ABSTRACT

Title of Dissertation:	THE ROLE OF THE VENTRAL STRIATUM AND AMYGDALA IN REINFORCEMENT LEARNING
	Craig A. Taswell, Doctor of Philosophy, 2021
Dissertation directed by:	Associate Professor, Daniel Butts, Department of Biology

Adaptive behavior requires that organisms choose wisely to gain rewards and avoid punishment. Reinforcement learning refers to the behavioral process of learning about the value of choices, based on previous choice outcomes. From an algorithmic point of view, rewards and punishments exist on opposite sides of a single value axis. However, simple distinctions between rewards and punishments and their theoretical expression on a single value axis hide considerable psychological complexities that underlie appetitive and aversive reinforcement learning. A broad set of neural circuits, including the amygdala and frontal-striatal systems, have been implicated in mediating learning from gains and losses. The ventral striatum (VS) and amygdala have been implicated in several aspects of this

process. To examine the role of the VS and amygdala in learning from gains and losses, we compared the performance of macaque monkeys with VS lesions, with amygdala lesions, and un-operated controls on a series of reinforcement learning tasks. In these tasks monkeys gained or lost tokens, which were periodically cashed out for juice, as outcomes for choices. We found that monkeys with VS lesions had a deficit in learning to choose between cues that differed in reward magnitude. Monkeys with VS lesions performed as well as controls when choices involved a potential loss. In contrast, we found that monkeys with amygdala lesions performed as well as controls across all conditions. Further analysis revealed that the deficits we found in monkeys with VS lesions resulted from a reduction in motivation, rather than the monkeys' inability to learn the stimulus-outcome contingency.

THE ROLE OF THE VENTRAL STRIATUM AND AMYGDALA IN REINFORCEMENT LEARNING

Craig A. Taswell

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Associate Professor Daniel Butts, Chair
Dr. Bruno Averbeck
Professor Matthew Roesch
Associate Professor Quentin Gaudry
Dr. Hugo Tejeda

Table of Contents

Table of Contents ii
Chapter 1: Introduction 1
1.1 Reinforcement learning 1
1.1.1 RL models 2
1.1.2 History of RL & dopamine 4
1.2 Behavior 8
1.2.1 Learning 8
1.2.2 Consequences 11
1.2.3 Conditioning 14
1.2.4 Pavlovian-instrumental transfer (PIT) 18
1.3 Dopamine 22
1.3.1 Dopamine & effort 26
1.3.2 Dopamine & learning 28
1.3.3 Dopamine's role? 30
1.3.4 Dopamine theory 32
1.3.5 Optogenetic activation of VTA dopamine 34
1.4 Studying learning systems 43
1.4.1 Response rates 43
1.4.2 Lesions & learning 46
1.4.3 Two-Process learning theories 53
1.4.4 Linking dopamine & learning 58
1.5 Proposal 61
1.5.1 Studying learning & our approach 61
Chapter 2: Effects Amygdala Lesions on Object-Based versus Action-Based Learning in Macaques 65
2.1 Introduction 65
2.2 Methods 68
2.2.1 Subjects 68
2.2.2 Surgery 68
2.2.3 Lesion assessment 69
2.2.4 Task & apparatus 70
2.2.5 Task training 71
2.2.6 Eye tracking 72
2.2.7 Bayesian model of reversal learning 73
2.2.8 Reinforcement learning model of choice behavior 76
2.2.9 ANOVA models 79
2.3 Results 80
2.3.1 Choice behavior 82

2.3.2 Reinforcement learning model 88
2.3.3 Reversals 91
2.3.4 Block type 98
2.4 Discussion 101
2.4.1 Conclusion 112
2.5 Supplemental Material 113
Chapter 3: The Ventral Striatum's Role in Learning from Gains and Losses 114
3.1 Introduction 114
3.2 Methods 117
3.2.1 Subjects 117
3.2.2 Surgery 117
3.2.3 Lesion assessment 118
3.2.4 Task 118
3.2.5 Images & eye tracking 122
3.2.6 Reinforcement learning models 123
3.2.2 ANOVA models 125
3.2 Results 126
3.2.1 Choice Behavior 128
3.2.2 Reinforcement Learning Models 135
3.2.3 Aborted Trials & Reaction Times 138
3.4 Discussion 141
3.4.1 Conclusion 150
3.5 Supplemental data 151
Chapter 4: The Amygdala's Role in Learning from Gains and Losses 157
4.1 Introduction 157
4.2 Methods 160
4.2.1 Subjects 160
4.2.2 Surgery 160
4.2.3 Lesion assessment 160
4.2.4 Task 161
4.2.5 Image & Eye Tracking 161
4.2.6 ANOVA Models 161
4.3 Results 161
4.3.1 Choice Behavior 164
4.3.2 Aborted trials & reaction times 174
4.4 Discussion 176
4.4.1 Conclusion 181
Chapter 5: What behavior can tell us about reinforcement learning 183
5.1 Introduction 183
5.2 Methods 186
5.2.1 Subjects 186
5.2.2 Task 187
5.2.3 Reinforcement learning models 188
5.2.4 ANOVA models 189
5.3 Results 189
5.3.1 Choice Behavior 189

5.3.2 Cash-out 212
5.3.3 Reinforcement learning model 217
5.4 Discussion 237
5.4.1 Conclusion 241
5.5 Supplemental Material 242
5.5.1 Reinforcement learning models 242
Chapter 6: Conclusions 246
6.1 Discussion 246
6.1.1 Conclusions 251
Bibliography 256

Chapter 1: Introduction

1.1 Reinforcement learning

Reinforcement learning (RL) is the behavioral process of learning to associate rewards and or punishments with actions or stimuli in particular states. States are comprised of a number of internal and external variables. Internal variables refer to things like drives; at its most reduced form, these variables stem from evolutionary important events for survival and reproduction (food, water, sex). This is in contrast to external variables, which are mainly comprised of a number of important elements in an organism's current environment. Both internal and external variables affect behavior, but the former does not have to be learned, while the latter can only be learned. That is to say, animals do not have to learn to be hungry, but animals do have to learn which environments provide food to be able to eat.

One of the most successful RL theories, and the one much of this introduction will focus on, is the temporal-difference reinforcement learning theory (TD). This theory suggests that phasic dopamine activity codes a reward prediction error (RPE), which is then used by striatal circuits to learn actions that maximize reward and minimize punishment (1, 2). RPEs act as the update rule for many RL models, and work in the following ways. When rewards and or punishments are

perfectly predicted the RPE is zero and no learning occurs. When outcomes are better than predicted/expected, the RPE is positive, and when outcomes are worse than predicted/expected, the RPE is negative. Thus, RPEs are defined as the difference between the value of the outcome/consequence (reinforcement or punishment) that is received, and the value of the outcome/consequence that is expected. It should be noted that the term outcome/consequence was used intentionally. The term reward (in lieu of outcome/consequence) is used almost exclusively in the literature -- its even the first word in the update term 'reward prediction error' -- but despite this fact most RL theories assume that this update rule is also true for aversive/punishing events (despite much less evidence).

1.1.1 RL models

Much of the support for this RPE theory comes from the formulation of RL models, specifically the Rescorla-Wagner (RW) algorithm (3). This learning algorithm was originally used to account for associative strength between a cue and reward. The term associative is important in this context because it describes a particular type of conditioning (Pavlovian/classical). It will become evident why this distinction is important. None the less, this algorithm was adopted and extended with a time component. This spawned temporal-difference (TD) RL algorithms, which compute prediction errors as the difference between the true

value of the reward, and the current state value of the reward. Thus, an update equation for a two-armed bandit choice task can be written as:

(1) $$v_i(k+1) = v_i(k) + p\big(R - v_i(k)\big)$$

The variable v_i is the value estimate for option i, R is the reward feedback for the current choice for trial k, and p is the learning rate parameter. This is only one way to write the update equation for this example task. For example, one could assume that there are two learning rates, one for each type of feedback. The TD model easily adapts to this idea by adding another learning rate parameter.

(2) $$v_i(k+1) = v_i(k) + p_f\big(R - v_i(k)\big)$$

In this case the only change to the equation above is addition of f, which indexes a separate learning rate for whether the current choice was rewarded ($R = 1$) or not ($R = 0$). In either case the larger p is, the faster values are updated. Regardless of the update equation, in choice tasks, values are converted into choice probabilities. This is done through a logistic function, which generates probabilities of choosing each option.

(3) $$d_1(k) = (1 + e^{\beta(v_2(k) - v_1(k))})^{-1}, d_2(k) = 1 - d_1(k)$$

The β parameter is the inverse temperature, which controls choice consistency. Specifically, higher values for this parameter indicate that the higher valued stimulus or action was chosen more often.

1.1.2 History of RL & dopamine

When it was discovered that dopamine is correlated with RPEs, this monopolized what the field believed dopamine's function to be. This monopolized view of dopamine's function led to a logical leap -- if dopamine is the 'value' neurotransmitter, then brain areas that receive large dopamine projections are well suited to compute the necessary computations needed for learning. Subsequent research provided evidence for this view. Specifically, it was found that midbrain dopamine neurons that project to the striatum, and provide reward prediction error (RPE) signals, increase their firing rates when rewards are unexpectedly delivered and decrease their firing rate when rewards are unexpectedly omitted (2, 5). In addition to these recording studies, fMRI studies have shown that the bold signal in the striatum correlate with RPEs (6, 7). These results led to the conclusion that the striatum underlines learning to select rewarding options.

The RPE theory extends to learning to minimize or avoid aversive events, which has led the field to assuming the striatum underlies this process as well. The major evidence for this being true comes from the finding that aversive outcomes cause midbrain dopamine neurons to pause their firing (8, 9). There is substantially less evidence for the aversive side of RL. However, it is crucial to

understanding the full spectrum of RL. So we will explore this in greater detail when appropriate.

Although this theory of dopamine has provided insight into some learning data, there are caveats to these results that are often overlooked. First, the type of learning the striatum is thought to underlie more closely resembles operant conditioning, and much of the evidence cited above comes from appetite Pavlovian conditioning (the importance of this distinction will become clear). Furthermore, when the striatum has been lesioned, operant conditioning has been impaired in some tasks (10, 11), but not others (12-14). Furthermore, often when these studies have found deficits, these deficits were the result of changes in the choice consistency parameter β, and not due to the learning rate parameter p. The choice consistency parameter is thought to be more of a function of motivation and not learning ability, thus even in these cases when deficits were found, the deficits were not consistent with the RPE theory.

The second caveat deals with the method used for most of the evidence above. Most of the evidence listed above comes from dopamine correlates in very similar tasks, in which there is no or relatively weak behavior. This is problematic because without behavior all we can say is that dopamine is correlated with parameters in the task. The purpose of learning is to acquire adaptive behaviors in new environments, so without behavioral changes, it is difficult to know what

dopamine is correlated with. This point is exacerbated by the fact that the more causal literature, like the results of the lesion studies listed above, are inconsistent with the RPE theory. A second point to this caveat is that certain studies have found dopamine responses to fully predicted rewards (15, 16). This is problematic for the RPE hypothesis because if rewards are fully predicted, there is no error and the RPE should be zero.

This level of inconsistency in the literature presents problems for current RL theories. We can further complicate the conclusions drawn from the RL literature by including results that have implicated other brain areas. For example single neuron studies in macaques have shown that the dorsolateral prefrontal cortex, as well as the anterior cingulate cortex, encode both losses and gains in a competitive game in which conditioned reinforcers could be gained and lost (17). In other work, the medial orbitofrontal cortex was found to encode gains and avoidance of losses, both of which have positive value (18). This study also found that appetitive RPEs in reward trials (i.e. increases with unexpected rewards) correlated with the extent of activation in the ventral striatum (VS), whereas RPEs in aversive trials (i.e. increases with unexpected punishments) correlated with activation in the insula, consistent with other work (19). These inconsistences have led to the idea that the brain is a redundant system, in which multiple brain areas are coding the same variables in a parallel fashion. I do not completely reject this notion -- this is

certainly possible -- but the proper work has not been done to make such a conclusion.

This is not unique to the striatum, many other brain areas follow this trend. For example, the amygdala is thought to underlie the formation of Pavlovian associations. However, animals with lesions to the amygdala can still learn to approach a food cup to obtain food during the presentation of a CS+ (20, 21). There are also results for the amygdala that mirror the results of the striatum mentioned above. Some instrumental conditioning tasks find deficits in animals with amygdala lesions (10, 22, 23), while other similar studies find no deficits in animals with amygdala lesions (24, 25). Again, the idea of parallel processing does not seem to fit in these cases. I submit that there is a more plausible reason for these conflicting results, and it has to do with differences in the tasks that find deficits versus the ones that do not. Simply put, certain brain areas are important for certain components that make up tasks, thus there are deficits when said task has those components, and no deficits when a task does not have those components. For example despite the fact that the VS is thought to underlie learned values, when the magnitude of rewards were tested against the timing of rewards, it was found that the VS was important for the timing and not the magnitude (26). As I will show throughout this chapter several factors contribute to the value of rewards (timing being one such factor), thus if these factors are not

controlled for it is easy to misattribute effects. Developing tasks with better behavioral control allows us to define more explicit hypotheses and draw stronger conclusions. The present review of the literature is concerned with understanding the process of RL.

1.2 Behavior

1.2.1 Learning

As we stated earlier, one of the most successful RL theories is the temporal-difference reinforcement learning theory (TD), which suggests that midbrain dopamine codes the temporal difference error from RL (27), which is then used to learn actions that maximize reward and minimize punishment (1, 2). In essence RL theories assume behaving organisms are optimal agents in the computer science/optimal control sense. To understand and make proper conclusions about the RL literature, we need to understand and clarify what it means to learn, and the behavioral components that make up the RL process. From a behavioral perspective there are a couple of things that are problematic for this theory. First, what does it mean to learn something, and how do we measure learning? Is it the case that once an organism learns something their behavior always displays it (ie they are optimal agents)? The answer is a resounding NO! There is considerable

evidence that organisms do not simply behave based on what they know, but instead behave based on internal and external motivational processes. It will become clear throughout this section that the distinction between learning and conditioning is an important one. Behavior is often controlled by the latter (28, 29). In essence theories of optimality do not properly account for motivation.

Formally there are at least two components to learning: acquisition of the stimulus/behavior-consequence relationship, and the maintenance of that relationship. In traditional learning theories the former is concerned with learning (there is a cap on how fast an organism can learn an association), while the latter has been assigned to motivation. Specifically, once an association is learned an organism can exploit it as much or as little as it wants depending on its motivational level. It is not trivial to dissociate these two components, and it will become clear later, that without dissociating these one can attribute behavioral results to the wrong process. In most cases an experimenter must judge learning from the behavior of the organism, which can be a mix of motivation and learning processes. However, this behavior is not fixed. Many things can alter this behavior, such as reward rate, reward schedule, and reward magnitude (30). The fact that environmental contingencies have such a profound and stable effect (even across species) on behavior suggests, at the very least, that behavior is not simply a display of what one knows. This means that an experimenter needs to design the

proper experiments to get at their question (ie most studies suggest they are studying learning, but they are not studying learning in isolation). To know what aspect of behavior one is studying, one needs to systemically manipulate environmental contingences. This is a prerequisite before one can make any conclusions about the behavior they are viewing (as we will show throughout this review this is one of the main reasons for the large discrepancy in results in the RL literature). For these reasons, behaviorists have long thought the term learning is misleading (28, 31, 32). This is easy to imagine when one considers the work that has shown that increases in associative strength on early trials do not translate into performance (33). This often leads to an abrupt onset of conditioned responding (34), starting much higher than trial and error learning would predict. All one truly knows is that under 'these' environmental contingencies, this is the performance that was witnessed. Instead of learning, behaviorists used terms like stamping in, performance, and conditioning. The former term speaks to ability, while the latter terms makes no such assumption. For example consider pigeon autoshaping data from (35). In this study they showed a well-known behavioral phenomena, increasing the inter-trial interval (ITI) speeds acquisition and promotes higher levels of conditioned responding. Is this result because of learning or motivation? To better understand the importance of this distinction, we first must explore the two major types of conditioning that make up most of the literature on learning.

1.2.2 Consequences

Before we explore the two types of conditioning, we must first define/explain a few aspects of behavior. By definition a reinforcer is any event that strengthens behavior, while a punisher is any event that weakens behavior. The positive and negative terms used before the words reinforcement and punishment just indicate if something was added (positive), or if something was removed (negative). Thus, a positive reinforcer is when something is added that increases the probability of some behavior, we often calls these things rewards. While a negative reinforcer involves the removal of an aversive event, the removal of this aversive event is reinforcing, as it strengthens the behavior that lead to this removal. For example, bringing an umbrella when there is a good chance of rain. In the past one has been reinforced by staying dry when it rains by bringing an umbrella, thus the future behavior of bringing an umbrella when rain is call for has been reinforced. Punishers are just the opposite. A positive punisher involves the addition of something, every time a dog barks, the dog gets shocked by their collar. In this procedure barking will soon be reduced. A negative punisher involves the removal of something someone finds reinforcing, the prison system is based on this one. When one breaks certain laws their freedom is taken away, in this case we assume freedom to be reinforcing.

There are two important concepts about these definitions that is crucial to understanding behavior. For the sake of simplicity I will just discuss these concepts for the reinforcers side, but the opposite appears to be true for the punishment side. The first concept is that these terms are discussed in relation to behavior. If a reinforcer does not increase the probability of the desired behavior it is not a reinforcer (by definition). It is easy to imagine why this is the case, the same rewards and their corresponding magnitudes are not universally reinforcing. In essence, all organisms do not find the same things reinforcing. Even in the same organism the value of rewards change based on their internal and environmental state. The former is associated with variables such as deprivation and satiation, these variables are well understood and typically controlled for in an appropriate manner. The latter, however, is more often overlooked, so in this section we are mainly concerned with how the environment affects reinforcement value. The effort side of how the task environments affect the value of rewards is well understood and has been extensively studied in delayed discounting and progressive fixed ratio lever pressing studies. These studies show that delay to rewards and the effort required change the value of the reward (we will explore this literature in more detail in the dopamine and effort section). But this is not exactly true because the reward has the same value in these cases, instead these rewards are just less reinforcing with time delays and increased effort cost. This

difference in definitions might seem trivial, but this difference is critical when discussing behavior because as we have just seen they mean different things (the same reward can have different reinforcing value depending on the contingency).

To make this concrete let's consider one more behavioral finding that is often overlooked in the RL literature due to task design. This finding has been called behavioral contrast and it describes a process that refers to the finding that rate of responding during a constant schedule of a multiple schedule task, may vary inversely with the reinforcement rate of the other schedule (36-38). For example, imagine a task with two different block types, A and B. To start suppose the reward rate for both block types is the same A (VI 1-min) and B (VI 1-min). After some training on these schedules and steady state behavior is reached. Following this, if one schedule is then changed, A (VI 1-min) and B (VI 3-min). It is normal to find increased responding in the A block type. This has implications about how motivation is related to performance, specifically the performance seen in one task is not necessarily the best the organism can do, it is simply the rate of behavior the organism is motivated to perform under current environmental conditions. The increased responding seen in A blocks is thought to be due to an increase in motivation for A blocks. This finding also speaks to another fact that it is easy to overlook, animals seem to be sensitive to the reward rate of the whole task

environment, and not just the block they are in. Similar to this finding, punishment has been found to facilitate responding to unpunished behavior (39, 40).

The second concept is what skinner called *probability of response* (28). The idea is that no behavior exist only in two states, one in which it always occurs, or one in which it never occurs. Instead behavior lives on a continuum of probability, this concept is essential to understanding behavior and the consequences that maintain behavior. Notice that the definitions for reinforcement and punishment are stated in this this view, they increase or decrease the probability of a response. We cannot readily predict when an organism will eat because there are many other factors that affect this behavior. Factors such as, how hungry the organism is, how costly it is for the organism to get food at the present time, how reinforcing the food is to the organism, and even more practically if an organism is busy doing something else. Instead all we can predict is that the probability that an organism will eat continues to increases from the time of their last meal to their next meal.

1.2.3 Conditioning

The two forms of conditioning that make up the much of the RL literature are Pavlovian conditioning and operant conditioning. RL theories were inspired by these behavioral forms of conditioning, but based on some of the conclusions drawn in the RL literature, it is clear that important concepts from the behavioral

literature have been lost in translation. Thus, the second problem in the TDRL theory has to do with the type of conditioning being investigated. Behaviorist have long classified these two forms of conditioning as distinct processes (41). This distinction is not as clear in the RL literature (beyond their definitions). In fact the definition and design of RL theory is most closely related to rules applying to instrumental conditioning (41), however most of the evidence supporting TDRL (RPEs) come from classical conditioning experiments (41, 42). When we discuss the overlap of these two learning systems it will become clear why this distinction is important.

In classical conditioning a previously neutral stimulus becomes a conditioned stimulus (CS) after being repeatedly paired with some biologically relevant unconditioned stimulus (US), such as food. After repeated pairings, organisms began to respond to the CS as if it was the US. In the famous Pavlovain experiment, Pavlov found that repeatedly pairing a bell, which was originally a neutral stimulus, with food (the US which dogs salivate in response to), lead dogs to salivate to the sound of the bell (43). In essence the bell began to predict the availability of food, thus the natural reflex of salivation that occurred at the sight of food now occurred when the bell was rang, which means at some level the dogs learned that the bell predicts food. This type of conditioning is just a predictive relationship, that is to say, the outcome (whether or not the animal gets food) is not

contingent on the animal's behavior. This feature of Pavlovian conditioning, which is often neglected in the RL literature, has to do with the topography of the response -- sometimes referred to as AutoShaping or sign tracking (37). Historically, the response that occurs as a result of classical conditioning (in this case salivation) is involuntary, dogs do not choose to salivate at the sight of food. This is something that occurs naturally. Thus, UR's have often been thought of as reflexes, autonomic, or preparatory behavior mainly concerned with the internal physiology of an organism (28). It is of great advantage for organisms to start to salivate before the food is in its mouth, but the animal does not voluntarily control this salivation. In fact some research has shown that there are two main classes of CRs (44). The idea is that the same conditioning process underlie the generation of preparatory and consummatory CRs (45). Preparatory CRs and thought to represent motivational emotive properties of the US, while consummatory CRs are thought to represent the sensory properties of the US. Preparatory responses in this model are not specific to the US, but to the activation of the motivational system. In the example above about food, the preparatory CR would be approach to food, while the consummatory CR would be to salivate as the organism gets closer to the food.

