animal behavior processes*. 2006; 32(1):87–90. DOI: 10.1037/0097-7403.32.1.87 [PubMed: 16435968]

- Niv Y, Edlund JA, Dayan P, O'Doherty JP. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2012; 32(2):551–562. DOI: 10.1523/JNEUROSCI.5498-10.2012 [PubMed: 22238090]
- 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(5669):452–454. DOI: 10.1126/science.1094285 [PubMed: 15087550]
- 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(2):329–337. [PubMed: 12718865]
- Pagnoni G, Zink CF, Montague PR, Berns GS. Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*. 2002; 5(2):97–98. DOI: 10.1038/nn802 [PubMed: 11802175]
- Parkinson JA, Crofts HS, McGuigan M, Tomic DL, Everitt BJ, Roberts AC. The role of the primate amygdala in conditioned reinforcement. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2001; 21(19):7770–7780. [PubMed: 11567067]
- Pearce JM, Hall G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*. 1980; 87(6):532–552. [PubMed: 7443916]
- Rescorla, RA., Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH., PWF, editors. *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts; 1972. p. 64-99.
- Rutledge RB, Dean M, Caplin A, Glimcher PW. Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2010; 30(40):13525–13536. DOI: 10.1523/JNEUROSCI.1747-10.2010 [PubMed: 20926678]
- Saddoris MP, Cacciapaglia F, Wightman RM, Carelli RM. Differential Dopamine Release Dynamics in the Nucleus Accumbens Core and Shell Reveal Complementary Signals for Error Prediction and Incentive Motivation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2015; 35(33):11572–11582. DOI: 10.1523/JNEUROSCI.2344-15.2015 [PubMed: 26290234]
- Schultz W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiological reviews*. 2015; 95(3):853–951. DOI: 10.1152/physrev.00023.2014 [PubMed: 26109341]
- Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, ... Frackowiak RS. Temporal difference models describe higher-order learning in humans. *Nature*. 2004; 429(6992):664–667. DOI: 10.1038/nature02581 [PubMed: 15190354]
- Shiflett MW, Balleine BW. At the limbic-motor interface: disconnection of basolateral amygdala from nucleus accumbens core and shell reveals dissociable components of incentive motivation. *The European journal of neuroscience*. 2010; 32(10):1735–1743. DOI: 10.1111/j.1460-9568.2010.07439.x [PubMed: 21044174]
- Sutton, Richard S., Barto, Andrew G. *Reinforcement learning : an introduction*. Cambridge, Mass: MIT Press; 1998.
- Thornton JA, Rothblat LA, Murray EA. Rhinal cortex removal produces amnesia for preoperatively learned discrimination problems but fails to disrupt postoperative acquisition and retention in rhesus monkeys. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 1997; 17(21):8536–8549. [PubMed: 9334426]
- Wilson CR, Gaffan D. Prefrontal-inferotemporal interaction is not always necessary for reversal learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2008; 28(21):5529–5538. DOI: 10.1523/JNEUROSCI.0952-08.2008 [PubMed: 18495887]

Figure 1.

60-trial discrimination task structure. The monkey chose between two objects. If the monkey chose S_{n+} , it was allowed to take the food reward hidden underneath. If the monkey chose S_{n-} , the trial was terminated. The position of the rewarded side was pseudorandomized across trials and across days. The reward was only associated with the object. The intertrial interval was always 20 s.

Figure 2.

60-pair discrimination across sessions. (A) Mean of rewarded fraction of options across sessions (thick line with error bars). Error bars are standard error of the mean (s.e.m.) across subjects for each day, where n =the number of monkeys that completed a given session. The mean of the probability of choosing one of the two objects, d from the RL model, is shown as a thinner line. Note the jagged nature of the line is because once animals reached criterion they had finished the study and the subsequent mean does not include that animal. (B) Mean learning rate parameter, α , per group (vs or controls). (C) Mean asymptote value, β , per group (vs or controls). Error bars for panels B and C are the standard error of the mean (s.e.m.) across subjects for each group (vs or controls), where $n=3$ for animals with VS lesions and $n=11$ for unoperated controls. Dots in panels B and C are the individual values for each animal.

Figure 3.

60-pair discrimination across the first ten sessions. (A) Mean of rewarded fraction of options across sessions (thick line with error bars). Error bars are standard error of the mean (s.e.m.) across subjects for each day, where n =all monkeys. The mean of the probability of choosing one of the two objects, d from the RL model, is shown as a thinner line. (B) Mean learning rate parameter, α , per group (vs or controls). (C) Mean asymptote value, β , per group (vs or controls). Error bars for panels B and C are the standard error of the mean (s.e.m.) across subjects for each group (vs or controls), where $n=3$ for animals with VS lesions and $n=11$ for unoperated controls. Dots in panels B and C are the individual values for each animal.