Sex is another and perhaps more powerful example of how this preparatory behavior is a great advantage for an animal's evolutionary fitness. For example, if

the precursor behaviors that lead to sex do not lead to arousal, the act of having sex (and thus reproducing) becomes more difficult. These precursor behaviors are likely learned, but there are some that seem to be innate. Research has shown that even though males are consciously unaware of when women around them are ovulating, their internal system is aware from olfactory cues (46). Furthermore, it has been shown that men rate ovulating women as being more attractive, when compared to their non-ovulating counterparts (47, 48). It also been found that testosterone levels are raised in men around ovulating women (49), and this chemical change leads to men performing different behaviors. Similar results have been found in women, it has been shown the type of men that women are attracted to changes when they are ovulating versus when they are not (50). The important thing about these chemical changes is that males and females are consciously unaware of them, and these chemical changes can alter behavioral. This seems to be the hallmark of the Pavlovian conditioning system.

In contrast to classical conditioning, in operant conditioning, an animal's behavior is acquired and maintained by the consequences (reinforcement and punishment) that follow said behavior. Thus, in operant conditioning the consequences are contingent on the animal's behavior. These two forms of conditioning are separate processes, in operant conditioning a reinforcer makes a response more frequent, while in classical conditioning a reinforcer increases the

magnitude of the response and shortens the time elapsed between stimulus and response (28). Historically, operant conditioning is considered voluntary behavior -- the animal is aware of the behavior and this behavior is effortful (28, 30, 37). There are more differences between these forms of conditioning, and we will touch on them as they become appropriate to gain an understanding of the RL literature. For now, I want to emphasize the contrast between voluntary and involuntary -- the former being effortful, while the latter does not require effort. As we will soon show (in the dopamine and effort section), this effort component is quite important.

1.2.4 Pavlovian-instrumental transfer (PIT)

We can further complicate this distinction by admitting that these two types of conditioning complement one another. Skinner believed that "Since the environment changes from generation to generation, particularly the external rather that the internal, appropriate reflex responses cannot always develop as inherited mechanisms. Since nature cannot foresee, so to speak, that an object with a particular appearance will be edible, the evolutionary process can only provide a mechanism by which the individual will acquire responses to particular features of a given environment after they have been encountered. Where inherited behavior leaves off, the inherited modifiability of the process of conditioning takes over." (Skinner, 1953). Specifically, Pavlovian associations add motivational value to

operant contingencies (51). This process is called Pavlovian-instrumental transfer (PIT). This can be shown in experiments where a CS (say a light) is paired with an appetitive outcome. Animals who are trained to press a lever for the same appetitive outcome paired with the CS, respond more to the lever in the presence of the CS (51, 52). This is an important concept because numerous lines of evidence suggest it is this Pavlovian boost that the RL literature is most often seeing with their dopamine recordings.

Evidence suggest that this form of PIT is mediated partially by the amygdala, and partially by the nucleus accumbens. In one study rats were trained to associate a light-noise compound stimulus with water. Following this half of the rats received excitotoxic lesions of the basolateral amygdala. Next both groups received intra-accumbens amphetamine infusions of d-amphetamine and began the test phase. In the test phase two novel levers were available. Neither lever produced water, but one did produce the conditioned reinforcer of the light-noise compound. The authors found that the amphetamine infusions increased responding on the lever that produced the conditioned reinforcer, and no change in responding on the lever that had no consequence for both the sham-controls and lesion animals (53). Thus, the lesion animals responded like the control animals (intra-accumbens amphetamine infusions of d-amphetamine produced amplified responding).

This result should be viewed in contrast to a study by De Borthgrave et al. (13). In this study experimenters examined the effects of cytotoxic lesions of the nucleus accumbens in rats across two instrumental conditioning experiments. When experimenters compared rats with lesions of the nucleus accumbens to sham-controls, they found that instrumental responding of lever pressing and chain pulling for food reinforcers was mildly suppressed in the lesion animals. However, this reduction in responding was not due to lesion animals having trouble learning the instrumental contingency, but instead due to a reduction in motivation. In a second experiment, d-amphetamine was administered into both the sham-controls and lesion rats, the authors found that the normally increased responding found when d-amphetamine is administered was significantly reduced in lesion animals. These results suggest that the nucleus accumbens' role in instrumental conditioning is to provide excitatory motivational effects of appetitively conditioned Pavlovian signals, instead of holding the value that is attached to instrumental outcomes. The fact that infusions of d-amphetamine to the accumbens made rats with amygdala lesions respond like controls, but not rats with nucleus accumbens lesions provide further support to the hypothesis that the nucleus accumbens plays this excitatory motivational role.

This overlap can make it difficult to distinguish what conditioning system is in control of a specific behavior. Recall the contrast of voluntary versus

involuntary as being one of the hallmark signs to distinguish Pavlovian from operant conditioning, but this voluntary versus involuntary behavior is not always apparent. In most cases the only way one can tell is by testing the target behavior. For example the behavior of approach. On the surface it is hard to think of approach behavior as involuntary, the key to understanding this is in the fact that organisms do not seem to have to learn this behavior. It seems to be as automatic as the huger example given at the start of this chapter (one does not have to learn to be hungry). Consider the (54) study, in which chicks were trained to expect food from a specific food cup. He then constructed an arrangement where if the chicks approached the food cup, the food cup retracted at twice the chicks approach speed. If chicks ran away from the food cup, the food cup approached them at twice their speed. Thus, the chicks had to learn to run away from the food cup to get the reward. This is an abnormal behavior because most organisms seem to inherently approach rewards. So if the chicks could learn to run away from the food tray, this would mean the chicks are sensitive to the consequences of the contingency, which would suggest that this behavior is under operant control. However, if chicks continue to approach the food cup despite the consequence of the contingency, it suggest that this behavior is under Pavlovian control. In this case the chicks have just learned an association between the food cup and rewards,

but not how their behavior affects the outcome (consequences). This is exactly what (54) found, chicks continued to chase the food away.

This result was not surprising, approach behavior has long been considered to be a "investigatory reflex" (32, 43), but it was important to show what Pavlovian responding looked like without the operant aspect. Just as Pavlovian associations affect operant contingencies (PIT), the consequences (in this case food from tray) can affect Pavlovian behavior. This is easy to imagine, the more an organism approaches the food tray and gets food, the more times this approach behavior is reinforced. One way to investigate this is to use omission schedules, in omission schedules the CS is followed by the US except when the organism produces the CR. So in our example the US of food always follows the CS except when the animal produces the CR of approaching the food tray during the CS. This ensures that the CR is not being reinforced. These studies have revealed that approach behavior is almost entirely Pavlovian and based on the reinforcement rate of the CS-US relationship (55, 56).

1.3 Dopamine

Now that we have some behavioral context, we are better equipped to review the RL literature. As we stated earlier, when it was discovered that dopamine is sometimes correlated with RPEs, this completely monopolized what the field

believed dopamine's function to be. This is the case despite two important facts. First, different types of tasks seem to engage the dopamine system in different ways. Specifically, many studies have found that dopamine is not always correlated with the learning signal described above, but instead correlated with other features of the task (i.e. the task matters). For example, dopamine has been found to encode action values, without changes to reward value (57). In another study dopamine was found to encode the identity of the reward in a prediction error fashion (58). And yet another study found that RPEs did not track learning -- RPEs instead tracked when actions should be taken (59). This is an important point, because it is not the stance of this thesis that dopamine does not (at the very least) encode RPEs, when learning is necessary. The question really becomes, what is this information used for? Do RPEs provide the learning signal, or is this information used to determine where and how much effort an animal should put forth in a particular environment? Or both?

Before the discovery that dopamine is sometimes correlated with RPEs, dopamine was thought to be important for motivation, specifically this aspect of motivation referred to as vigor. Vigor is defined as the propensity to work harder, longer, and faster. Specifically, the vigor theory suggests that perceived opportunity cost determines how and how much effort should be distributed based on a cost/benefit analysis of the environment (60-63). This is a compelling theory

for dopamine's function because it provides a unifying function for dopamine that explains much of the seemingly contradictory literature that is produced under the scope of the RPE theory. For example in T-maze studies, dopamine signals have been found to ramp up as the animal gets closer to completing the maze, and thus the reward (64). This result provides problems for the RPE theory, but fits well for the opportunity cost/vigor theory. From an opportunity cost point of view as the rat gets closer to completing the maze, it becomes more valuable for the rat to complete the maze and receive its reward.

The idea of vigor is more closely linked to motivation than learning. This distinction should seem familiar from the section on learning, in which learning is linked to acquisition, while motivation is linked to performance. The difference between motivation and learning is not always apparent, because both motivation and learning have value signals. However, these forms of value differ. There are many factors that affect reinforcement value (65). Most often these factors are thought to affect an animal's motivation and strength of engagement in tasks. Factors, such as level of deprivation, amount of effort required to receive reinforcement, magnitude of reinforcement, richness of the reward environment, and schedule of reinforcement (to name a few). Consider how much work a poor person might do for 10 dollars, in contrast to that same person if they were rich. While, the latter is more fixed (10 dollars is 10 dollars) and concerned with simply

linking a stimulus or behavior to a particular consequence. Admittedly, the former can affect the latter to a point (think about how paying attention and how the amount of effort put forth can affect the amount of rewards one receives), thus the question becomes; using the literature can we divorce these two features?

This is a unique problem for operant conditioning, in particular operant learning paradigms, because the variable (reinforcement) that is thought to influence responding, depends on responding. Since in these paradigms reinforcement is contingent on response, the reinforcement rate is dependent on the response rate. There are ways to disentangle this reciprocal relationship, especially when designing the experiment, but they are seldom done in the RL literature. There are hints in the literature, but the evidence comes from a range of different experimental approaches, so we have to establish operational definitions to make proper conclusions. The RPE theory of dopamine implicitly implies that increasing/decreasing dopamine should increase/decrease the RPE signal and result in faster/slower learning rates. This theory predicts that manipulations to dopamine should affect the acquisition rate of learning, but not the maintenance phase of learning. In essence this theory suggests that increased dopamine should make one learn faster. This is a different statement than saying increased dopamine makes one more motivated to learn. As we stated earlier, separating these two phases is not trivial, because motivational variables affect both phases.

1.3.1 Dopamine & effort

One way to divorce motivational value from a learning signal is to remove learning from the task. The studies of the effects of dopamine on effort fulfill this requirement. The research on effort is usually conducted using two types of tasks. The first, measures performance on a progressive fixed ratio lever pressing task. In these tasks the number of lever presses required to receive a reward is increased over time. Consistently studies have found that dopamine depletion from the nucleus accumbens leads to deficits on higher fixed ratio requirements (66-69). There are two important facts about these studies. First, there were no deficits on an FR1 schedule, but as the schedules increased (FR4, FR16, FR64), so did the deficits. Second, these studies showed that the dopamine depletion did not affect primary food reinforcement. For free or low effort cost, the animals with dopamine depletion would consume as much as their control counterparts. Taken together these results suggest that food was still reinforcing to these animals, just not as reinforcing at the higher work requirements.

The second way research on effort is typically done is with the T-maze task. In T-maze tasks animals are given a choice between a small reward in one arm and a large reward in the other arm. The arm with the large reward has a barrier, the reward can only be obtained by climbing the barrier (a measure of effort). Both

lesions to the nucleus accumbens and the administration of haloperidol (antagonist of dopamine receptors) consistently lead to rats choosing the low effort/low reward arm more often than their control counterparts (70-72). Similar to T-maze tasks, delayed discounting tasks, in which animals are offered the choice between a small immediate reward or a larger delayed reward (where it is assumed that waiting longer for a reward is more effortful), lead to similar results. Animals given haloperidol or who have received lesions to the accumbens, choose the small immediate reward more often than their control counterparts (72-74).

There is also evidence that microinjections of d-amphetamine into the nucleus accumbens produces increased lever responding (75). In this study rats were trained to associate a light (conditioned reinforcer) with water. After the initial training water was no longer presented with the light. Rats were then presented with two novel levers, one of which produced the light. Rats that received microinjections of d-amphetamine into the nucleus accumbens were found to have selective, dose dependent, increases in responding to the lever that produced the light. In similar studies, d-amphetamine injections into the nucleus accumbens has also been found to enhance sexual arousal in male rats (76).

In another series of studies where rats were trained to lever press for food pellets and sucrose, it was found that rats with excitotoxic lesions to the nucleus accumbens did display impairment in lever press performance when compared to

sham-controls. However, these deficits were not due to lesion animals having problems learning the action-outcome contingency, but instead due to the lesioned rats not being as motivated as the sham-controls, and this decrease in motivation is what lead to the lever pressing deficits (12, 13).

Furthermore, injections of d-amphetamine into nucleus accumbens has been shown to invigorate a range of behaviors (77). The term invigorate is important, because no matter the behavior or task, animals perform the behaviors with more vigor. Specifically, animals perform these behaviors faster, this point, along with the data on effort, has led some to believe that dopamine provides general energizing/motivational effects. Some have described this effect as a gain/incentive amplification of learned responses (78, 79). This would mean the effects we see in regard to dopamine, are of motivational value and not learning. And indeed there have been models proposed by (41, 77) on how this would work. However, there is no learning in these effort studies, so next I will review some of the learning literature before we consider any models.

1.3.2 Dopamine & learning

Most of the RL literature does not study learning. Furthermore, it is not so clear that the small portion of studies that claim to study learning are studying learning. We will discuss some of these caveats later in this section. For now we

consider some of the learning research that the field believes support the RPE theory.

In a task where humans could earn or lose money, Pessiglione et al found that subject given L-dopa (which is a metabolic precursor of dopamine) earned more money in an instrumental conditioning two-armed bandit task, when compared to subjects given haloperidol (which is an antagonist of dopamine receptors) (19). This drug effect only occurred in the appetitive condition. The groups did not differ in the loss condition. The authors combined their drug approach with functional imaging and found that the bold response in the VS was enhanced in subjects given L-dopa, when compared to subjects given haloperidol. Another study looked at the effects of sulpiride (D2 antagonist) on learning from gains and losses (80). When compared to a placebo group, the authors found that the drug group had reduced performance when choosing rewarding options. Consistent with (19) this difference in performance between the groups was limited to the appetitive condition, the drug group showed no performance deficits when learning to avoid losses. Importantly the RL model that the authors fit indicated that this reduced performance was not due to the learning rate (acquisition), but due to choice consistency (maintenance), which is also consistent with (19).

In a study with non-human primates (within monkey) comparing the effects of L-dopa, haloperidol, and saline on a two-armed bandit reversal learning task, it

was found that L-dopa and haloperidol led to increased performance, when compared to saline (81). Importantly, this increase in performance due to both drugs was a result of the choice consistency/maintenance phase and not the learning/acquisition phase. None the less the fact that haloperidol lead to increased performance is difficult to reconcile with our current framework. In a more recent study some evidence was presented that helps explain the haloperidol effect found in the previous study. In a 5-choice serial reaction time task (which is a well-validated measure of attention and impulsivity), rats were given the choice between an easy or hard discrimination (82). Successful hard trials were reinforced with double the reward (sugar pellets). Consistent with work in humans, researchers found individual variations in the rat's willingness to work. The term 'workers', refers to rats who naturally chose the hard trials significantly more than other rats, called 'slackers'. This study found that when workers were given d-amphetamine sulfate, they tended to slack off. In contrast, when slackers were given amphetamine, they tended to work harder (83). This is a very interesting finding, and we will explore this more below.

1.3.3 Dopamine's role?

Taken together these results presented on effort and learning seem to clearly point to dopamine having a motivational function versus providing the learning

signal. So then how has the RPE theory persisted despite this somewhat counterintuitive evidence? In all the cases above (in effort and learning sections) dopamine was altered in some artificial way, and in the cases when dopamine was increased, it was the tonic (slow) level of dopamine that was increased. The fact that the RPE theory states that phasic (fast) dopamine is what drives learning seems to provide enough of a difference (for some people in the field) to allow this theory to stay in place and remain unaltered. Regardless of whether one finds this reasoning good enough, the research above presents major problems for RPE theories.

First, the effort data explicitly shows that motivational value can be altered, without affecting learning. This effect is further supported by the learning data. When RL models were fit to the learning data above it was found that the results on learning were due to the choice consistency parameter β, and not due to the learning rate p. As we stated earlier, the choice consistency parameter is thought to be more of a function of motivation and not learning ability. However, it is true, that it is unclear how well the models separate these two parameters, so this part of the evidence is not clear. What is perhaps more problematic is that in these learning studies, 1) only the appetitive side was affected, 2) L-dopa did not help participants learn better, and 3) sulpiride did not make participants learn worse from aversive outcomes. Thus, the second problem for RPE theories, is the

inconsistent/lack of effect on the aversive side of learning. The third problem has to do with the ramping effects of dopamine seen during T-maze tasks. This is not unique to T-maze tasks. Similar results were found in a task where rats lever pressed for cocaine. Nucleus accumbens dopamine rapidly ramps as the rats gets closer to pressing the lever for access to cocaine (84).

1.3.4 Dopamine theory

It is unclear if we should consider tonic and phasic dopamine as distinct processes that have different functions. The different dynamics of dopamine are not well understood. However, there are some theories that coincide with behavioral data that suggest a role for both. One such theory was presented by (77). The cornerstone of this theory starts with agreeing that tonic dopamine levels are a measure of an animal's current motivational profile. This idea is appealing for several reasons. Perhaps the biggest reason is, this provides a mechanism for which an animal's internal state can differentially select how it should behave. For example, a food deprived animal will likely have higher levels of tonic dopamine in a food learning task, when compared to a sated counterpart. Most animal researchers have experienced this in one form or another. The animal performing a task at the beginning of a session seldom behaves the same at the end of the session. This is also consistent with the effort and learning literature stated above.

However, this is only part of the equation for adaptive behavior. Animals still needs to read their environment and decide based on their needs how much energy and effort to expend. To do this, animals must have a way to track the reward rate of their current environment, this is where RPEs are important. RPEs provide an estimate of reward rate for the current environment. Holding tonic levels of dopamine constant, in rich reward environments, animals should expend more energy. This is in contrast to lean reward environments, where animals should expend less energy. Importantly, this phasic dopamine RPE signal affects tonic levels of dopamine, raising it in rich reward environments, and lowering it in lean reward environments (this sounds like the cost benefit function described above). And there is evidence that this is the signal dopamine is conveying. Consider the study from (62). In this study the authors measured dopamine release in the nucleus accumbens across multiple time scales using voltammetry. They found that minute-by-minute dopamine co-varied with reward rate and vigor, and this change in dopamine immediately altered willingness to work.

This framework is appealing because not only is there evidence supporting dopamine in this role -- this is how animals behave. It is well known that response rates are higher in rich reward environments, when rewards have higher values and lower work or time requirements , and response rates are lower in lean reward environments where rewards have larger work, or time requirements (85). It has

been found that the optimal solution in standard operant reinforcement schedules that have rich reward environments is to respond faster for all possible actions. That is to say that optimal latency of all actions is inversely proportional to the average reward rate (77). The idea is that in rich reward environments, animals should act more quickly regardless of the chosen behavior because rich reward environments have higher opportunity cost. Reward are readily available, so even if the animal chooses to do a behavior that is not rewarded, like grooming, they should do it quickly, so they can return to the behaviors that readily produce rewards.

1.3.5 Optogenetic activation of VTA dopamine

This provides a framework that helps explain several behavioral findings. Findings such as the speed accuracy trade off and explore/exploit trade off (to name a few), but most importantly it accounts for the "stronger" RPE evidence. The stronger evidence typically optogenetic activation and inactivation in two types of tasks. The tasks are either conditioned place preference and conditioned place aversion, or Pavlovian approach tasks. We will discuss the latter. The (86) study is a result that is considered to be strong evidence for the RPE theory. It is not clear why this is the case because the results they get are consistent with the vigor theory and exactly what the research on effort and PIT suggest. In this

experiment they found that optogenetic activation of VTA dopamine neurons given with reward caused cue-elicited reward seeking behavior in an associative blocking and extinction paradigm (will just discuss the former). In blocking paradigms the association between a cue and reward is prevented (blocked) if another cue presented at the same time already predicts the reward. In the first training phase both groups of rats were trained to respond to an auditory cue (A) for sucrose. In the next training phase, the compound training phase, both groups of rats were trained to a compound auditory (A) and visual cue (L), and the identical reward, sucrose was delivered. For the blocking group the same auditory cue that was used in the first training phase (A) was used in the compound phase. For the control group a different auditory cue (A') was used in the compound phase. The only difference between the groups was the predictability of sucrose. Since the blocking group had the same auditory cue in the compound phase (previously trained), they should expect sucrose, while the control group should be surprised by sucrose. The idea is that since the blocking group had already been trained with the auditory cue, they would not learn that the visual cue predicts reward. The opposite is true for the control group, since their auditory cue was changed to a novel cue, they would learn that the visual cue predicts reward. This is indeed what they find, conditioned responding to the visual cue was reduced in the

blocking group as compared with the control group. This supports the RPE idea because learning should only occur when rewards are unpredicted.

Following this in a new set of rats they had three groups and trained them all using the procedure described above for the blocking group. One group received optogenetic activation of VTA dopamine neurons at the time of the reward delivery during compound training trials, and the other two groups were control groups. The idea is that the group that received this activation should learn about the light because the extra stimulation from the optogenetic activation will drive an RPE. And they found that this group responded more strongly to the visual cue on the first test trial when compared to both control groups (they approach the food cup more). Two things to note about this experiment. One, this task was not a traditional Pavlovian approach task because in this version the US delivery was contingent on the rats being in the reward port (without additional information and test it is unclear if and how this might affect behavior in this task). Two, the effect they found on the first trial was dramatically reduced by the third trial.

On the surface this effect seems to provide good evidence for the RPE theory, but let's consider all the things we have discussed thus far and see if we can understand what is occurring here. To do this we first have to consider another sensory conditioning paradigm (87). In this study two separate pairs of cues are paired with one another without the delivery of reward. So for group one, A + B,

and for group two, C + D. After this preconditioning phase if B is paired with rewards and D paired with nothing, and then look at dopamine responses to A and C. What you find is that there are higher dopamine responses to A compared to C, signifying that A does in fact have some value. This effect that the authors found, is a kin to a well-known effect in the associative learning literature, this effect is known as retrospective revaluation (88, 89). This effect describes a phenomenon that was first discovered in an overshadowing experiment (88). Overshadowing refers to the finding that conditioned responding is reduced if the CS is reinforced in compound compared to if it were reinforced alone. If A + L is trained as a compound stimulus and reinforced as such, conditioned responding to L (overshadowed cue in this case) is suppressed when compared to conditioned responding to A. However, if A is presented alone without the US (extinction), responding to L increases. The value of L is increased due the value of A being decreased. This is important because since this was discovered it has provided problems for the RW model. Recall that the TD model is an extension of this learning theory, thus it cannot explain this phenomenon. In the (87) experiment A was never paired with rewards, thus dopamine is not able to go back in time to make this association with A and B. This invalidates the assumption that a prediction error is needed for learning, which is the assumption that (86) and the RPE theory is based on. The rationale for this in the RL literature has been that

this is a different type of learning -- this is the result of model-based RL, and not the trial and error leaning encompassed by model-free learning. Importantly, dopamine is thought to be responsible for the latter and not the former.

Let's assume the effect presented in (87) is not convincing, because this experiment did not use blocking. Let's consider another study (90). In this study authors merged the methods of the previous two papers. They used the same sensory preconditioning procedure described above but added blocking and optogenetic activation of VTA dopamine neurons. On the first day of training two groups of rats were conditioned to pair two novel cues, A and X (sensory preconditioning). Following this initial pairing rats were trained to pair AC and AD with X (blocking component). In addition to the pairing of these cues, rats were also conditioned on the same procedure to associate a compound stimulus EF with X (compound stimulus control conditioning group). The idea is that C and D should be blocked because A already predicts X. E was not presented without F, so F should not be blocked. So far this procedure is similar to (86), only differing in two ways. First, X has not been paired with primary reinforcement (US) yet. Second, when A was presented with C (AC) one group (ChR2) of rats received optogenetic activation of VTA dopamine neurons when X was presented following AC. The second group (eYFP) of rats were the activation control group. The authors found no differences in food cup approach behavior between the two

groups for any of the cues (AC, AD, EF, A, or X). In fact, across all rats there was minimum approach behavior across all the cues.

Following the blocking procedure and test phase, rats were separately conditioned to pair X with sucrose pellets. Consistent with conditioning using primary reinforcement, after this procedure both groups of rats approach the food cup more in the presence of X. Importantly, both groups responded at similar rates to X. Next, the authors tested to see how rats would respond to the cues C, D, and F. The authors found that both groups learned that F predicts X. In addition, the authors found differences across the groups in approach behavior to C. Specifically ChR2 rats responded more to C when compared to eYFP rats. In addition ChR2 rats responded more to C when compared to D. So the activation of VTA dopamine during the preconditioning phase reversed the normally seen blocking effect that was present with D. Interestingly, the "value" that C gains in ChR2 rats is related to the primary reinforcer (sucrose pellets) that was paired with X. The authors tested this by using a devaluation procedure. After devaluing the sucrose pellets by pairing it with lithium chloride injections, which makes the rats sick, the rats approach behavior to the food cup is reduced. This is an important finding because along with the sensory preconditioning effect, which should not be able to happen if we subscribe to the RPE theory, it shows that animals are able to

do some sort of backpropagation linking C to the primary reinforcer that X would later come to predict.

The results from (90) were replicated and extended in (91). In this study the authors show that we should be careful about the assumptions we make in regards to the type of values the cues in the above study have. Sticking with the cues from (90), X and C, even though it was found that C elicits approach behavior and is sensitive to the devaluation of the sucrose, the authors in the follow up study found that C does not have value (not cached value anyway). In essence, the value that C and X have differ. X (the cue that was paired with primary reinforcement) has this cached-value and act as a conditioned reinforcer, while C has just picked up this valueless association that is predicting a specific outcome. One way to asses if a cue has predictive value is to see if animals will work to produce it. The authors found that rats will lever press to gain access to X, but they will not lever press to gain access to C. The authors also show that optogenetic activation VTA dopamine neurons facilitated cue learning without endowing cues with value. The authors concluded artificial induction of dopamine transient seem to supports valueless associative learning, rather than cached-value learning. This result provides additional support for the role of dopamine in attention and motivation and not learning value.

These four optogenetic studies taken together along with some of the other literature presented suggest a very different conclusion than the one drawn from (86), and thus the dopamine RPE theory of learning as a whole. As we stated earlier, sensory preconditioning and retrospective revaluation provide problems for the RPE theory of learning. In addition to this it appears that the RL literature is working off a false conclusion about what occurs in blocking experiments. In blocking experiments, it is not the case that the blocked group does not respond to the blocked cue, the responding is just suppressed when compared to the unblocked cue. Which suggests that the blocked group did in fact learn something. Furthermore, the optogenetic activation of dopamine does not just 'unblock'. Responding is heightened when compared to unblocked controls. There is a boost in responding, which speaks to something other than learning. Which is consistent with modern views on blocking. In blocking experiments, it has long been known that the second cue is not actually blocked, instead the suppression of responding is due to a group of phenomena call cue-interactions. Whenever a compound cue is presented, the value of one cue depends on the value of the other cue, a phenomena called *relative validity effect* (3). The hallmark experiment was done by (92) and showed that if a compound stimulus CD is followed by the US, the value is split between the cues. If you present another compound stimulus DE, but the US does not follow this compound stimulus, responses to D is much more suppressed than

if both compounds were reinforced 50% of the time. In essence in CD trials, D competes with the always reinforced C, which leads to D being less valued. This also explains the retrospective revaluation phenomena. Recall that the RW algorithm has no mechanism that can explain either of these findings.

The Pearce-Hall (PH) algorithm was developed as an extension of the RW model that could account for changes in learning rates under different conditions due to changes in attention (93). This model was developed to account from data coming from unblocking experiments. The PH model has an extra parameter for the associability of the CS, this new parameter is modifiable with experience. Recall, that the RW model predicts that learning about B in an AB blocking experiment is prevented because A already predicts the US. We have already suggested that B is not in fact blocked, but let's consider one more example to solidify this point. In the AB blocking experiment, what happens if one was to lower the magnitude of the US from conditioning trials with just A (higher magnitude) to conditioning trials with AB (lower magnitude). For the blocking theory to hold, what the RW model suggest should happen in this case is that there should be a negative prediction error, which should lead to lower associative strength. Instead what you find is that B gains strength. The surprise in the different magnitude of the US lead animals to associate it with the other cue (B). This is what the associability parameter in the PH model accounts for, the model

assumes that individual prediction errors influence the CS associability, while the aggregate prediction error influence the degree of learning that is available to all present CSs (34).

1.4 Studying learning systems

1.4.1 Response rates

Regardless of the theory or model, we should be hesitant about the assumptions we make about rates approach behavior and value. As (90, 91) have shown one can get increased approach behavior, without having changes in cue value, measure by if an animal will work (lever press) to produce a cue. This finding illustrates two points. One, the topography of the behavior matters. In general approach behavior (especially in early trials) is thought to be a Pavlovian behavior, while lever pressing is thought to be an operant behavior. Two, this suggest we should have multiple measurements of value before any substantial conclusions can be drawn.

Behavior momentum theory suggest that response rates is not a good measure of associative strength (94, 95), this theory suggest that resistance to change is a better measure of associative strength. This theory has been supported by work that has shown that increases in associative strength on early trials do not translate into performance (33). Furthermore, examination of individual learning

curves often have a more abrupt onset of conditioned responding (34). (96) Have shown that these abrupt changes in acquisition are often missed when data is averaged across subjects. Consider the results from (56), which used the omission procedure described above to show conditioned magazine approach behavior for the US of food reinforced at different rates (100%, 50%, 25%, 12.5%, pre-CS). In this procedure each session contained 64 trials, 16 trials of each of the four CSs were presented randomly and approach behavior was recorded. Despite how often the CS was reinforced, before steady state behavior was achieved, responding was similarly high in early sessions. This is potentially the result of the threshold requirement (perhaps this is the threshold to activate preparatory system) that lead to the more abrupt onset of conditioned responding mentioned above. I point this out because often in the optogenetic studies, there is not very much behavior, they only track responding for a couple of trials, and there has been constant debate as to what these early trials signify (33, 34).

Furthermore it has been shown that the behavioral requirements of the organism under study is crucial for any assessment of learning. Much like the previous distinction between approach behavior and lever pressing behavior, it is important to know if the consequences are Pavlovian (does not matter what the organism does, consequences are delivered regardless), or operant (the organism

must perform a particular action to receive consequences). Consider the results from (97), where they show different rates of approach response behavior for a stimulus-contingent group compared to a stimulus-response-contingent group. In this study for the stimulus-contingent group, rewards were delivered at the end of CS regardless of behavior (true Pavlovian). For the stimulus-response-contingent group, rewards were only delivered if the rats made at least one approach response during the CS (operant). Following this testing procedure (first 36 sessions) both groups were tested on an omission schedule (last 24 sessions), in which rewards were only delivered if no approach response was made during the CS. Approach behavior was recorded for four separate time bins: (a) the inter-trial interval (ITI), (b) the 20 sec pre-CS period, (c) the 8 sec CS period, and (d) the 20 sec post-CS period. In general responding was higher for the stimulus-response-contingent group across all time bins except the post-CS period. For our purposes the time bin to pay attention to is (c) the 8 sec CS period, the responding difference was the highest during this time. This was driven by heightened responses during the first 36 session and slower extinction in the last 24 sessions for the stimulus-response-contingent group. We will come back to this point below, but for now it is clear that the behavior requirement contingency affects responding behavior. The difference between the groups in the last 24 sessions (omission/extinction) is what behavior momentum theory suggest is a better measure of associative strength

(resistance to change). Notice that the stimulus-contingent group responds at a steady rate across the sessions, until the omission sessions, in which responding stops. This is in contrast to the stimulus-response-contingent group whose behavior never reached the low levels of the stimulus-contingent group.

1.4.2 Lesions & learning

The number of different methods used to study RL, even without accounting for task differences, makes it difficult to make any clear conclusions. A better method is to perhaps break the literature down in a way, where at least conclusions about certain methods can be drawn. This thesis is primarily concerned with understanding what lesions can tell us about the function and role the ventral striatum (VS) and amygdala in RL. In this section we will review some of the literature on learning and see if the proposed model can help explain some of the seemingly conflicting results.

In one study three groups of non-human primates were compared on a two-armed bandit visual discrimination reversal learning task. One group received excitotoxic VS lesions, another group received excitotoxic amygdala lesions, and the final group consisted of un-operated controls. These groups were compared across 4 reward schedules (100%/0%, 80%/20%, 70%/30%, 60%/40%), it was found that monkeys with VS lesions only had learning deficits in the stochastic

schedules (10). It was also found that monkeys with VS lesions made choices much faster (reaction time) than controls, and monkeys with VS lesions were displaying a speed-accuracy trade off, which accounted for errors in deterministic learning. In this same study, it was found that monkey with amygdala lesions had learning deficits across all schedules. This led the authors to conclude that the VS is important for stochastic RL, while the amygdala is important for both deterministic and stochastic RL.

In a follow up study monkeys with VS (excitotoxic) lesions (importantly these were the same VS monkeys from above) were compared to un-operated controls on a two-armed bandit reversal learning task (11). In this task monkeys were compared across 3 reward schedules (80%/20%, 70%/30%, 60%/40%), in randomly interleaved blocks where for one block type (object learning) the stimulus (regardless of location) was the rewarded feature, and another block type (action learning) in which the location (regardless of the stimulus) was the rewarded feature, it was found that monkeys with VS lesions only had deficits in the object learning blocks. This led the authors to conclude that the VS is only important for stimulus based RL. Importantly, in this second study the ability to learn to select the more rewarding visual stimulus was almost non-existent. This is interesting because in the first study the VS monkeys discriminated visual stimuli well, just not as well as control monkeys on stochastic schedules. In the context of

motivation (what one is motivated to do) versus ability (what one can do), this finding needs to be considered more carefully. But let's explore more results before we jump to any conclusions.

In another study monkeys with lesions to the VS were compared to un-operated controls on a task in which monkeys had to learn (deterministic) stimulus reward outcomes for 60 pairs of objects. Monkey only received one trial per day with each pair of objects. This study found that monkeys with VS (excitotoxic) lesions had no learning deficits when compared to un-operated controls (14). Similar results were found for monkeys with amygdala lesions. In a task where monkeys had to learn which cue predicted reward (food) in 40 novel visual cue pairs, it was found that monkeys with amygdala excitotoxic lesions had no learning deficits when compared to un-operated controls (98). One thing to note about these tasks is that the learning environment was deterministic (100, 0) -- one object/one cue was always rewarded and the other was never rewarded. The other difference to note is that in the first study animals only had one trial with each cue pair per a day, while in the second study, animals were given at least 4 trials to learn the correct cue.

This is a small subset of experiments, but this was done intentionally because it minimizes some external variables that likely affect results. These variables include who did the lesions, the method of the lesion, the monkeys used

(NIH has their own colony), and the staff who trained the monkeys (to name a few). Most of these variables are not typically worth trying to track, but when the literature is inconsistent it is advantageous to minimize as many variables as possible. The studies mentioned above were all done by the same two groups, albeit over different time frames and with different staff who trained the monkeys. It still presents a unique situation. Finding inconsistent results in this type of analysis speaks directly to the point made at the beginning of this introduction -- failing to understand what the task environments predicts, leads to false or overstated conclusions. In addition, this set of results is descriptive of the overall literature involving these two structures.

If a structure is responsible for a function, removal of that structure should eliminate or diminish an animal's ability to perform that function. As we said at the onset of this chapter, this is not exactly true, because other structures could pick up the slack when one structure loses a function. However, to find effects in some tasks, and not others, points to a more direct answer. For example the same VS animals were used in (10) and (11) but the deficits seen in the 'stochastic' stimulus based RL with the same reward schedules are larger in the latter study compared to the former. So, can monkeys with VS lesions learn 'stochastic' stimulus based RL, or not? Recall the concept of behavior contrast discussed above, this seems like another logical possibility. The learning results stated above

for the amygdala are hard to reconcile with one another and the standard RPE theory. However, these results make complete sense if dopamine signals motivation. We know from the behavior field that several things affect an animal's motivation and thus their behavior. What we established with the dopamine and effort section of this chapter, is that dopamine affects motivation and thus effort. These structures may play highly specific roles in RL but because of task design. It is always hard to detect.

Let's use the VS data mentioned above as an example, the conclusion from the data mentioned above is that the VS is important for stochastic but not deterministic learning (the authors of these papers concluded this). The problem with this conclusion is the experiments run did not actually test this. The RL literature as a whole has a curious way of testing stochastic learning, and it is unclear if the field is aware of this fact. All of the learning studies mentioned in this chapter (seems to be indicative of the literature as a whole) use a similar the type of reward structure. They all use stochastic concurrent reward schedules, schedules like (80/20,70/30, & 60/40). This is different than a purely stochastic schedule, which would look like (80/0, 70/0, & 60/0). This is even true in the two human drug studies mentioned above (19, 80). The former study used an 80/20 schedule, while the latter used a 75/25 schedule. However, there are better reasons to be skeptical about the conclusions these types of schedules produce. These two

types of reinforcement schedules lead to different types of behavior, and thus the conclusions that can be drawn from each are different. Deficits in the latter schedules are more indicative of learning or ability when compared to the former schedules. While deficits in the former schedules are more indicative of motivation, when compared to the latter schedules. While, both effects are informative and important, it is essential to know what is driving these effects to make proper conclusions. This is something behaviorist have long been aware of, but it something that gets lost in the RL optimality theory. Let's explore why this might be the case.

Often time conclusions drawn from RL optimality theory assumes that animals are trying to exclusively pick the best option in these concurrent reward schedules. This is interesting because behavioral choice theory states and shows this is not how animals behave. There is considerable evidence for theories like the matching law (31, 99, 100), which simply states that in concurrent reward environments, animals respond to cues approximately at the rates that they are reinforced. Ironically one of the environmental predictors for undermatching, in which animals respond to the less rewarded cue more than they should, is a rich reward environment (99). The conclusions drawn from studies that show undermatching are usually indicative of an organism simply being less sensitive to the reward environment. Most tasks looking at the matching law are choice and

not learning tasks, but even in learning tasks, at some point it becomes a choice task. We have already established that the optimal solution in standard operant reinforcement schedules that have rich reward environments is to respond faster for all possible actions. This RL optimality theory assumes that as the probability of reward for the two options get closer to one another it makes learning more difficult, so a 60/40 reward schedules is much harder to learn than an 80/20 schedule. Although this is one theory, it is unlikely when actual behavioral data is considered. The flipside of this idea theory is that as the reward probabilities get closer, the task gets easier, because no matter what I do/choose, my reward rate will be pretty constant (ie the reward probability being closer together, my incentive value goes down and signals that I do not need to have to try as hard). In essence as the reward schedules get closer to one another, all that has been done is to make responding in general more valuable by making rewards easier to get (less effort is required to get rewards). Ironically, a rich reward environment is often defined as an environment where it does not matter what the animal does -- rewards are everywhere and the effort cost for them is low. This view is consistent with the cost benefit analysis for dopamine described above. If this is the case, the learning literature stated above is consistent with the literature on effort and the vigor theory stated above. This is how the very same animals can sometimes show

remarkably different levels of deficits across tasks. The task change is likely manipulating motivation and not ability.

1.4.3 Two-Process learning theories

In one form or another it has long been thought that conditioning (sometimes referred to as learning) is under the control of two major systems. Behaviorist have called it *The Two-process learning theory* (101). Psychologist have referred to these two systems as a *fast and slow system* (102). Neuroscientist have referred to these two systems as an *actor and critic* (103). On the surface these theories seem quite different, but overall they are actually quite similar. In the following section we will focus on *The Two-process learning theory*, but in the correct lens the things discussed below apply to these other theories. The basic principle behind these two learning systems is that a discriminative stimulus acquires both incentive-motivational and discriminative-response properties.

In the conditioning section above we discussed how Pavlovian associations can affect operant behavior in things like PIT. However, this connection goes much deeper. Recall that Pavlovian conditioning is concerned with making associations for two classes of CRs, preparatory and consummatory. This means that in any operant procedure it would be advantageous to have these Pavlovian influences. This is exactly what the *two-process learning theory* suggest. This

theory suggest that an organism's preference for goals and objects is a combination of incentive and the reinforcement value of the discriminative-responses (36, 92, 104). Incentive value is thought to come from Pavlovian associations, while the reinforcement value of discriminative-responses comes from the interaction (actions required) of the operant contingency. We started this chapter by stating that an organism does not need to learn to get hungry (incentive value), but an organism does need to learn what behaviors best lead to acquiring food (the best reinforcement value). The response-dependent reinforcement value is determined by the factors listed above. Factors like effort, magnitude of reinforcement, richness of the reward, delay to reinforcement, and the resulting incentive motivation (36, 65). As with Pavlovian conditioning the incentive value is also affected by many of these factors. This value is just comprised of more factors (internal ones), and as I will show below stored as a separate form of value. This form of value compliments the response-dependent reinforcement value and is used to energize different classes of behaviors in particular environments (36, 60).

It has long been thought that responding in Pavlovian conditioning is a product of motivational state, sometimes called incentive value (105). Consider the (106) study in which Pavlovian approach behavior was shown to be directly affected by the basic motivational state of rats. They exposed food deprived rats to a CS that predicted peanut oil (fat US), and a CS that predicted sucrose

(carbohydrate US). Following this one group of rats were placed in a lipoprivic state (fat deprived), and another group of rats were placed in a glucoprivic state (sugar deprived). Rats in the glucoprivic state responded more to the CS that predicted sucrose compared to the CS that predicted peanut oil, and the opposite trend was found for rats in the lipoprivic state. Importantly, this experiments showed a link between the internal needs (these internal needs are likely not consciously known) of the organism and their current environment. This is important because these incentive values differ from the discriminative-response values. This can been seen is devaluation experiments. When a US is separately devalued in Pavlovian tasks, animals reject the US (they do not value it anymore) and their reaction changes to the CS (51, 107). This turns out not to be the case for discriminative-response value. In operant task, independently devaluing the US does not change the value of the US until the animal receives the devalued US as a part of the response contingency (51, 108). That is to say, animals will still work for the US until their actions lead to the devalued US. This provides strong evidence that these forms of value are stored differently and there is evidence for the forms of value being stored in distinct circuitry (51). Furthermore these forms of value (at least in early conditioning) are responsible for different classes of behavior, they follow the distinction between Pavlovian and operant behaviors discussed above (36, 51).

The important concept about this incentive system is it follows the rules of Pavlovian conditioning, thus it is thought not to be consciously controlled and often tied to emotional motivation (36, 51, 101). This incentive value can make behavior more probable by internal chemical changes. Recall the sex example given at the beginning of this chapter. I discussed research that shows that even though males are consciously unaware of when women around them are ovulating their testosterone levels are raised, this chemical change makes a certain operant class of behavior more probable (28). The operant class in this case is made up of all the behaviors in the organism's repertoire that has led to sex, with the strength (reinforcement value) of these behaviors being largely determined by the organism's history of reinforcement (28, 36). In this example the increase in testosterone is the incentive value that make behaviors that have previously lead to sex more probable.

This incentive (Pavlovian) value system associates appetitive and aversive events with environments, the smells, the visual stimuli, and time of day. It is this system that is more susceptible to being hijacked and can drive desirable and undesirable behavior. In fact this is one of the ways drug addiction, drug relapse (51), and drug overdoses are thought to happen (36). It is a well-known strategy for recovering drug addicts to stay away from environments where they have previously used drugs. The thought is that the sight and smell of the environment

has been associated with drug use (It is has become a CS), this association elicits chemical CRs with previous drug use (109). In the cause of drug use, one way this works is through a negative reinforcement contingency, the CR (internal preparation to deal with the drug) is aversive without drug use, thus the addict relapses to escape this aversive feeling. The CR in this case is thought to be protective, this preparatory behavior has been linked to drug tolerance (110). In fact, recall that Pavlovian conditioning is mainly concerned with the internal physiology (maintaining homeostasis) of an organism (28). Furthermore, drug relapse is associated with heightened drug-cue response in the same brain regions most often associated with RL (111). Feeding behavior is another example of how this incentive value system can promote a particular class of behaviors. Much like the drug example, whenever an organism eats (the US), there are internal chemical responses, and these internal responses are associated with the external environment to the point where the environment becomes a CS for the CR of eating. Indeed there is a considerable amount of research showing that a major part of *learned overeating* is due to Pavlovian conditioning (112). Consider how the smell of one's favorite baked good might increase the probability of eating, even though that before this smell one was not hungry or even thinking of this baked good.

1.4.4 Linking dopamine & learning

Now that we have linked this incentive value to Pavlovian conditioning, recall that this system is thought to have two major systems, the appetitive and aversive system. When the literature is examined in this light It becomes clear that it this appetitive incentive value that dopamine codes. Importantly, this Pavlovian incentive value is what I have referred to as motivation throughout this chapter. This incentive value is motivational because its value is based off elements outside of the learning contingency. At the beginning of this chapter I stated one of the problems with the current dopamine RPE theory is, "the type of learning the VS is thought to underlie more closely resembles operant conditioning, and much of the evidence cited above comes from appetite Pavlovian conditioning".

Understanding this *Two-process learning theory* makes it is easy to see how this happened, despite the considerable evidence showing that the striatum is not needed to learn operant contingencies. The discriminative-response and incentive-motivational properties co-vary in most experiments, making the contributions of each process difficult to appreciate (36). It's only when they are teased apart does this role for dopamine become clear. Dopamine serving this function unties and explains the conflicting RL literature. Let consider some of the results we have discussed above in the light of this theory and see if the role I suggested for dopamine above holds.

We have shown that optogenetic activation VTA dopamine neurons can increase approach behavior to cues (90) and facilitate configural cue learning without endowing cues with value (91). This result of association versus value on its own explains results that up until now, the dopamine RPE learning theory could not answer. In a task where rats were conditioned to expect one flavor of milk, dopamine recording revealed prediction errors to an unexpected different flavor of milk (113). Importantly, rats had no preferences for either flavor of milk, so the prediction error was to the identity of the milk. Recall in the dopamine section I stated that it is not my position that dopamine does not encode RPEs, the question becomes, what is this information used for? Making association is most certainly a form of learning, and it would require RPEs to better predict associations. This association theory also explains why dopamine has been found to respond to novel cues. The problem is this associative learning seems to be for estimates for the incentive-motivational value system, thus not the form of learning dopamine is said to underlie. This is the type of value I discussed in the dopamine section, the type of value associated with the vigor theory. In which case this type of motivational value is more important for the level of effort an organism will want to put forth in its current environment.

The view of dopamine explains the effort literature reviewed above, the fact that lesions to nucleus accumbens negatively affect effort (dopamine depletion =

less motivation = less effort). This view of dopamine explains the decrease in performance effects that haloperidol had in learning tasks. We also showed that this haloperidol effect extended toward effort tasks, animals given haloperidol chose the low effort/low reward arm more often than their control counterparts. This view of dopamine explains the increase performance effects that L-dopa had in learning tasks. This effect was consisted with the increased effort results from rats who received microinjections of d-amphetamine into the nucleus accumbens (increased dopamine = more motivation = more effort).

This theory also provides a frame work to understand the lesion results discussed above. I started this introduction by stating that task design is likely the reason for the conflicting results in the RL literature. This theory makes it easy to understand how this could be the case. The research on PIT suggests that this incentive value works more in a boost (more effort) fashion and can increase performance in operant tasks. Consistent with what I put forth in this section, I showed that this effect of PIT is at least partially mediated by the nucleus accumbens. This role of dopamine also provides a possible explanation for the conflicting results in that section and provides a way to investigate.

1.5 Proposal

1.5.1 Studying learning & our approach

A note of contention, most of the RL literature is composed of rich reward environments. This is not done for some systematic reason. It is instead done for a more practical reason. RL experimentalists realized what behaviorists have known -- rich reward environments keep animals motivated, and motivated animals do more trials. This approach has biased our understanding of RL. This bias is exacerbated in neurophysiology recording data, where more trials are always desired. A better approach to studying learning is to follow in the footsteps of Skinner (29). To understand and separate the environmental effects of learning tasks, we must run a series of tasks, where each task varies by just one parameter. This is how we can separate the motivational effects of the environment from learning ability, thus, make strong statements about learning.

This is the route this thesis will take. We will compare three groups of monkeys across of series of tasks, to make stronger statements about the contributions that the VS and amygdala make to RL. The three groups consist of monkeys with VS lesions, monkeys with amygdala lesions, and un-operated controls. The specific goal of this project is to use a series of tasks to evaluate and gain a better understanding of the role that these structures (VS & amygdala) play

in RL. As can be seen from above there is a lot of literature examining and identifying neural systems underlying RL. However, this literature is not wholly consistent and a number of different neural systems have been identified as being crucial for some component of RL.

This diversity of information can result from one of three reasons:

1. Difference in learning environments (tasks). Specifically, most of the literature supporting the current view of RL has used strictly appetitive environments. Behaviorists have shown that in many ways environment is one of, if not the most important attribute when it comes to predicting behavior. Consider a learning environment where the worst thing that can happen is the animal does not get rewarded, in contrast to an environment where if an animal does not pick the rewarded option, they lose a reward -- it is not difficult to see that the value of the reward in the latter environment is more than it is in the former. Finding a learning deficit in the former and not the latter leads to different conclusions than if data just exists for the former. These environments taken together change the conclusion from a learning deficit to a motivation deficit.
2. Definitions! Learning systems and their anatomical substrates can be dissociated in various ways. For example, learning is often studied using Pavlovian or instrumental paradigms (114). Formation of Pavlovian CS-US

associations is mediated, to some extent, by the amygdala (51, 79). Formation of instrumental associations, on the other hand, is thought to be mediated by frontal-striatal systems (115). Although there is considerable interaction between these behavioral processes in tasks like Pavlovian Instrumental Transfer (116) and conditioned reinforcement (117) it is important to realize that behaviorists separated these two types of conditioning because they result in different conditioning (learning) profiles which in turn leads to different behaviors. The fact that these two forms of conditioning can be separated behaviorally likely means these are separate processes that are likely mediated by different brain structures.

3. A number of the results that led to the strong belief of the current theory of RL come from physiology experiments. This is important to note because in these complex learning tasks it is possible that one could be mistaking a factor that correlates with value but is not in fact value (for example motivation). This is not a new idea and there has been some work explicitly examining this (118). In addition, just because one finds a correlate from some process does not mean that correlate is necessarily responsible for the identified process.

However, despite the diversity/conflicting literature, two areas that are often associated with being important for RL are the VS and the amygdala. Testing monkeys with lesions to these areas on a series of tasks can shed more light on the exact role each of these areas contributes to RL. Using monkeys with lesions helps control for point 3 listed above because if one of the identified areas is responsible for some component of RL, removing said area should result in learning deficits. It is possible that once one of these areas is lesioned a different area can take over its role, however if we find deficits in one task and not the other it is more likely that the area is contributing to one aspect of RL (likely the difference between the two tasks). Testing the same monkeys on a series of tasks that emphasize different components of RL helps control for point 1 and point 2. In addition, testing these monkeys on a number of tasks can help narrow down the role each of these areas contributes to RL.

Chapter 2: Effects Amygdala Lesions on Object-Based versus Action-Based Learning in Macaques

2.1 Introduction

Learning to execute actions or select objects that lead to rewards is critical for survival. While formal models of reinforcement learning (RL) do not distinguish between these (4) there is considerable evidence to support the view that separate neural circuits mediate learning about the value of actions versus objects. Starting in the visual system, there is a distinction between spatial vision and object vision, that has been referred to as the dorsal (spatial) and ventral (object) visual streams hypothesis (119). A related view suggests that the distinction between the two systems involves processing information for action versus perception (120). The anatomical separation between these systems continues into prefrontal cortex (121, 122) and also through the frontal-basal ganglia-thalamo-cortical loops (123, 124). There is also interaction between these circuits (125), especially when object information is required to select spatially directed actions (126). But to some extent these processing streams are segregated. The anatomy, therefore, suggests that learning to associate rewards with actions

may rely more on dorsal circuitry, and learning to associate rewards with objects may rely more on ventral circuitry (127).

RL has often been linked to the ventral striatum (VS). This suggests that the VS underlies both learning to associate rewards with actions and stimuli. There is considerable evidence for the role of the VS in object based RL (7, 10), particularly when it comes to learning to choose between two positive outcomes that vary in magnitude (128). There has been less evidence for the role of the VS in action selection. When action and object learning have been studied in the same experiment, monkeys with lesions to the VS had deficits in object but not action based RL (11). Other work has shown that the dorsal striatum (DS) plays a role in learning to associate actions (129-131) and action sequences (132, 133) with rewards. This suggests that different neural circuits underlie these two different types of learning, at least in the striatum.

The amygdala has also been shown to play an important role in visual object based RL (10, 23, 51, 134-136) and other forms of reward learning (137, 138). Studies in monkeys have shown that lesions of the amygdala lead to learning deficits in a probabilistic reversal learning task (10) and a reward magnitude learning task (23). And, amygdala lesions lead to a decrease in the information about stimuli associated with rewards, relative to pre-lesion recordings, in the orbital prefrontal cortex (23, 139). Therefore, there is considerable evidence that

supports a role for the amygdala in learning to associate objects with rewards. Furthermore, the amygdala has strong anatomical connections with the ventral visual pathway, and less pronounced connections with dorsal pathway structures (127, 140, 141). Although the amygdala contributes to object reward learning, and has strong links to the ventral visual pathway, it also has links to the dorsal pathway. For example, single neurons in the amygdala code the locations of chosen objects independent of reward expectation (142, 143). In addition, the amygdala projects to cingulate motor areas (144), which provides a potential route for the amygdala to influence action learning. Whether the amygdala makes a causal contribution to learning to choose rewarded locations, however, has not been directly examined.

To determine the amygdala's role in action- versus object-based RL, we tested four monkeys with excitotoxic lesions of the amygdala on a two-arm bandit reversal learning task used previously to examine learning following VS lesions (11). This task involved two different types of learning, carried out in blocks of trials. In one block type the monkeys had to learn to pick the location (action based) that yielded the most rewards, and in the other block type monkeys had to learn to choose the stimuli (object based) that led to the most rewards. We found that the amygdala plays a role in both action- and object-based RL.

2.2 Methods

2.2.1 Subjects

The subjects included 10 male rhesus macaques with weights ranging from 6-11 kg. Four of the male monkeys received bilateral excitotoxic lesions of the amygdala. The remaining six monkeys served as unoperated controls. Four out of the six unoperated control monkeys were the same monkeys used in a previous study (10). Five out of the six unoperated control monkeys were the same monkeys from an additional study (11). All remaining monkeys were not previously used in the studies mentioned above. In particular none of the amygdala lesion monkeys (n = 4) were previously used in the studies mentioned above. For the duration of the study, monkeys were placed on water control. On testing days monkeys earned their fluid from their performance on the task. Experimental procedures for all aspects of the study 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.

2.2.2 Surgery

Four monkeys received two separate stereotaxic surgeries, one for each hemisphere, which targeted the amygdala using the excitotoxin ibotenic acid (for

details, see (10). Injection sites were determined based on structural magnetic resonance (MR) scans obtained from each monkey prior to surgery. After both lesion surgeries had been completed, each monkey received a cranial implant of a titanium head post to facilitate head restraint. Unoperated controls received the same cranial implant. Behavioral testing for all monkeys began after they had recovered from the implant surgery.

2.2.3 Lesion assessment

Lesion volume estimates were taken by first transforming each subject's T2-weighted scan acquired one week post-operatively to the standard NMT (NIMH macaque template; (145)) using AFNI's 3dAllineate function (146). We then applied thresholding to identify the area of hyperintensity on the transformed T2-weighted object to isolate a binary mask that corresponded to the area of damage. The masks were visually inspected and manually edited to ensure that they fully captured the areas of hyperintensity on the T2-weighted object. A lesion overlap map was created by summing the binary masks for each hemisphere and displaying the output on the NMT (Fig. 1C).

As intended, all operated monkeys sustained extensive damage to the amygdala, bilaterally; the estimated percent damage ranged from 86 to 95% (Table S1, Fig 1C). Surrounding structures, mainly the entorhinal cortex, sustained

inadvertent damage that varied widely in extent (Table S1). Based on prior work, the percent damage to entorhinal cortex as estimated from the T2-weighted scans is almost certainly an overestimate (147).

2.2.4 Task & apparatus

We tested rhesus macaques (*Macaca mulatta*) on a probabilistic two-arm bandit reversal learning task. During the experiment animals were seated in a primate chair facing a computer screen. Eye movements were used as behavioral readouts. In each trial, monkeys first acquired central fixation (Fig 1A, B). After a fixation hold period of 500 ms, we presented two objects, left and right of fixation. Monkeys made saccades to one of the two objects to indicate their choice. After holding their choice for 500 ms, a reward was stochastically delivered according to one of three reward schedules: 80%/20%, 70%/30%, 60%/40%. In an 80%/20% reward schedule one of the choices led to a reward 80% of the time and the other choice led to a reward 20% of the time. The reward schedule and stimuli were used for a total of 80 trials, which constituted one training block. At the beginning of each block, two novel objects were introduced and the block was randomly assigned a reward schedule; this assignment remained constant throughout the entire block. In addition, on each trial the location of 'best' object, left or right of fixation, was randomized.

There were two different block types: What and Where. In 'what' blocks, the higher-probability option was one of the two objects independent of which side it was presented on. In 'where' blocks, the higher-probability option was one of the two saccade directions independent of the object that was selected. There was no cue to indicate block type; monkeys determined block type by making choices and getting feedback. As with the reward schedule, the block type remained constant throughout the entire 80-trial block. In each block, on a randomly selected trial between 30 and 50, inclusive, the reward mapping was reversed, making the previously lower probability option the higher probability option. The reversal trial was not cued; monkeys had to learn through trial and error that the reward mapping switched.

2.2.5 Task training

All animals were trained on the task using the same procedure. Eight out of ten monkeys (5 of the 6 controls and 3 of the 4 lesion monkeys) had a more extensive training history. These monkeys completed other tasks before beginning training for the current task. In the previous tasks they learned only object based reward associations. After the remaining two monkeys (1 control and 1 lesion) learned to make saccades to fixate on targets they were trained on a simple

two arm bandit RL task in which they learned only object based reward associations.

Next, all monkeys were trained with a deterministic schedule (100/0) in both the What and Where conditions. Monkeys were first introduced to one block type, either What or Where, with block type randomly assigned and balanced across the group. Once monkeys could successfully perform 15-24 blocks per session, we introduced the other block type by itself, and then upon stabilized performance in that block type, we mixed the two block types. Once the monkeys reached stable performance in the deterministic setting, we gradually introduced probabilistic outcomes; probabilities were lowered until the final schedules of 80/20, 70/30, 60/40 were reached.

2.2.6 Eye tracking

Objects provided as choice options were normalized for luminance and spatial frequency using the SHINE toolbox for MATLAB (148). All objects were converted to grayscale and subjected to a 2D FFT to control spatial frequency. To obtain a goal amplitude spectrum, the amplitude at each spatial frequency was summed across the two object dimensions and then averaged across objects. Next, all objects were normalized to have this amplitude spectrum. Using luminance histogram matching, we normalized the luminance histogram of each color channel

in each object so it matched the mean luminance histogram of the corresponding color channel, averaged across all objects. Spatial frequency normalization always preceded the luminance histogram matching. Each day before the monkeys began the task, we manually screened each object to verify its integrity. Any object that was unrecognizable after processing was replaced with an object that remained recognizable. Eye movements were monitored and the object presentation was controlled by PC computers running the Monkeylogic (version 1.1) toolbox for MATLAB (149) and Arrington Viewpoint eye-tracking system (Arrington Research).

2.2.7 Bayesian model of reversal learning

We fit a Bayesian model to estimate probability distributions over several features of the animals' behavior as well as ideal observer estimates over these features (10, 11, 150, 151). The Bayesian ideal observer model inverts the causal model for the task, so it is the optimal model. Using the ideal model we estimated probability distributions over reversal points to estimate when a reversal occurred. To estimate the Bayesian model we fit a likelihood function given by:

(1) $$f(x, y|r, p, h, b) = \prod_{k=1}^{T} q(k)$$

Where r is the trial on which the reward mapping is reversed ($r \in$ 0-81) and p is the probability of reward of the high reward option. The variable h encodes

whether option 1 or option 2 is the high reward option at the start of the block ($h \in 1, 2$) and b encodes the block type ($b \in 1, 2$ – What or Where). The variable k indexes trial number in the block and T is the current trial. The variable k indexes over the trials up to the current trial so, for example, if $T = 10$, then $k = 1, 2, 3, \ldots 10$. The variable r ranges from 0 to 81 because we allow the model to assume that a reversal may not have happened within the block, and that the reversal occurred before the block started or after it ended. In either scenario where the model assumes the reversal occurs before or after the block, the posterior probability of reversal would be equally weighted for r equal to 0 or 81. The choice data are given in terms of x and y, where elements of x are the rewards ($x_i \in 0, 1$) and elements of y are the choices ($y_i \in 1, 2$) in trial, i. The variable p is varied from 0.51 to 0.99 in steps of 0.01. It can also be indexed over just the exact reward schedules (i.e. 0.8, 0.7 and 0.6), although this makes little difference as we marginalize over p for all analyses.

For the ideal observer model used to estimate the block type in the Bayesian analysis, we estimated the block type probability at the current trial, T, based on the outcomes from the previous trials. Thus, the estimate is based on the information that the monkey had when it made its choice in the current trial. For each schedule, the following mappings from choices to outcomes gave us $q(k)$. For estimates of What (i.e. b = 1), targets 1 and 2 refer to the individual objects

and saccade direction is ignored; whereas for Where (i.e. b = 2), targets 1 and 2 refer to the saccade direction and the object is ignored. For $k < r$ and $h = 1$, (when target 1 is the high probability target and the trial is prior to the reversal) choose 1 and get rewarded $q(k) = p$, choose 1 and receive no reward $q(k) = 1 - p$, choose 2 and get rewarded $q(k) = 1 - p$, choose 2 and have no reward $q(k) = p$. For $k \geq r$ these probabilities are flipped. For $k < r$ and $h = 2$ the probabilities are complementary to the values where $k < r$ and $h = 1$. To estimate reversal, all values were filled in up to the current trial, T.

For the animal's choice behavior, used to estimate the posterior over b for each group, the model is similar, except the inference is only over the animal's choices, and not whether it is rewarded. This model assumes that the animal had a stable choice preference which switched at some point in the block from one object to the other. Given the choice preference, the animals chose the wrong object (i.e. the object inconsistent with their choice preference) at some lapse rate 1-p. Thus, for k < r and h = 1 choosing option 1: q(k) = p, choosing option 2: q(k) = 1 – p. For k >= r and h = 1, choosing option 1: q(k) = 1 – p, choosing option 2: q(k) = p. Correspondingly for k < r and h = 2, choosing option 2: q(k) = p, etc. The choice behavior model is therefore similar to the ideal observer, except p indexes reward probability in the ideal observer model and 1-p indexes the lapse rate in the

behavioral model. The reward outcome also does not factor into the behavior model.

Using these mappings for $q(k)$, we then calculated the likelihood as a function of r, p, h, and b for each block of trials. The posterior is given by:

(2) $$p(r,p,h,b|x,y) = f(x,y|r,p,h,b)p(r)p(p,h,b)/p(x,y)$$

For r, p, h and b, the priors were flat. There is general agreement between the ideal observer estimate of the reversal point and the actual programmed reversal point (10, 150).

With these priors, we calculated the posterior over the reversal trial by marginalizing over p, h and b.

(3) $$p(r|x,y,M) = \sum_{p,h,b} p(r,p,h,b|x,y)$$

The posterior over block type could correspondingly be calculated by marginalizing over r, p and h.

2.2.8 Reinforcement learning model of choice behavior

We fit 6 different reinforcement learning models that varied in the number of parameters used to model the data. In the results we focus on the two models that most often accounted for the behavior. All models were based on a Rescorla-Wagner (RW), or stateless RL value update equation given by:

(4) $$v_i(k+1) = v_i(k) + \alpha_f\big(R - v_i(k)\big)$$

We then passed these value estimates through a logistic function to generate choice probability estimates:

(5) $$d_j(k) = (1 + e^{\beta(v_i(k) - v_j(k) + h_i(k) - h_j(k))})^{-1}, \quad d_i(k) = 1 - d_j(k)$$

The variable v_i is the value estimate for option i, R is the reward feedback for the current choice for trial k, and α_f is the learning rate parameter, where f indexes whether the current choice was rewarded ($R = 1$) or not ($R = 0$). For each trial, α_f is one of two fitted values used to scale prediction errors based on the type of reward feedback for the current choice. Note that models M1, M2, and M3 described below do not have the h_j factors. The variable $h_j(k)$ implemented a choice autocorrelation function, which increased the value of a cue that had occurred in the same location, recently. This allows us to model a tendency to repeat a given choice, independent of whether it was rewarded. Because we wanted to use the same model across both What and Where, we implemented the choice autocorrelation functions as repetitions of choices when the same object occurs in the same location, which results in autocorrelations across 4 terms (i.e stim 1 left, stim 1 right, stim 2 left, stim 2 right). A model which used only choice repetition across location could not be fit to the Where condition, since the animals should "perseverate" on location. Thus, the use of object-location terms for the autocorrelation allows us to use the same model in both tasks.

The autocorrelation function was defined as follows:

(6) $$h_i(k) = \kappa e^{-\lambda(k-k_{l(i)})}$$

where the variable κ and λ were free parameters scaling the size of the effect and the decay rate, respectively. The variable $k_{l(i)}$ indicates the last trial on which a given object was chosen in a given location, i. There were four separate values for $k_{l(i)}$ as it tracked two cues across locations. The values entered into equation 5 were the two (of the 4) that corresponded to the object/location pairs actually presented in the current trial. These parameters allowed us to characterize choice-perseveration across the interaction of object and action choices.

The likelihood was given by:

(7) $$f(x, y|\beta, a, \kappa, \lambda) = \prod_k [d_1(k)c_1(k) + d_2(k)c_2(k)]$$

Where $c_1(k)$ had a value of 1 if option 1 was chosen on trial k and $c_2(k)$ had a value of 1 if option 2 was chosen. Conversely, $c_1(k)$ had a value of 0 if option 2 was chosen, and $c_2(k)$ had a value of 0 if option 1 was chosen for trial k. We used standard function optimization methods to maximize the likelihood of the data given the parameters. Note, not all parameters were present in all models.

Three of the models (M1, M2, & M3) had different numbers of learning rate and inverse temperature parameters. M1 had one inverse temperature and two learning rate parameters (indexed by the subscript *f* on α), one for positive feedback and one for negative feedback. M2 had one inverse temperature and one

learning rate parameter. M3 had two inverse temperatures, one for the acquisition phase and one for the reversal phase, and four learning rates, two for the acquisition phase (one for positive feedback and one for negative feedback), and two for the reversal phase(one for positive feedback and one for negative feedback). The remaining three models are the plus versions of the models discussed above (M1+, M2+, & M3+). The plus models have the same number of parameters as the basic (i.e. M1, M2, M3) model with the addition of two more parameters, one for the coefficient on the autocorrelation factor, κ, and one for the decay factor on the autocorrelation, λ. Models M2 and M2+ predicted behavior most often across groups, so to simplify presentation we show results for these two models only, and the plots for these models show only the number of times these were the best model.

2.2.9 ANOVA models

To quantify the difference between choice behavior in each group, we first flipped the data following the reversal when we entered it into the ANOVA. Data is plotted unflipped. Next, we performed an arcsine transformation on the choice accuracy values from each session, as this transformation normalizes the data (152). Data were then averaged across sessions within monkey. We then carried out an N-way ANOVA (ANOVAN). Monkey was included as a random factor.

All other factors were fixed effects. For all reported ANOVAs, we always ran an omnibus model with all factors and interactions of all order. Non-reported interactions were not significant. The ANOVA on win-stay lose-switch, entropy and reversal trial difference were done in the same way as above without the arcsine transformation. For the choice strategy model, we entered both win-stay and lose-switch as dependent variables and included a factor in the model for choice-strategy (i.e. either win-stay or lose-switch). Effect size is reported using ω^2 (153).

2.3 Results

We tested rhesus macaques on a two-armed bandit reversal learning task with three different stochastic reward schedules: 80%/20%, 70%/30%, 60%/40%. In addition to the three different reward schedules, there were two different block types: What and Where. In 'what' blocks, the higher-probability option was one of the two objects independent of the chosen location. In 'where' blocks, the higher-probability option was one of the two saccade directions independent of the chosen object. There was no cue to indicate block type. Therefore monkeys determined block type by making choices and getting feedback. In each block, on a randomly selected trial between 30 and 50 (inclusive) the reward mapping was reversed, making the previously lower probability option the higher probability option and

vice versa. The reversal trial was not cued and therefore monkeys had to learn through trial and error that the reward mapping switched.

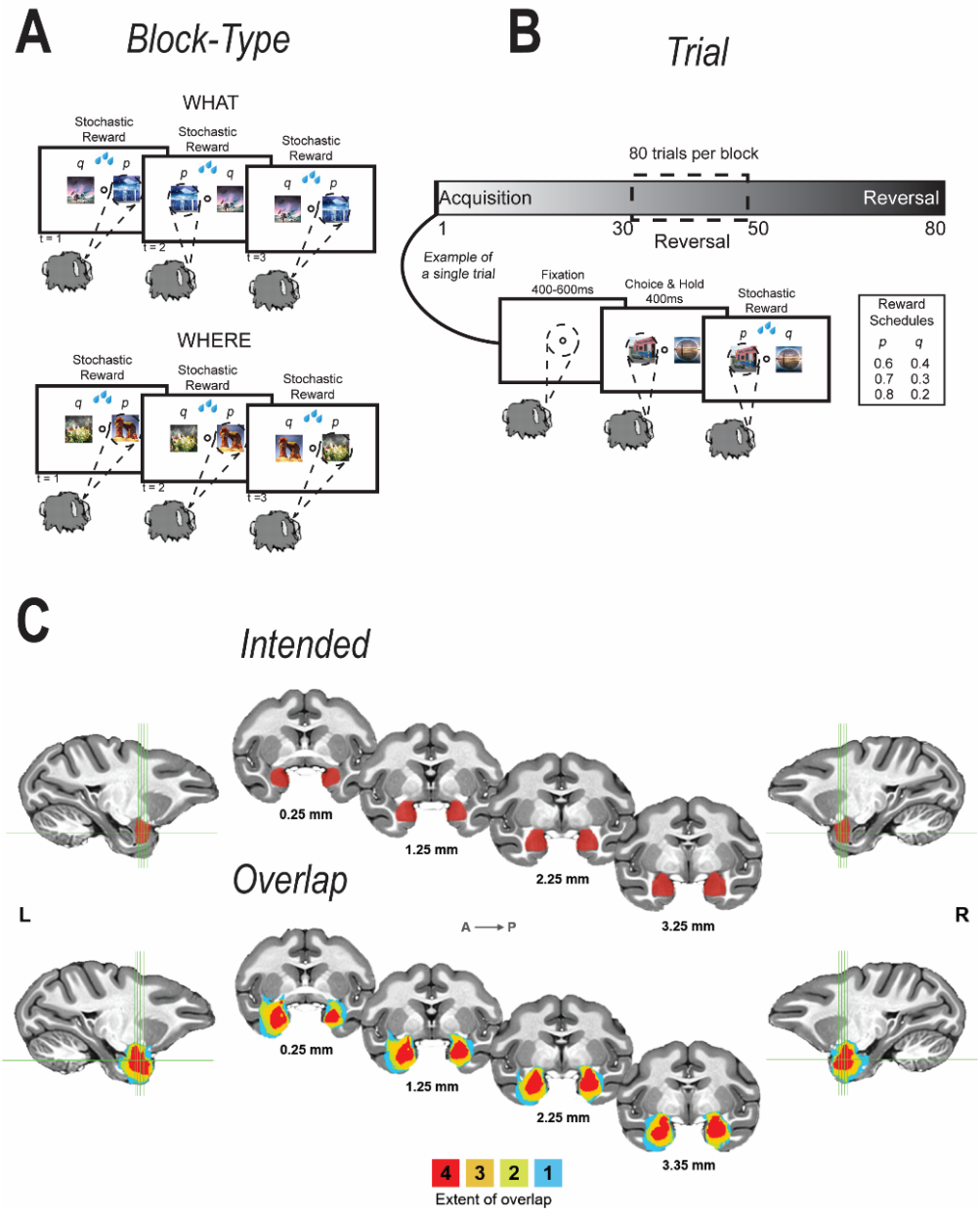


Figure 1. Task and lesion extent. A, B. What and Where Task. The task was divided into 80-trial blocks. At the beginning of each 80-trial block we introduced two new objects that the animal had never seen before. Each block of trials was either a What block or a Where block. If it was a What block, we assigned a high reward probability to one object and a low reward probability to the other. If it was a Where block we

assigned a high reward probability to one of the locations and a low reward probability to the other. We used three different reward schedules (80/20, 70/30 and 60/40). The reward schedules were randomly assigned to the block and remained fixed for the block. The block type was also randomized and remained fixed for the entire block. There was no cue to indicate block type. Therefore, the animals had to infer the block type. In addition, on a randomly chosen trial between 30 and 50 we reversed the choice-outcome mapping, such that the better choice became the worse choice and vice-versa. C. Extent of lesion and number of animals with shown extent, overlaid on a standardized macaque brain template.

2.3.1 Choice behavior

We began by analyzing the monkeys' choice behavior. Because the reversal trial differed across blocks, we first aligned each block to the true reversal point and interpolated the trials in the acquisition and reversal phases so there were 40 "trials" in each. We then carried out ANOVAs on this data, where the dependent variable was the fraction of times the animals chose the best initial option (Fig. 2). We first carried out an ANOVA across both block types (What and Where). There was no average difference in performance across block type (Block-Type; $F(1,8) = 0.05$, $p = 0.83$, $\omega^2 = 0$). We did however, find differences in reward schedule (Schedule; $F(2,16) = 107$, $p < 0.001$, $\omega^2 = 0.171$) on choices. There were also differences in these factors by trial (Block Type x Trial; $F(78,624) = 4.1$, $p < 0.001$, $\omega^2 = 0.007$ and Schedule x Trial; $F(156,1248) = 23.4$, $p < 0.001$, $\omega^2 = 0.026$), which reflects both the initial learning, and the reversal of choices after the reversal in reward mapping. We also found that control monkeys performed better than amygdala monkeys and this varied by trial (Group x Trial; $F(79,632) = 2.3$, $p < 0.001$, $\omega^2 = 0.014$), and by schedule (Group x Schedule x Trial; $F(156,1248) =$

1.3, p = 0.01, ω^2 = 0.001). The groups did not, however, differ by block type (Group x Block Type x Trial; F(79,632) = 0.6, p = 0.99, ω^2 = 0.001).

Although there were no group differences by block type, we carried out planned comparisons on the data from each condition. In What blocks both groups chose more accurately in the richer reward schedules (Schedule x Trial; F(156,1248) = 16.3, p < 0.001, ω^2 = 0.027). We also found that control monkeys performed better than the amygdala lesioned monkeys and this varied by trial (Group x Trial; F(78,624) = 2.1, p < 0.001, ω^2 = 0.013), and by schedule (Group x Schedule x Trial; F(156,1248) = 1.4, p = 0.003, ω^2 = 0.002). Similarly, in Where blocks, both groups chose more accurately in the richer reward schedules (Schedule x Trial; F(156,1248) = 8.9, p < 0.001, ω^2 = 0.029). Control monkeys performed better than the amygdala monkeys which varied by trial (Group x Trial; F(79,632) = 1.6, p < 0.001, ω^2 = 0.018), but there was no difference across schedule (Group x Schedule x Trial; F(156,1248) = 1.1, p = 0.27, ω^2 = 0.003).

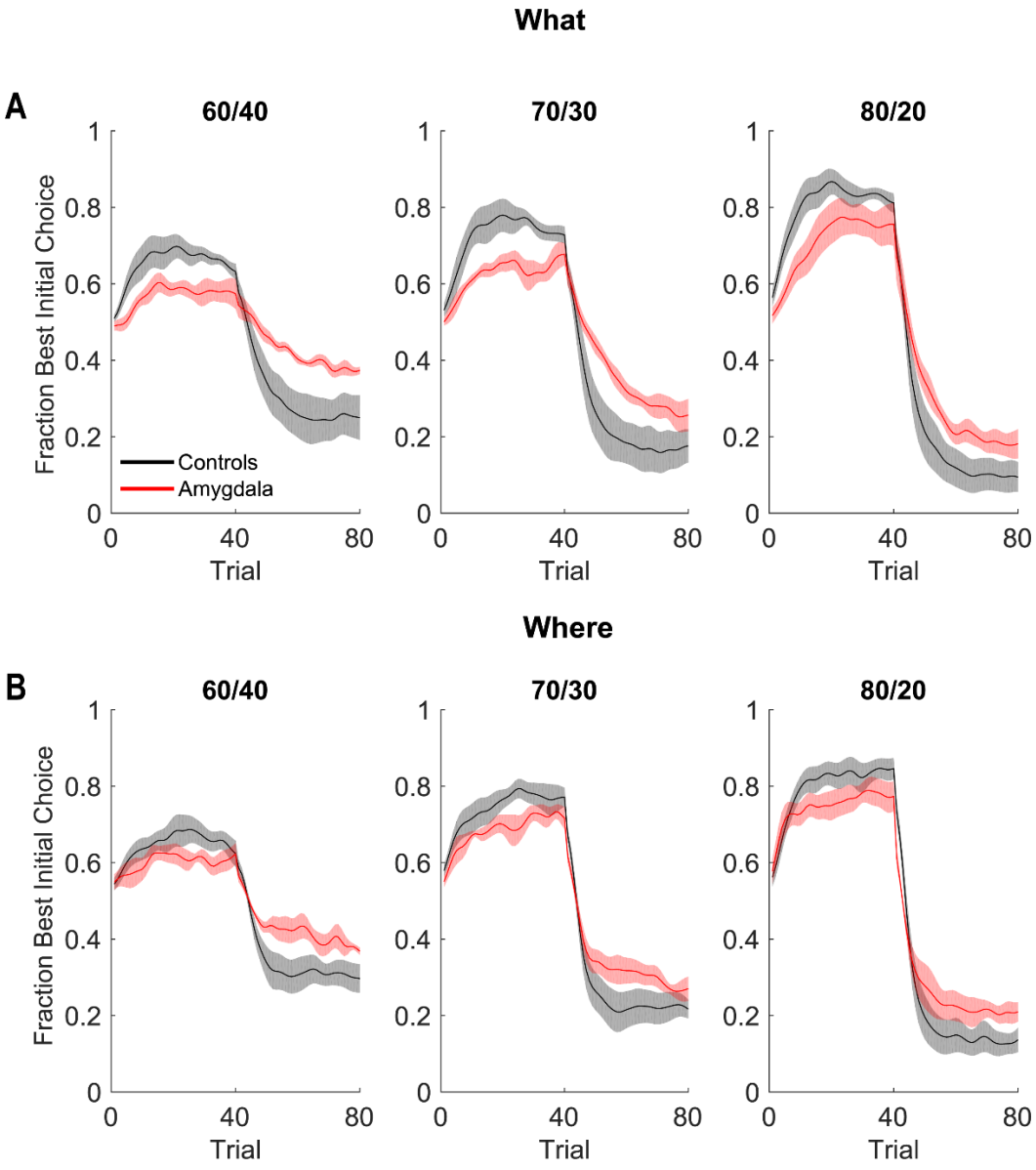


Figure 2. Behavioral performance in What and Where conditions. A. Fraction of times the animals chose the best initial cue in the What condition. Shaded region indicates +/- 1 s.e.m., where the N = the number of animals in each group (4 lesion, 6 control). B. Same as A for the Where condition.

To further characterize the learning behavior, we analyzed the win-stay, lose-switch performance (Fig. 3). Win-stay is the probability that the animals chose the same option after a positive outcome in the previous trial and lose-switch

is the probability that they chose the other option after a negative outcome in the previous trial. For purposes of the ANOVA we analyzed only win-stay and lose-switch probabilities. The difference between win-stay and lose-switch was coded as a choice-strategy effect. We found differences across block types (Block type; $F(1, 8) = 23.7$, $p = 0.001$, $\omega^2 = 0.006$). Consistent with the decreased overall accuracy of the lesioned animals, they also had lower win-stay strategies relative to higher lose-switch than controls (Group x Choice Strategy; $F(1, 8) = 6.4$, $p = 0.036$, $\omega^2 = 0.035$). These group strategies did not differ by block type (Group x Choice Strategy x Block-type; $F(1, 8) = 1.8$, $p = 0.217$, $\omega^2 = 0.004$). We then ran the analysis separately for Win-stay and Lose-switch strategies and found that there were no group differences for Win-stay (Group; $F(1, 8) = 3.6$, $p = 0.095$, $\omega^2 = 0.101$). However, lesioned animals more often switched following a negative outcome (Group; $F(1, 8) = 11.1$, $p = 0.010$, $\omega^2 = 0.393$). Therefore, across block types, lesioned animals switched after a negative outcome more frequently than the control animals. The groups also differed by block type (Group x Block type; $F(1, 8) = 16.3$, $p = 0.004$, $\omega^2 = 0.004$). When we ran the analysis separately for each group, controls did not differ by block type (Block type; $F(1,5) = 0.6$, $p = 0.46$, $\omega^2 = 0$) but the lesioned animals did (Block type; $F(1,3) = 21.7$, $p = 0.018$, $\omega^2 = 0.37$).

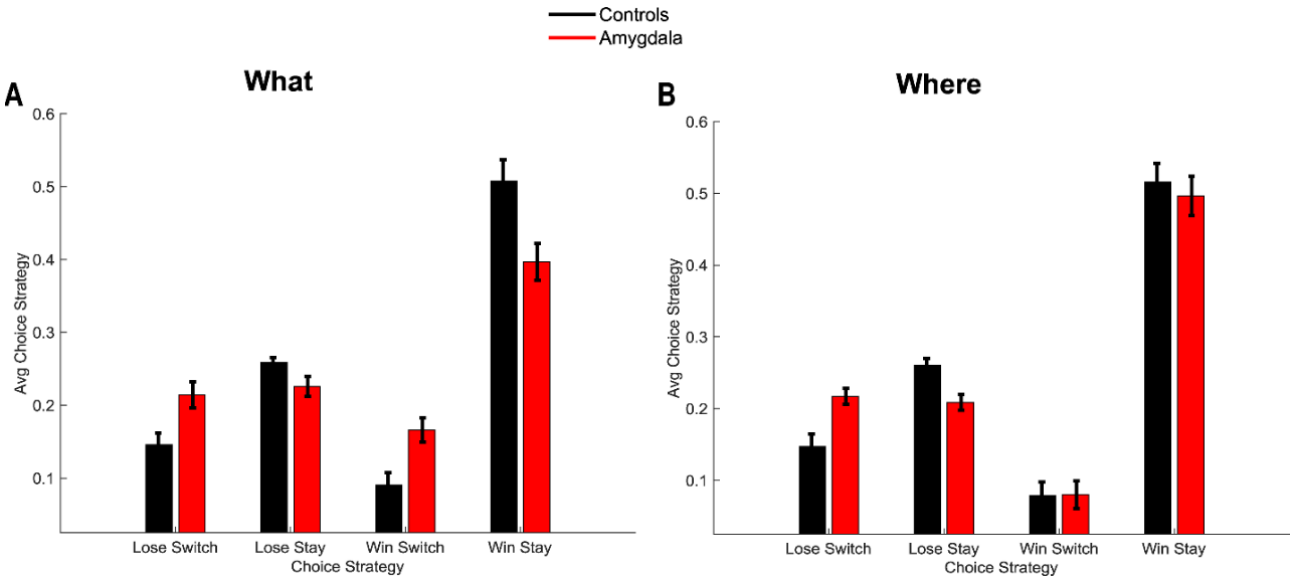


Figure 3. Win-stay, Lose-switch. A. Win-stay, lose switch performance for the two groups in the What condition, averaged across schedules. B. Win-stay, lose-switch performance for the two groups in the Where condition, averaged across schedules. Error bars are +/- 1 s.e.m. (N = 6 control, 4 lesion).

Next, we looked at the probability that while animals were in one block type they were making choices consistent with the other block type (Fig. 4). For What blocks we quantified the probability of choosing the most frequently chosen location and for Where blocks we quantified the probability of choosing the most frequently chosen object. On average, monkeys should be choosing each action at chance levels in What blocks, and each object at chance levels in Where blocks. However, even if they infer the correct block type, at the beginning of the block, they may make choices consistent with the wrong block type for several trials, and this can persist into the block.

We started by analyzing both block types together and found that animals were closer to chance, and therefore were making choices more consistent with the appropriate block type in easier schedules (Schedule; $F(2,16) = 2.6$, $p < 0.001$, $\omega^2 = 0.009$). In addition, we found that the groups differed across schedule and block type (Group x Schedule x Block-type; $F(2,16) = 11.2$, $p < 0.001$, $\omega^2 = 0.018$). The amygdala lesioned animals were making relatively more location choices in What blocks than Control animals, when compared to object choices in Where blocks and this differed by schedule. To examine this in more detail we analyzed each block type separately. In What blocks (Fig. 4A) we found that animals performed better in easier schedules (Schedule; $F(2,16) = 37.5$, $p < 0.001$, $\omega^2 = 0.35$). We also found that the groups differed across schedule (Group x Schedule; $F(2,16) = 8.2$, $p = 0.003$, $\omega^2 = 0.039$). In Where blocks (Fig. 4B) we again found that animals performed better in easier schedules (Schedule; $F(2,16) = 19.6$, $p < 0.001$, $\omega^2 = 0.042$). There were, however, no group differences across schedules (Group x Schedule; $F(2,16) = 2.9$, $p = 0.081$, $\omega^2 = 0.006$).

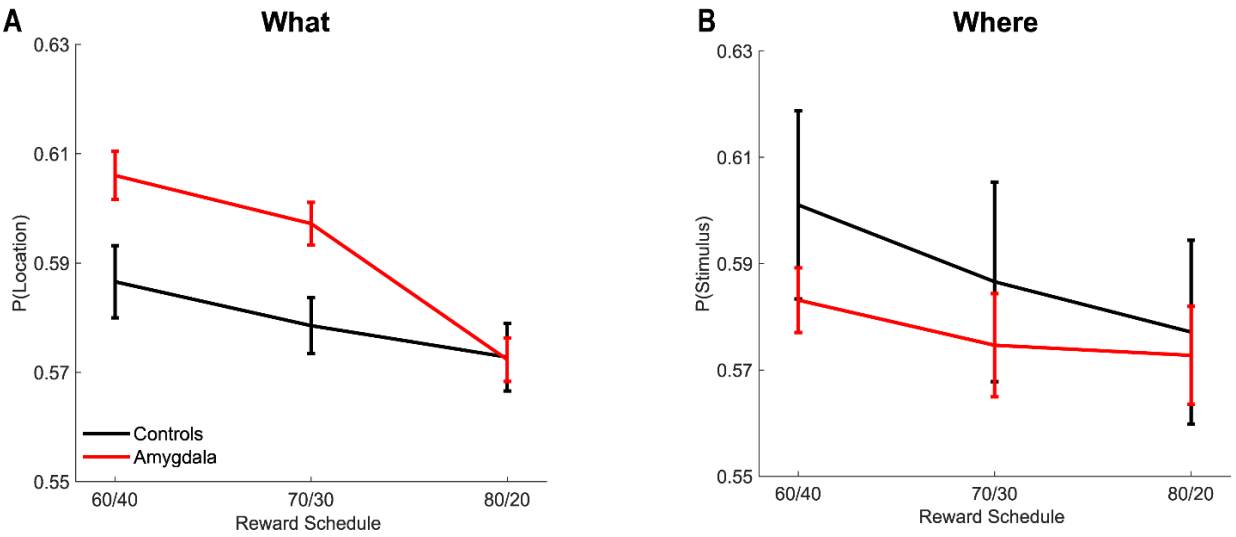


Figure 4. Cross condition choice frequencies. A. Probability of choosing the most frequently chosen location in the What condition, averaged across reversals (with reversal data flipped). B. Probability of choosing the most frequently chosen object in the Where condition. Error bars are +/- 1 s.e.m., where the N = the number of animals in each group (4 lesion, 6 control).

2.3.2 Reinforcement learning model

To further investigate why the monkeys with amygdala lesions behave differently we fit several RL models which varied in the number of free parameters used to model the choice behavior (see methods). We used the Bayesian Information Criterion (BIC) to assess which model fit best in each session for each animal (Fig. 5). Across monkeys, the model which most frequently fit best had 4 parameters (M2+). The M2+ model had one learning rate, one inverse temperature, one autocorrelation parameter and one decay parameter which decays the autocorrelation perseveration effects. The autocorrelation and decay parameters characterized the tendency to perseverate on choices, independent of whether they were appropriate to the current block. The model which fit second

most frequently had 2 parameters (M2). The M2 model had one learning rate and one inverse temperature parameter but no perseveration parameters. The M2+ model captures perseverative choice biases driven by choices consistent with the opposite block type. Therefore, a preference for the plus model suggests the monkeys choices are not driven by the choice-outcome effects for the current block type, to the same extent.

The relative preference for the M2+ model was larger in amygdala animals than controls in the What condition than the Where condition (Fig. 5; Group x Block-Type x Model; $F(1,8) = 6$, $p = 0.040$, $\omega^2 = 0.083$). Next we split the analysis by block type. In the What condition there was a preference for the M2+ model (Model; $F(1,8) = 21.5$, $p = 0.001$), but there was no preference in the Where condition (Model; $F(1,8) = 0.7$, $p = 0.434$). However, there were no group effects or interactions with group in either the What or Where conditions ($p > 0.05$). Therefore, there was a shift towards the M2 model, relative to the M2+ model, in the amygdala animals in the Where conditions. Since the plus model captures a tendency to repeat a response to an object at a specific location, independent of reward, this suggests that there is a shift towards a less object dependent strategy in the Where condition, relative to a location dependent strategy in the What condition, in the amygdala animals.

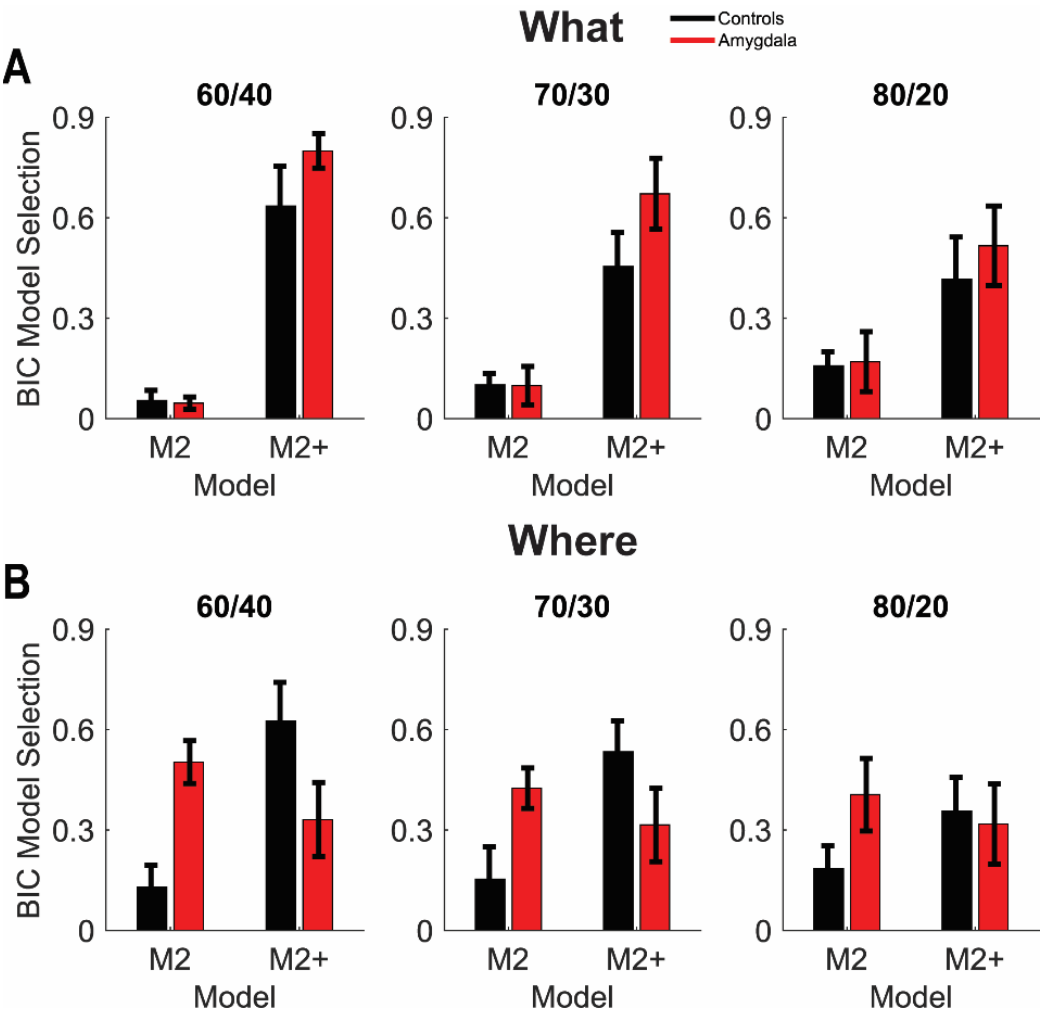


Figure 5. Bayesian Information Criterion (BIC) model selection. A. Percentage of sessions BIC selected models M2 and M2+ (out of all 6 models) in the What condition. Error bars are +/- 1 s.e.m., where the N = the number of animals in each group (4 lesion, 6 control). B. Same as A for the Where condition.

Next, we examined the parameters for the M2+ model. The only parameter that the groups differed on was the autocorrelation coefficient, κ (Fig. 6), which was larger in controls across both conditions (Group; $F(1,8) = 15.6$, $p = 0.004$, $\omega^2 = 0.582$). The autocorrelation factor captures the tendency to repeat choices of

objects at specific locations, and therefore captures perseveration across actions and objects (154, 155).

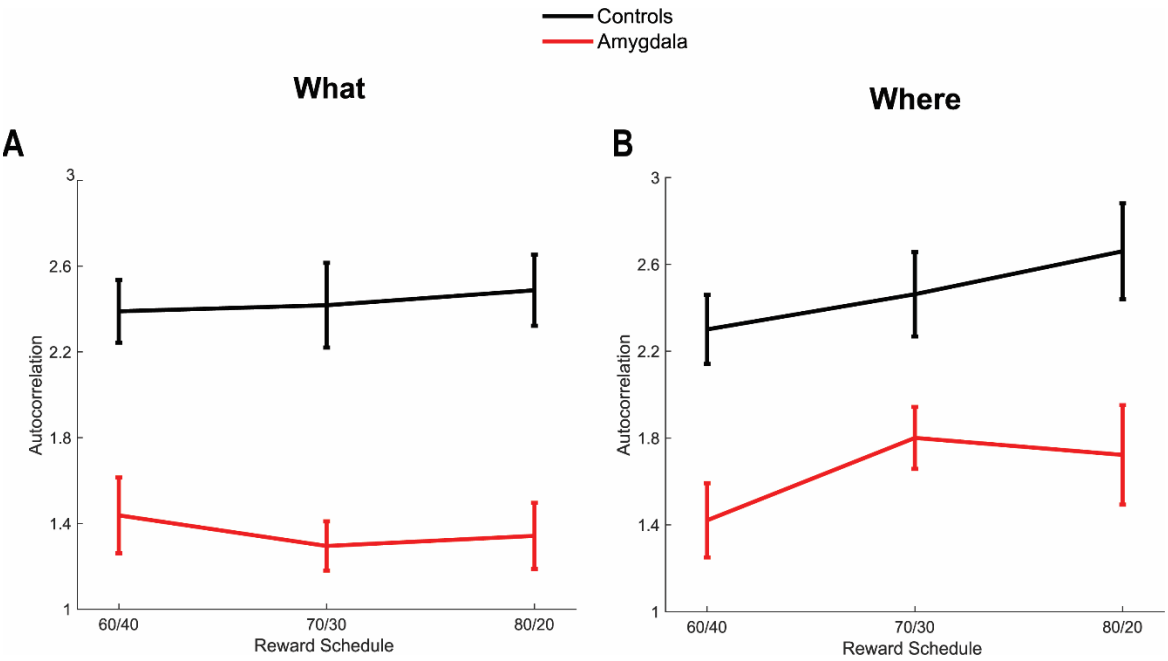


Figure 6. Autocorrelation parameter for model M2+**.** A. Autocorrelation parameter in the What condition. Error bars are +/- 1 s.e.m., where the N = the number of animals in each group (4 lesion, 6 control). B. Same as A for the Where condition.

2.3.3 Reversals

Learning in this task is governed by three processes, which may or may not map onto different neural systems. Monkeys have to infer the block type, they have to infer the correct option within each block type, and they have to reverse this preference when the outcome mapping reverses. The animals have extensive experience on the task before we collect behavioral data and, at least control animals, learn that reversals happen in the middle of the block (150). The

monkeys use the acquired task knowledge to improve performance on the task. The results above show that monkeys with amygdala lesions have deficits in both the What and Where conditions. However, it is not clear whether animals with amygdala lesions have general deficits in forming associations between actions or objects and rewards, or whether they have deficits in reversing these preferences. Therefore, we next addressed the reversal performance directly.

We used a Bayesian model to analyze the reversal behavior. The model assumes that the animals develop an initial preference for one option, and then reverse this preference at some point in the block. Because behavior is stochastic, the animals do not pick one option exclusively, and then switch at some point in the middle of the block to picking the other option. However, they tend to pick one of the options more often and this tendency switches in the middle of the block (Fig. 2). The model generates the probability that the animal reversed its choice behavior on each trial of the block – a probability distribution over reversal trial ($p(r)$, Fig. 7). On average, these probability distributions were better centered around the actual reversal points for the easier than harder reward schedules (Fig. 7A, B, D, E). We tested this by taking the value of the probability distribution on the average expected reversal trial (40). When we compared these values across both block types, we found that animals had higher values in easier schedules (Schedule; $F(2,16) = 11.4$, $p < 0.001$, $\omega^2 = 0.127$).

To characterize the distributions and examine group differences, we calculated the entropy (i.e. $\hat{h} = -\sum_{r=0..81} p(r) log p(r)$) of the posterior distribution over reversals in each block (Fig. 7C, 7F). The entropy generalizes the concept of variance to non-Gaussian distributions. It is a measure of how concentrated the distribution is around the mean or mode. Higher entropy indicates broader reversal distributions and therefore noisier, less precise reversals. When we compared the entropy across both block types, we found an overall effect of schedule on entropy (Schedule; $F(2,16) = 47.5$, $p < 0.001$, $\omega^2 = 0.107$). Therefore, the switch in choice preference was more clearly defined for the easy than hard schedules. We also found that the entropy for control monkeys was significantly lower than for the monkeys with amygdala lesions (Group; $F(1,14) = 9.4$, $p = 0.015$, $\omega^2 = 0.197$). Next, we analyzed the What and Where blocks separately. In What blocks we found an overall effect of schedule (Schedule; $F(2,16) = 25.2$, $p < 0.001$, $\omega^2 = 0.07$) and a group effect (Group; $F(1,8) = 9$, $p = 0.017$, $\omega^2 = 0.252$). In Where blocks we also found effects of schedule (Schedule; $F(2,16) = 31.1$, $p < 0.001$, $\omega^2 = 0.157$) and group (Group; $F(1,8) = 6.9$, $p = 0.030$, $\omega^2 = 0.147$).

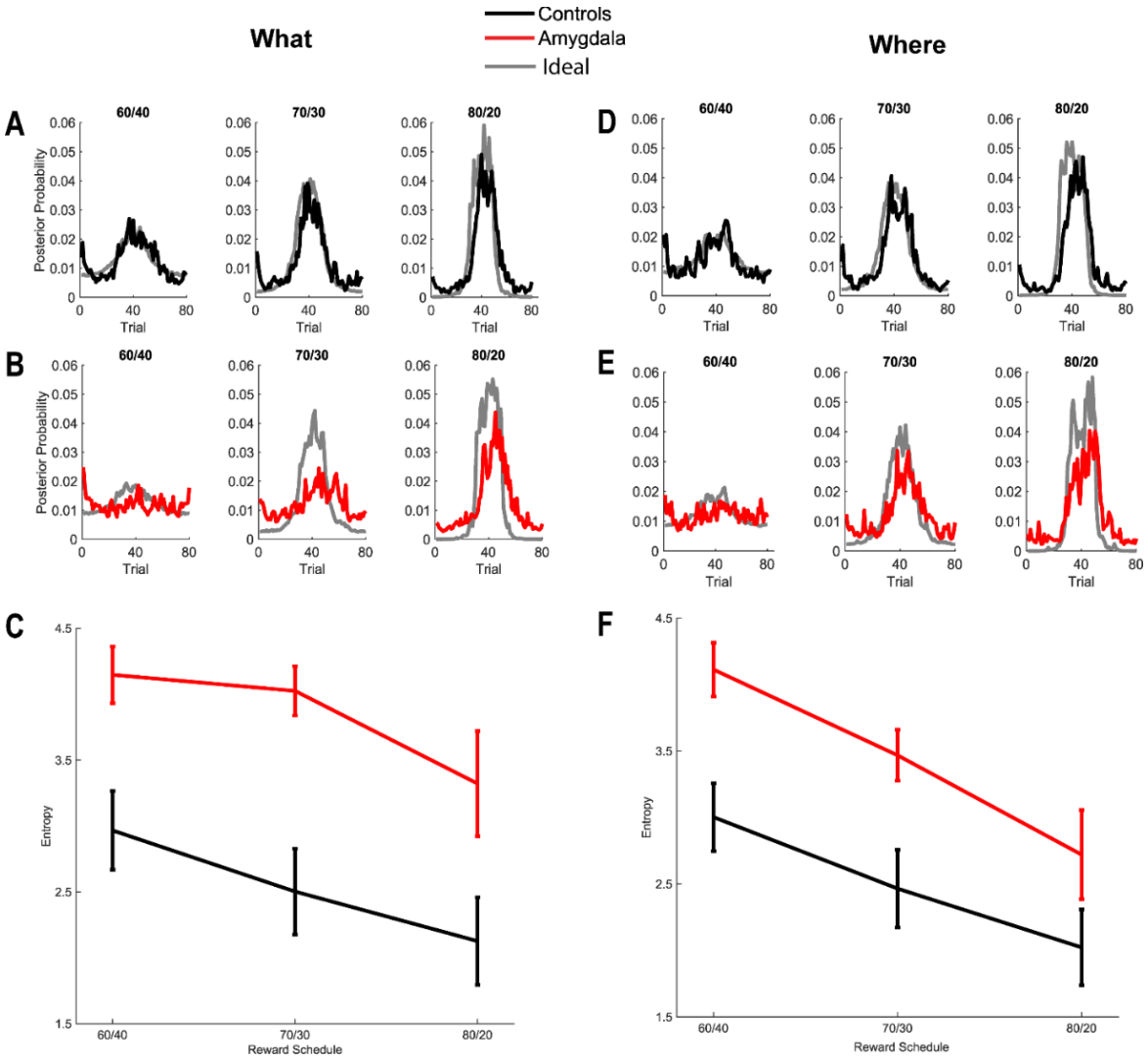


Figure 7. Posterior distributions and entropy of posterior. A. Posterior distribution for control group in the What condition, overlaid on ideal observer posterior for each schedule of the What condition. B. Same as A for the lesion group. C. Entropy of posterior distribution for both groups for the What condition, broken out by schedule. Error bars are +/- 1 s.e.m. (N = 6 control, 4 lesion). D. Same as A for the Where condition. E. Same as B for the Where condition. F. Same as C for the Where condition.

Two distributions can have different entropy but the same mean. Therefore, we next examined whether the estimated reversal trial differed between lesion and control groups, to see if the groups tended to reverse on the same trial. To do this, we calculated the expected value (i.e., the mean) of the reversal distribution in each

block (i.e. $\hat{r} = \sum_{r=0..81} rp(r)$). This gives us a single number for each block, estimating the trial on which the animal reversed its choice preference. This number can be compared to where the actual reversal occurred, which we refer to as the reversal trial difference (i.e. $\hat{r} - r_{actual}$; Fig. 8A, 8B). The variable $\hat{r}$ characterizes our estimate of where the monkey reversed and r_{actual} is the programmed reversal trial. When we analyzed both block types together, we found no effect of block type on the difference between the reversal trial of the animals and the actual reversal trial (Block Type; $F(1,8) = 0.01$, $p = 0.92$, $\omega^2 = 0$). However, there was an effect of schedule (Schedule; $F(2,16) = 17.1$, $p < 0.001$, $\omega^2 = 0.24$), with animals reversing before the actual reversal trial in the harder conditions, consistent with previous work (11, 150). There were no group differences (Group; $F(1,8) = 1.2$, $p = 0.3$, $\omega^2 = 0.033$). When we analyzed the where block by itself, we found that the groups differed in reversal behavior across schedule, reflecting the difference in the 60/40 condition (Group x Schedule; $F(2,16) = 5.2$, $p = 0.018$, $\omega^2 = 0.121$).

Following this we looked at the absolute value of the difference between the monkey reversal and the actual reversal (i.e. $|\hat{r} - r_{actual}|$; Fig. 8A, 8B). Unlike the difference in reversal trials (Fig. 8A and 8B), the absolute value of the difference (Fig. 8C and 8D) characterizes how close the animals were to the actual reversal, either before or after. As above we found an effect of schedule

(Schedule; $F(2,16) = 64.0$, $p < 0.001$, $\omega^2 = 0.35$). There were, however, no differences across block type (Block Type; $F(1,8) = 0.03$, $p = 0.87$, $\omega^2 = 0$) and there were no differences between groups (Group; $F(1,8) = 0.15$, $p = 0.71$, $\omega^2 = 0.01$). Overall, therefore, despite their generally noisier behavior, the monkeys with amygdala lesions tended to reverse on the same trial as the controls, and they were as close to the actual reversal, in absolute value.

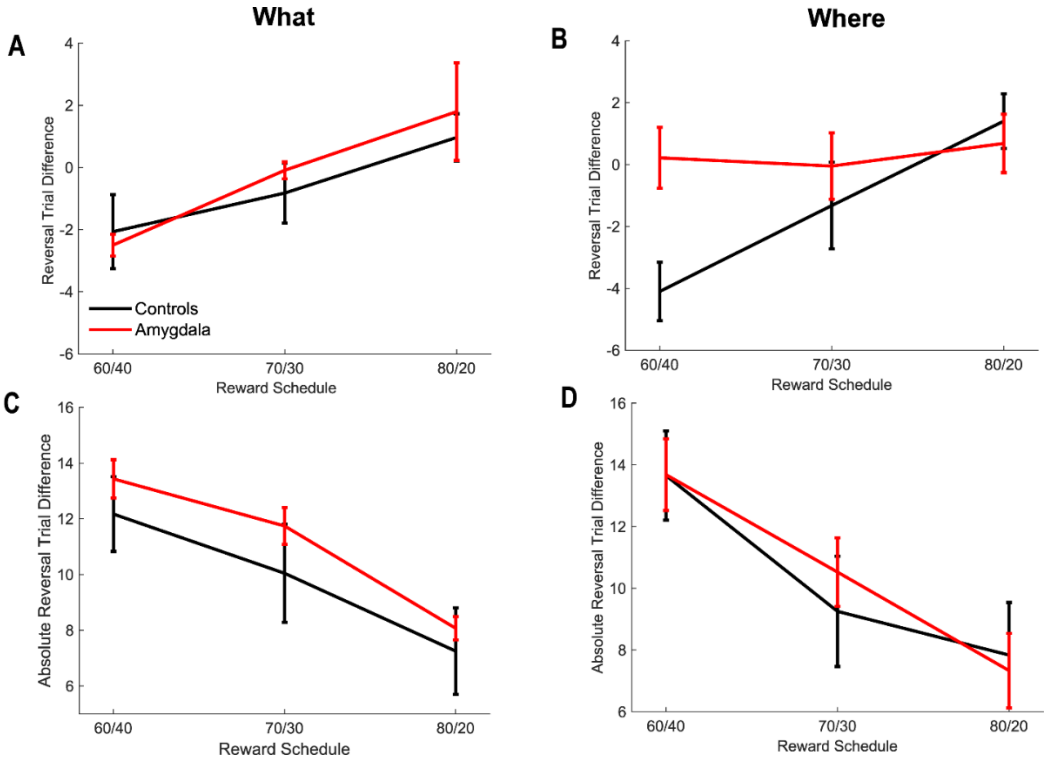


Figure 8. Relative and absolute difference in reversal behavior. Error bars are +/- 1 s.e.m. (N = 6 control animals and 4 lesion animals). A. Relative reversal trial in the What condition. The relative reversal trial is given by the difference between the point estimate of the monkey's reversal trial in each block and the actual reversal trial. Negative numbers indicate that the monkey reversed before the actual reversal trial. B. Relative reversal trial in the Where condition. C. Absolute value of the difference in the reversal trial in the What condition. In each block we computed the difference between the estimated reversal trial of the animal, and the actual reversal trial. We then took the absolute value of this difference.

The increased entropy of the reversal distribution may be driven by noisy choice behavior. The algorithm assumes that any choice not consistent with the dominant choice in a phase is possibly a reversal. Therefore, noisy choices broaden the reversal distribution. To characterize this in more detail, we first calculated the average fraction of correct choices for each animal relative to the currently most rewarded object, across the block. We then correlated the average fraction of correct choices with the average entropy of the reversal distribution (Fig. 9A), the average reversal trial difference (Fig. 9C) and the average absolute value of the reversal trial difference (Fig. 9B). The correlation between fraction correct and the entropy was large and negative across animals (ρ = -0.920, $p < 0.001$), as would be expected as entropy depends on choice accuracy. The correlation was also significant with the absolute value of the difference between the monkey and actual reversal trial (ρ = -0.773, $p < 0.01$). However, the correlation between the fraction correct and the signed difference in the reversal trial was not significant (ρ = 0.117, $p > 0.05$). The correlation between fraction correct and entropy was significantly larger than the correlation between fraction correct and the reversal trial difference ($Z = 3.19$, $p = 0.001$) but the difference between the correlation of the fraction correct and the entropy, and the correlation between the fraction correct and the absolute value, was not different ($Z = 1.05$, $p = 0.300$). It is not surprising that these correlations do not differ, because the

absolute value is related to the entropy. The entropy characterizes the width of the posterior over reversal trials, and the absolute value characterizes how far samples from this distribution are, from the mean, on average.

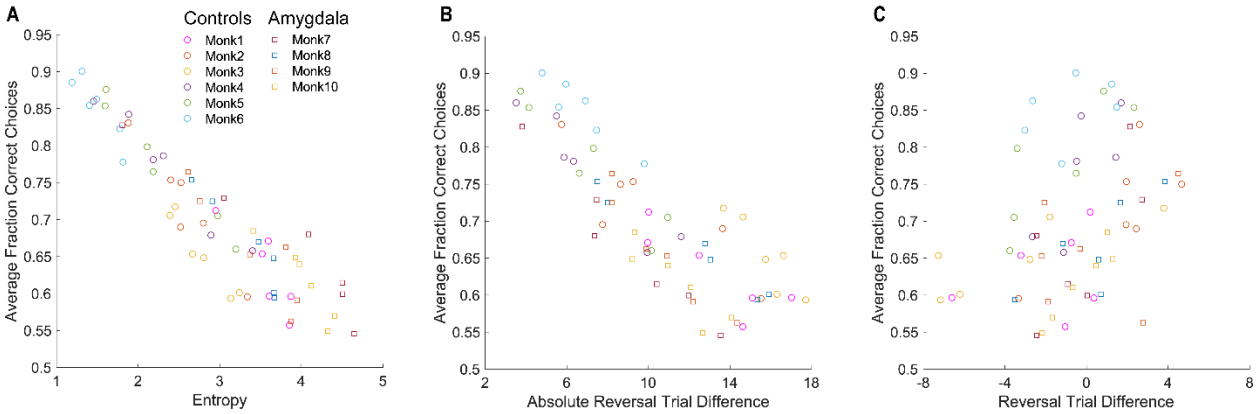


Figure 9. Correlation plots for all 10 monkeys (6 control, 4 lesion). A. The correlation between fraction correct and entropy. B. Same as A, but the correlation is between fraction correct and absolute reversal trial difference. C. Same as A, but the correlation is between fraction correct and reversal trial difference.

2.3.4 Block type

The Bayesian model also estimates whether the monkey's choices were more consistent with choosing one of the objects (What block) or one of the saccade directions (Where blocks). These estimates provide evidence for the block type the monkeys thought they were in based on their choice strategy (Fig. 10). Across conditions there was a fourth order interaction (Group x Trial x Schedule x Block-Type; $F(158,1264) = 1.3$, $p = 0.007$, $\omega^2 = 0.001$). To examine this in detail we analyzed each block-type separately. In What blocks we found that posteriors

were higher for easier schedules (Schedule; $F(2,16) = 42.7$, $p < 0.001$, $\omega^2 = 0.12$). We also found that group differences varied across schedules and trials (Group x Trial x Schedule; $F(158,1264) = 2.4$, $p < 0.001$, $\omega^2 = 0.004$). When we analyzed effects in the What blocks separately for each schedule, we found the groups differed across trials in all schedules, (Group x Trial; 60/40 $F(79, 632) = 3.5$, $p < 0.001$, $\omega^2 = 0.028$); 70/30 $F(79, 632) = 6.6$, $p < 0.001$, $\omega^2 = 0.036$); 80/20 $F(79, 632) = 1.31$, $p = 0.044$, $\omega^2 = 0.009$). In Where blocks (Fig. 10B) we found that posteriors were higher in easy schedules, reflecting increased consistency in the monkey's choice behavior (Schedule; $F(2,16) = 29.9$, $p < 0.001$, $\omega^2 = 0.117$). There were, however, no group differences. Therefore, in What blocks the monkeys with amygdala lesions were less consistently choosing one of the objects relative to the controls. However, in Where blocks the groups did not differ.

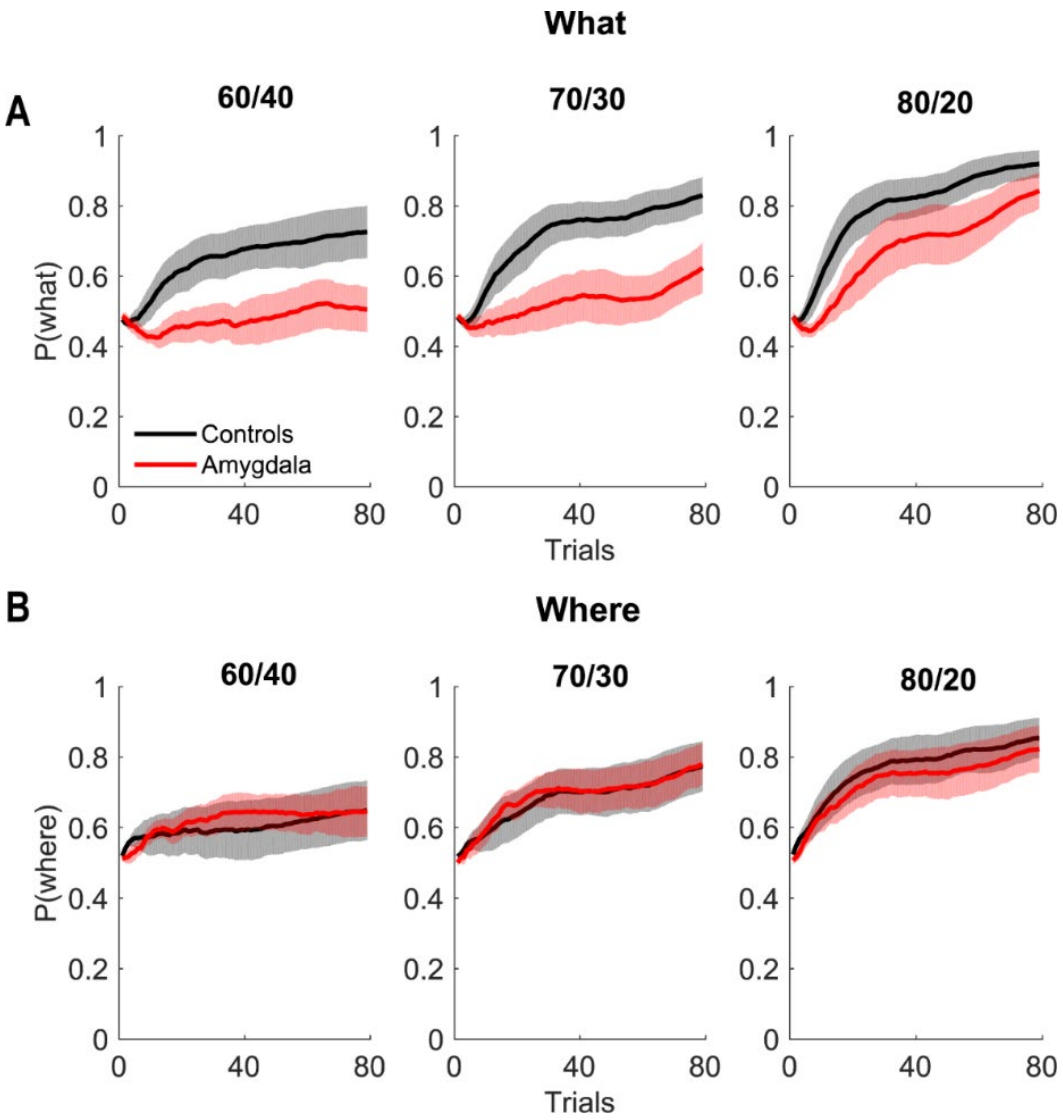


Figure 10. Posterior probability of the choice strategy used by the monkeys. A. Probability that the monkeys were using a What strategy in What blocks. A What strategy implies that the monkeys are consistently picking one of the objects. B. Probability that the animals were using a Where strategy in Where blocks. A Where strategy implies that the monkeys are consistently picking a location. Shaded region indicates +/- 1 s.e.m., where the N = the number of animals in each group (4 lesion, 6 control).

2.4 Discussion

In the present study, we found that lesions of the amygdala affected learning to select rewarding stimuli (what) and rewarding actions (where). In both block types, we found that controls more often chose the better option than the monkeys with amygdala lesions. The choice accuracy deficit in the animals with amygdala lesions was not significantly different in one block type versus the other. When we analyzed win-stay, lose-switch measures of the monkey's choices, we found that the lesioned animals more often switched after a negative outcome, which decreased performance due to the stochastic schedules. Therefore, much of their decreased accuracy, overall, followed from switching after negative outcomes, and these effects were significant in the What condition, although we did not find that the groups differed significantly across conditions. We also found that the animals with amygdala lesions tended to consistently select locations more often in What blocks, relative to control animals in harder schedules. This was consistent with the finding that animals with amygdala lesions were better fit by a model with object by location perseveration in the What condition than the Where condition. Because the perseveration term in the RL model is independent of reward, it will tend to lower performance and therefore it lowered performance relatively more in the What condition than the Where condition.

Because the operated monkeys sustained a variable amount of damage to structures adjacent to the amygdala, in addition to the substantial, planned damage to the amygdala, we considered the possibility that the behavioral impairments arose from the inadvertent, extra-amygdala damage. Notably, there was no apparent correlation of behavioral scores with the amount of inadvertent damage to a particular structure. For example, cases M2 and M3 had similar scores on acquisition yet ranked 4th and 1st among lesion subjects in extent of damage to the entorhinal cortex, respectively. In addition, based on prior work, we can be confident that the amount of damage estimated from T2-weighted scans, reported in Table S1, is an overestimate (147). These two factors militate against the possibility that extra-amygdala damage is responsible for the behavioral impairments we observed.

We also examined the reversal behavior in detail, to see whether the animals with amygdala lesions had specific deficits in reversing their choice-outcome preferences. For both block types, we found that the lesion group had higher entropy in their reversal distributions. This would be expected if the animals less consistently chose the better option in both the acquisition and reversal phase, as each time the animals choose the less preferred option, there is a small probability that they are reversing their choice preference. This is, therefore, consistent with the increased lose-switch probability of the lesioned group. When we compared

the mean and absolute values of the estimated reversal trials, we found no average differences between the groups. Therefore, the monkeys with amygdala lesions reversed, on average, as effectively as the controls. In other work we have found a correlate of the reversal inference in dorsal lateral prefrontal cortex (156), which suggests it may be playing an important role in the reversal process, although we have not yet looked for such a correlate in the amygdala. As stated earlier, learning in this task requires three processes. Monkeys have to infer the block type, they have to figure out the best choice within each block type, and they have to reverse this preference when the outcome mapping reverses. (It is possible that inferring the block type and figuring out the best option are done as one process.) Of these three processes, only the ability to consistently pick the best option was significantly impaired in monkeys with amygdala lesions, and this was primarily driven by more frequently switching after a negative outcome. The fact that our results are not statistically distinguishable across the different block types suggests that the amygdala plays a general role in forming associations between both objects and actions with rewards. Whether this is a deficit in representing the choices, the rewards, or in forming associations between them is not clear from the current results. Previous work would suggest the deficit, at least in part, is in forming the association (10, 51, 157). These data are also consistent with previous work, examining learning to reverse, which showed that monkeys with amygdala

lesions learned to reverse faster (151). The current results suggest that the faster reversals in those studies followed from weaker object outcome associations, not stronger prior probabilities on variability in the environment (151). Also relevant is the finding discerned from fMRI that, in intact monkeys, amygdala activity during both deterministic and probabilistic learning specifically predicts lose-shift behavior, and adaptive win-stay, lose-shift signals are evident in ventrolateral prefrontal cortex area 12o (157), a region necessary for probabilistic discrimination learning (158). Future studies could address whether probabilistic learning like that examined here requires the functional interaction of the amygdala with the ventrolateral prefrontal cortex.

Learning systems and their anatomical substrates can be dissociated in various ways. For example, learning is often studied using Pavlovian or instrumental paradigms (114). Formation of Pavlovian CS-US associations is mediated, to some extent, by the amygdala (51, 79). Formation of instrumental associations, on the other hand, is thought to be mediated by frontal-striatal systems (115). Both forms of conditioning were developed from purely behavioral considerations and, therefore, they do not necessarily map cleanly onto separable neural systems. Furthermore, there is considerable interaction between these behavioral processes in tasks like Pavlovian Instrumental Transfer (116) and conditioned reinforcement (117). The bandit tasks often used to study

reinforcement learning (7, 10, 159), do not map cleanly onto Pavlovian or instrumental constructs. Actions are required to select options in bandit tasks. However, when the reward values of objects are being learned, the required action varies depending on the location of the object. Furthermore, when the reward values of actions are being learned, it is likely that learning the reward values of arm movements may engage different neural systems than learning the reward values of eye movements, given the differing neural systems engaged by each type of action (160). It is also possible that learning deficits following amygdala lesions may depend on the type of motor response required to register choices. Additional work will be required to clarify this hypothesis.

From a psychological perspective, it is of interest that the amygdala is essential for both object-outcome and action-outcome associations as assessed with devaluations tasks (98, 161). Together with the present data, these findings show that the amygdala is important for learning about both objects and actions as they relate to reward probability (present study) and current reward value, including reward magnitude (98, 139, 161). An earlier study on object reversal learning found that, relative to unoperated controls, monkeys with amygdala lesions benefitted more from correct choices that follow an error in a deterministic setting (162). While this finding would seem to be at odds with the present findings, we note that the object reversal tasks differ in more ways than use of deterministic vs.

probabilistic outcomes. For example, the standard object reversal learning task employed by Rudebeck and Murray employed a small number (nine) of reversals whereas in the present study, all monkeys had received extensive training in reversals. As a result, unlike in the present study, monkeys in the deterministic task experienced unexpected uncertainty, at least in early reversals. In addition, the present and earlier task differ in the type of response required (manual vs. eye-movement), in the location and type of reward (food reward under object vs. fluid reward delivered to mouth), and in the number of trials administered per session (30 trials vs. Massed trials). These task differences might account for the somewhat different picture gained from assessing amygdala contributions to the two kinds of reversal learning. Thus, the amygdala makes an essential contribution to reversal learning in probabilistic and even deterministic settings with massed trials in an automated apparatus (10) but not to reversal learning in deterministic settings with a small number of trials in a manual test apparatus (25).

The What vs. Where task used in the current study was developed to separate neural circuits underlying learning rewards associated with objects whose locations vary, vs. learning to associate rewards with actions independent of the object at the saccade target location. The hypothesis that such a dissociation should be possible follows from work in the visual and auditory systems (119, 163) based on the separable anatomical organization of visual cortex (119, 120), as well

as proposed frontal extensions of this circuitry (164). Parietal cortex processes information about the spatial locations of objects in the environment, and the motor actions required to interact with these objects (165-167). The ventral visual cortex, on the other hand, processes information about object features that allow for object identification and discrimination (168, 169). This separable organization continues into prefrontal cortex (121, 122, 164, 170), and correspondingly into the striatal circuitry (123, 124, 127). While there is evidence for anatomical segregation, neurophysiological recordings have shown integration of What and Where information in both prefrontal (171) and parietal (125) cortex. Thus, in neural circuits that are less proximal to sensory processes it is not clear that these separate streams differentially process behaviorally relevant information, particularly in frontal-striatal systems.

The task was developed to separate learning about actions vs. learning about objects. However, there are other differences between the conditions that may lead to behavioral effects. For example, in the What condition, the preferred object is present on the screen, whereas in the Where condition, although the response zones are indicated on the screen, the animals have to internally generate the action that will most likely lead to reward. Still, there is no reason to think monkeys with amygdala lesions are impaired in the ability to internally generate actions. An earlier study that examined the effects of amygdala lesions on conditional motor

learning found no impairment in learning new conditional problems in a deterministic setting, even though the responses were internally generated (172). It is also possible that the animals find one or the other condition to be more difficult. We have not systematically studied this, but in other work we have seen that some animals do better in either the Where relative to the What (11) or What relative to Where conditions (156, 173). Therefore, animals do not consistently show a clear preference for one or the other condition.

We have previously shown that learning oculomotor action sequences depends on a dorsal-lateral prefrontal, dorsal striatal circuit (132, 133). The prefrontal and striatal nodes in this circuit processed sequence related information (132), and local injections of dopamine antagonists into the dorsal striatum led to deficits in performance during sequence learning (133). The dorsal striatum also contains a stronger representation of action value than prefrontal cortex (130, 131). This suggests that the dorsal circuit is important for action learning, when actions are eye movements. We have also found that lesions of the ventral striatum yield deficits specific to learning to select rewarding objects, without affecting learning to select rewarding actions, using the same What vs. Where task used here (11). In other work we found that amygdala lesions affect learning to choose rewarding stimuli (10, 23). We have not, however, carried out a double dissociation

experiment, using the same task with manipulations of either the dorsal or ventral striatum.

Anatomy is often a guide to function. Anatomically, the basolateral amygdala is strongly interconnected with the ventral, visual object system (127, 140, 174). It receives substantial projections from high level visual cortex (140), and correspondingly projects to the ventral striatum (175), and ventrolateral and orbital prefrontal cortex (176, 177). Both prefrontal areas also receive input from temporal lobe, and not parietal lobe, visual areas (122). The amygdala also interacts with the medial portion of the mediodorsal thalamic nucleus (178), which also projects to orbitofrontal cortex (179).

Given the anatomical connections of the amygdala with the ventral visual pathway, we had hypothesized that it would be mostly related to learning to choose objects and not actions. However, we found that lesions of the amygdala led to deficits in both learning to select actions and objects. Numerically the choice accuracy effect was larger for selection of objects. Although the amygdala has minimal connectivity with dorsal prefrontal areas that underlie oculomotor control (177) and no connections to the dorsal striatum, it does have substantial back-projections across the visual hierarchy, including early visual cortex (140). Given that early visual areas have minimal bilateral visual representation, these back-projections may affect lateralized spatial representations. Furthermore,

neurophysiology studies have shown that amygdala neurons contain representations of the spatial locations of rewarded objects (142, 143, 180-182). These spatially selective responses are also in contrast to the ventral striatum, which contains no spatial information (142). Therefore, neurophysiology suggests a possible role for the amygdala in spatial-attentional processes, and the effects of these representations on behavior may be mediated by back-projections to early visual areas, or other pathways that connect these representations, polysynaptically, to areas that underlie eye movements. In related work, the amygdala is also involved in revaluing arm-motor responses in devaluation paradigms, where the value associated with a specific motor response changes following a selective satiation procedure (161).

Amygdala interactions with orbitofrontal and ventrolateral prefrontal cortex are also likely important for the learning processes we have examined (158, 183). Recent work in rats has shown that ablation of amygdala neurons that project to the OFC impairs reversal performance on a probabilistic spatial learning task (184). This deficit was due to the rats losing their ability to use positive outcomes to guide their choice behavior. In this same study it was shown that ablation of OFC neurons projecting to amygdala enhanced reversal performance by destabilizing action values. Related to this, it has been shown that lesions of the OFC impair reversal behavior, but subsequent lesions of the amygdala in the same animals

restore performance (185). These results suggest that, under some circumstances, amygdala-OFC interactions may be detrimental to learning.

Our Bayesian model assumes that a state inference process underlies reversal learning. The model assumes the animals have a preference for one option (i.e., state A) which reverses at some point in the block (i.e. state B). The posterior distribution over reversals is the estimate of where the animals switch states in each block. Our data suggests that the amygdala is not involved in inferring state or state switches in our task. Although monkeys with amygdala lesions have deficits in consistently choosing the better option in both conditions, they reverse their choice preference as well as controls. Therefore, it is possible that state inference processes are being carried out by prefrontal cortical areas, as suggested by previous work (156, 186-192). The interaction between cortical state-inference processes, and amygdala learning processes, may lead to deficits in some conditions, when there is a conflict between the best choices predicted by each process. Under these conditions, lesions of the OFC to amygdala pathway may improve performance. Future work can examine this possibility in more detail.

2.4.1 Conclusion

We found that lesions of the amygdala led to deficits in consistently choosing the more frequently rewarded options. We found these deficits in both the What condition, when animals had to learn to choose the best visual object, and in the Where condition, when the animals had to learn to choose the best action. The deficits in choice accuracy followed primarily from switching after a negative outcome, which led to decreased performance due to the stochastic schedules. We did not find deficits in reversal accuracy; thus, monkeys with amygdala lesions were able to reverse their choice-outcome mappings, in both conditions, as well as controls. Inferring reversals in choice-outcome mappings may, therefore, be more dependent on other brain areas, including cortex. Overall, this suggests that the amygdala is important for consistently choosing a rewarded option. Future work should focus on understanding how the network of areas that are important for learning orchestrates the multiple processes involved in learning in dynamic environments.

2.5 Supplemental Material

	Left				Right			
Monkey	Percent Estimated Damage from (Amygdala)	Percent Estimated Damage from (Entorhinal Cortex)	Percent Estimated Damage from (Perirhinal Cortex)	Percent Estimated Damage from (Hippocampus)	Percent Estimated Damage from (Amygdala)	Percent Estimated Damage from (Entorhinal Cortex)	Percent Estimated Damage from (Perirhinal Cortex)	Percent Estimated Damage from (Hippocampus)
M1	97.21%	68.92%	60.26%	27.14%	92.01%	46.18%	11.01%	24.99%
M2	81.57%	10.69%	8.15%	3.25%	90.84%	4.00%	0.28%	3.22%
M3	90.30%	42.62%	34.33%	14.26%	92.02%	81.25%	34.89%	41.89%
M4	89.61%	55.76%	32.08%	41.35%	85.14%	57.68%	27.12%	14.38%
Mean	89.67%	44.50%	33.71%	21.50%	90.00%	47.28%	18.33%	21.12%
SD	6.40%	24.97%	21.30%	16.44%	3.29%	32.33%	15.61%	16.45%

Figure S1. Estimates of the volume of lesion damage for intended (amygdala) and unintended (entorhinal cortex, perirhinal cortex, and hippocampus) brain structures.

Chapter 3: The Ventral Striatum's Role in Learning from Gains and Losses

3.1 Introduction

Adaptive behavior requires that organisms choose wisely to gain rewards and avoid punishment. Reinforcement learning refers to the behavioral process of learning about the value of choices, based on choice outcomes. From an algorithmic point of view, rewards and punishments exist on opposite sides of a single value axis. Simple distinctions between rewards and punishments, however, and their theoretical expression on a single value axis, hide the considerable complexities that underlie appetitive and aversive reinforcement learning. Most notably, both rewards and punishments come in many forms. Food, sex and ascending the social hierarchy are rewarding. Correspondingly, loss of cached food, pain and social defeat are punishing (193). Whether threat, pain, and loss of accumulated reward drive learning via the same neural systems, at any level, is unclear. Furthermore, even when gains and losses are expressed with money, which has objective value, they can have differential subjective effects on behavior (194-196).

Studies of reinforcement learning (RL) often use paradigms in which participants learn to choose options on the basis of reward frequency or reward

magnitude (10, 19, 159). These studies have shown that the striatum, and the dopamine input to the striatum, underlies learning to select rewarding options. Theoretical models of RL extend directly to learning from losses, and therefore striatal mediated learning may generalize to these conditions (4). This hypothesis is supported by work that has shown that dopamine neurons, which provide reward prediction error (RPE) signals to the striatum, increase their firing rates when rewards are unexpectedly delivered and decrease their firing rate when rewards are unexpectedly omitted (2, 5). However, some studies have explicitly examined learning from gains and losses (as opposed to reward omission) and found that they are mediated by partially overlapping, but partially distinct systems that cross cortical and subcortical circuits. For example, single neuron studies in macaques have shown that the dorsolateral prefrontal cortex, as well as the anterior cingulate cortex, encode both losses and gains in a competitive game in which conditioned reinforcers could be gained and lost (17). In other work, the medial orbitofrontal cortex was found to encode gains and avoidance of losses, both of which have positive value (18). This study also found that appetitive RPEs in reward trials (i.e. increases with unexpected rewards) correlated with the extent of activation in the ventral striatum (VS), whereas RPEs in aversive trials (i.e. increases with unexpected punishments) correlated with activation in the insula, consistent with other work (19). In addition to the work in macaques and humans, work in

rodents, which has used various paradigms including conditioned place aversion and Pavlovian threat of shock, has shown that basolateral amygdala circuits through the VS encode reward mediated approach behavior, whereas circuits through the central nucleus of the amygdala encode avoidance (197-199). Related experiments focusing on circuitry have found that dopamine inputs to the IL/PL regions of medial prefrontal cortex also encode avoidance behavior (200). Thus, there is evidence that both overlapping and distinct systems underlie learning from rewards and punishments, using some paradigms.

To examine the role of the VS in learning from both gains and losses, we adapted a previously used token reward system (17) to two-armed bandit RL tasks. In the tasks, rhesus monkeys made choices among options, and received tokens for their choices. The tokens were represented by circles on the bottom of the screen and the animals periodically received juice in exchange for accumulated tokens. The use of tokens, which are secondary reinforcers, allowed us to study the effects of gains and losses on choices using one and the same unit of value. We ran four variants of the task to address specific questions. Three variants used deterministic outcomes, and one used stochastic. We compared the behavioral performance of 3 monkeys with lesions of the VS and 4 unoperated controls.

3.2 Methods

3.2.1 Subjects

The subjects included 6 male and 1 female rhesus macaques with weights ranging from 6-11 kg. Three of the male monkeys received bilateral excitotoxic lesions of the VS. The remaining four monkeys served as unoperated controls (3 males and 1 female). One of the male control animals was not able to complete all 4 tasks, and therefore task 4 only has 3 controls. For the duration of the study monkeys were placed on water control. On testing days monkeys earned their fluid from their performance on the task. Experimental procedures for all monkeys 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.

3.2.2 Surgery

Three monkeys received two separate stereotaxic surgeries, one for each hemisphere, which targeted the VS using quinolinic acid. After both lesion surgeries, each monkey received a cranial implant of a titanium head post to facilitate head restraint. Unoperated controls received the same cranial implant.

Behavioral testing for all monkeys began after they had recovered from the implant surgery. Lesioned animals were used in three previous studies (10, 11, 201).

3.2.3 Lesion assessment

Lesions of the VS were assessed from postoperative MRI scans. We evaluated the extent of the damage with T2-weighted scans taken after the initial surgeries. For the lesioned monkeys, MR scan slices were matched to drawings of coronal sections from a standard rhesus monkey brain at 1 mm intervals. We then plotted the lesions onto standard sections.

3.2.4 Task

We tested rhesus macaques on three deterministic and one stochastic two-arm bandit learning task. We conditioned tokens as reinforcers, which allowed us to assess learning from both gains and losses within the same dimension. All animals completed the 4 tasks in the same order. Each experimental session was composed of nine novel and three familiar blocks that were randomly interleaved. In each novel block we introduced images the animal had never seen before and they had to learn the cue-outcome associations. The images in the familiar blocks were kept constant for the duration of a task. We completed testing on each task before beginning the next.

During the experiment the animals were seated in a primate chair facing a computer screen. Eye movements were used as behavioral readouts. In each single trial, the animals first acquired fixation (Fig. 1A). After a fixation hold period, we presented two images, left and right of fixation. The animals made an eye movement to one of the images to indicate their choice. They were allowed to make an eye movement as soon as the targets appeared. After a hold period, the number of tokens associated with their choice was added or subtracted from their accumulated tokens. Every 4-7 trials, with the interval randomly selected, the animals received one drop of juice for each token they had at the time of cash-out. When each drop of juice was delivered one of the tokens disappeared from the screen.

TkD- Token task 1 (Deterministic learning)

In the first task (TkD), novel blocks consisted of 108 trials and familiar blocks of 36 trials. Novel blocks consisted of 4 images the animals had never seen before. Associated with each image was a value (+2, +1, -1, -2), such that if that image was chosen, the animal gained or lost the corresponding number of tokens. On each trial monkeys had to acquire and hold central fixation for 500 ms. After monkeys held central fixation two of the images would appear to the left and the right of the fixation point. The animal chose one by making a saccade to the image and holding for 500 ms. The number of tokens associated with the image was then

added or subtracted from their total count, represented by circles at the bottom of the screen. The animals could not have less than zero tokens, however. Therefore, if they had one token and they chose a -2 image, they were reduced to 0 tokens. Every 4-7 trials their tokens were cashed out. At cash-out, the animals were given one drop of juice for each token. When each drop of juice was delivered, one token was removed from the screen. There were six individual conditions in this task, defined by the possible pairs of images. The conditions within a block of 108 trials were presented pseudo-randomly. The animals saw each condition twice, once on the left and once on the right, every twelve trials before seeing any condition a third time. At the end of each 108 trial block we introduced 4 new images and the animals began the learning from scratch.

NtK- Token task 2 (Deterministic learning)

In the second task we included an image in the set which, if chosen, led to no change in the number of tokens. Thus, at the beginning of each novel block we introduced 5 new images. The images had associated token outcomes of +2, +1, 0, -1 and -2. There were, therefore, 10 different pairs of objects which we refer to as conditions. These were administered in blocks of 120 trials. Each pair of images was seen twice every 20 trials, with each image presented once on the left and once on the right. As before, the conditions were randomly interleaved within a mini-block.

TkS – Token task 3 (Stochastic learning)

In the third task, we examined performance when feedback was stochastic. In this task, at the beginning of each block we introduced 4 new images with associated reward magnitudes of +2, +1, -1 and -2. The design was otherwise the same as the first token task. Except, in this task, in 75% of the trials the number of tokens was adjusted by the magnitude associated with the chosen option, but in 25% of the trials there was no change in the number of tokens. This makes learning more difficult, and information has to be integrated across a larger number of trials to learn the correct choice.

TkL- Token task 4 (Deterministic learning)

In the final task, we again used deterministic feedback to examined performance. This version of the task was similar to task 1 (TkD) with two differences. First, we changed the value of the -2 cue to -4. So the cues for this task were +2, +1, -1, -4. Second, we gave the animals an endowment of 4 tokens after every cash-out to maintain motivation, and to increase the number of trials on which they would experience the actual 4 token loss.

In some trials the animals had zero tokens and chose a loss cue. In this case they had no change in tokens. Therefore, the animals would know that they had not chosen a gain token, but they would not know the magnitude of the loss. In task 1

this happened 12.5% of the time, in task 2 17.5% of the time, in task 3 13.2% of the time and in task 4 2.6% of the time.

3.2.5 Images & eye tracking

Images provided as choice options were normalized for luminance and spatial frequency using the SHINE toolbox for MATLAB (148). All images were converted to grayscale and subjected to a 2D FFT to control spatial frequency. To obtain a goal amplitude spectrum, the amplitude at each spatial frequency was summed across the two image dimensions and then averaged across images. Next, all images were normalized to have this amplitude spectrum. Using luminance histogram matching, we normalized the luminance histogram of each color channel in each image so it matched the mean luminance histogram of the corresponding color channel, averaged across all images. Spatial frequency normalization always preceded the luminance histogram matching. Each day before the monkeys began the task, we manually screened each image to verify its integrity. Any image that was unrecognizable after processing was replaced with an image that remained recognizable. Eye movements were monitored and the image presentation was controlled by PC computers running the Monkeylogic (version 1.1) toolbox for MATLAB (149) and Arrington Viewpoint eye-tracking system (Arrington Research).

3.2.6 Reinforcement learning models

We fit a large set of models that varied in the number of parameters they used to model the conditions. In the results we focus on 4 models that most often accounted for behavior. All models were built around a Rescorla-Wagner, or stateless RL value update equation given by:

(1) $$v_i(t+1) = v_i(t) + \alpha_j\big(R - v_i(t)\big).$$

These values were then passed through a soft-max function to give choice probabilities for the pair presented in each trial:

(2) $$d_j(t) = (1 + e^{\beta_k\left(v_i(t) - v_j(t) + h_i(t) - h_j(t)\right)})^{-1}, d_i(k) = 1 - d_j(k).$$

The variable v_i is the value estimate for option i, R is the change in the number of tokens that followed the choice in trial t, and α_j is the condition dependent learning rate parameter, for condition j. In addition, we also used, for some models, condition dependent values of the choice consistency or inverse temperature parameter, β_k. The variable $h_i(t)$ implemented a choice autocorrelation function (202), which increased the value of a cue that had occurred in the same location, recently. The autocorrelation function was defined as:

(3) $$h_i(t) = \kappa e^{-\lambda(t - t_{l(i)})},$$

Where the variables, κ and λ were free parameters scaling the size of the effect and the decay rate, respectively. The variable, $t_{l(i)}$, indicates the last trial on which a given cue, i, was chosen in a given location. There were eight separate values for $t_{l(i)}$ as it tracked the four cues across locations, except for Task 2 (TkN) which had ten values.

We then maximized the likelihood of the animal's choices, *D*, given the parameters present in the model under consideration, using as a cost function:

$$f\left(D\middle|\alpha_j, \beta_k, \kappa, \lambda\ \right) = \prod_t [d_1(k)c_1(k) + d_2(k)c_2(k)]. \quad (4)$$

Where $c_1(k)$ was an indicator variable that took on a value of 1 if option 1 was chosen and zero otherwise, and $c_2(k)$ took on a value of 1 if option 2 was chosen and 0 otherwise.

The VALENCE model had one inverse temperature, and two learning rates, 1 for positive cues and one for negative cues. The CUE model had one inverse temperature, and one learning rate for each cue. Note that the null cue in Task 2 would always have a 0 reward prediction error, because the reward associated with this cue was 0, and its values started at 0. Therefore, it does not need a learning rate.

We also explored addition models that had: i. one inverse temperature and one learning rate, ii. two inverse temperatures, one for the loss-loss condition and one for the rest of the conditions and two learning rates, one for positive outcomes

and one for negative outcomes. iii. two inverse temperatures, one for the 2 v 1 condition, and one for the rest of the conditions, and two learning rates, one for positive feedback, one for negative feedback. None of these models predicted behavior well, however, so to simplify presentation we do not show their results.

3.2.2 ANOVA models

To quantify differences between choice behavior in each group, we performed an arcsine transformation on the choice accuracy values from each session, as this transformation normalizes the data (152). We then carried out an N-way analysis of variance (ANOVAN). Monkey and session were included as random effects with session nested under monkey. All other factors were fixed effects. The ANOVA on learning rate parameters across experiments was also done as a mixed effects ANOVA with session and monkey as random effects and experiment and cue as fixed effects.

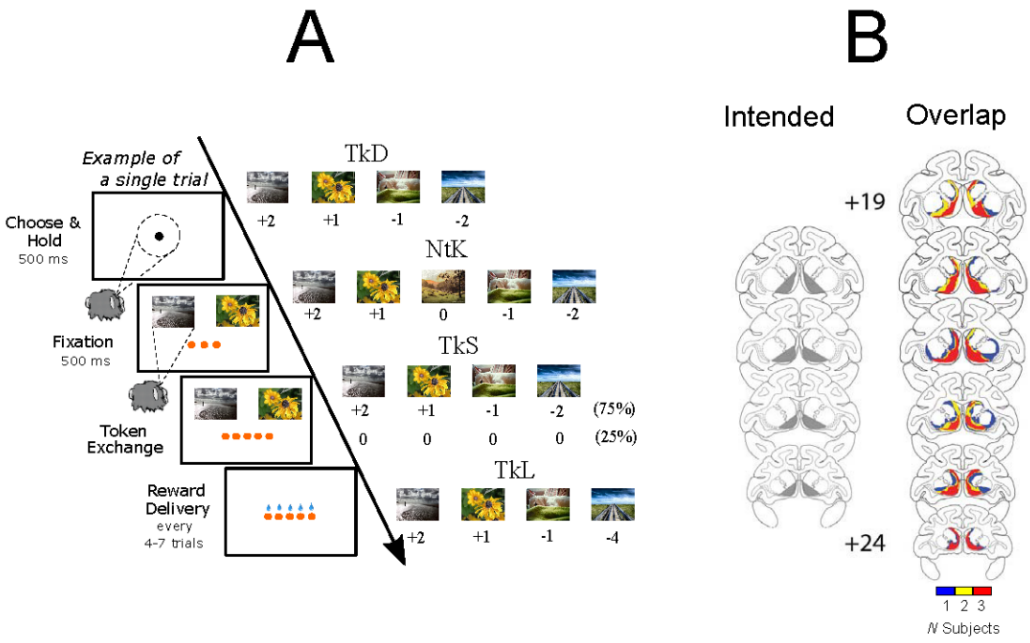
All within group post-hoc analysis of aborted trials and reaction time was done using the multcompare function in MATLAB, specifically using the Bonferroni method. Unless otherwise stated, multcompare within group stats will only be reported for the condition that is the least significant.

3.2 Results

The monkeys were run on a series of four tasks. In each task, trials involved a forced choice between two images. Selection of a particular image led to increases or decreases in accumulated tokens (Fig. 1A). The outcome of each trial following a choice was realized on the monitor screen as a change in the number of tokens the animal had accumulated. Every four to seven trials, with the interval chosen randomly, we cashed out the accumulated tokens. During cash-out the monkeys received one drop of juice for each token. The animals had to learn over trials to select the image from the pair that maximized their gains and minimized their losses. When the monkeys had no tokens and they chose a loss cue there was no change in the tokens. The animals could also not incur negative token counts.

The tasks were run in a fixed sequence (Fig. 1A). Each task evaluated the monkeys' choices during learning and performance on novel or familiar stimulus-outcome associations. In the novel blocks, the monkeys learned stimulus-reward associations for a novel set of images. In the familiar blocks (See SI Appendix Figs. S1-S4), the monkeys chose between stimuli they had repeatedly sampled over the course of prior experimental sessions. The stimulus-outcome associations of these familiar choice options were fixed for the duration of the experiment. Novel and familiar blocks were randomly interleaved each day. The novel blocks allowed us to examine the rate at which cue-reward associations were learned,

whereas the familiar blocks allowed us to examine asymptotic performance with overlearned cue-reward associations.



3.2.1 Choice Behavior

TkD

We first evaluated the ability of the monkeys with or without lesions of the VS to learn deterministic stimulus-outcome associations. In this task, at the beginning of each novel block, the monkeys encountered four images they had not seen before. Each image was associated with a fixed, deterministic gain or loss of tokens (+2, +1, -1 or -2 tokens). Two of the images were presented as choice options on each trial. This resulted in six unique pairs of images, which we refer to as conditions. In each block, conditions were randomly interleaved over intervals of twelve trials until each condition occurred 18 times in novel (Fig. 2) blocks and 6 times in familiar (See SI Appendix Fig. S1) blocks.

In novel blocks, the monkeys learned the stimulus-outcome associations efficiently. With experience, they were able to choose the better option of the pair on a high proportion of trials (Fig. 2). There were differences in performance across conditions (Condition; $F(5, 20) = 140$, $p < 0.001$) and differences in performance across trials in the different conditions (Condition x Trial; $F(85, 38) = 5.7$, $p < 0.001$). The monkeys performed best in the conditions in which there was a loss paired with the largest reward. For example, they most often picked the best cue when choosing between the +2 and -1 and +2 and -2 conditions. This effect was driven largely by the frequency with which they experienced the outcomes

associated with each cue and the differences in the values of the cues. The animals most frequently picked the +2 cue across all conditions, and therefore most frequently received feedback on its value, and the value of this cue would also asymptote at +2.

In task 1, there were no differences between groups (Group; $F(1, 9) = 0.1$, $p = 0.7611$) and no differences between the groups across conditions (Group x Condition; $F(5, 22) = 0.5$, $p = 0.7801$). The monkeys did not perform well in the -1 v -2 condition, although across the groups there was a significant positive correlation between choice accuracy and trial, which indicates learning ($t(6) = 9.1$, $p < 0.001$). When we examined the groups individually, we found that both groups learned to choose the smaller loss more often with experience (Control: $t(3) = 6.9$, $p = 0.006$), VS: $t(2) = 13.4$, $p = 0.005$).

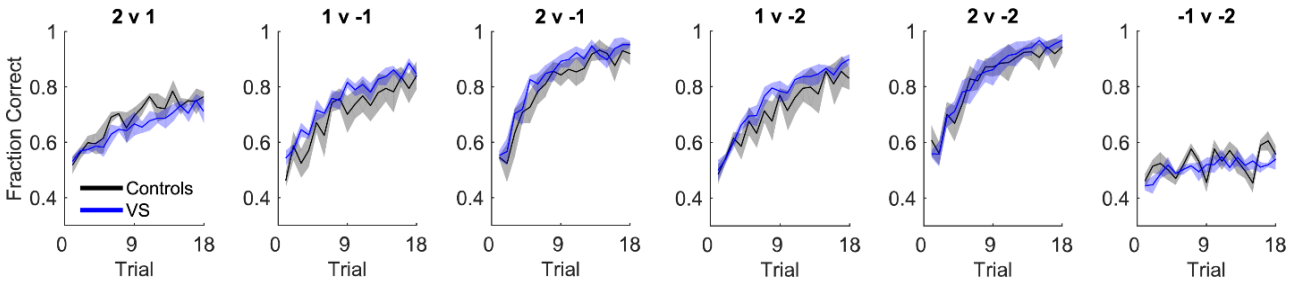


Figure 2. TkD Deterministic Reinforcement Learning of Stimulus-Outcome Associations. Task 1 choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

NtK

In task 2, we used five cues in each block with cue-outcome mappings of +2, +1, 0, -1 and -2. This resulted in 10 pairs of cues and therefore 10 conditions. In task 2, both the novel (Fig. 3) and familiar (See SI Appendix Fig. S2) blocks were composed of 120 trials, 12 per condition. Therefore, in the novel blocks the monkeys saw each pair of cues 12 times. Inclusion of the null cue allowed us to test two specific hypotheses. First, does the absolute difference between the value of the cues drive performance independent of the reward value associated with the cues? Second, can animals learn to select the null cue when it is paired with a loss cue?

In the novel blocks (Fig. 3), there was again a difference in performance across conditions (Condition; $F(9, 19) = 54.7$, $p < 0.001$) and also a difference in performance across trials in the different conditions (Condition x Trial; $F(99, 398) = 8.2$, $p < 0.001$). There were no differences between groups (Group; $F(1, 9) = 0.1$, $p = 0.778$) and no differences by condition (Group x condition; $F(9, 14) = 0.6$, $p = 0.793$). There was also no difference between groups when we examined only the 2 v 1 condition (Group; $F(1,5) = 3.6$, $p = 0.117$).

Similar to task 1, when we grouped all the animals together there was a significant correlation between trial and performance when the animals had to choose between the two loss cues ($t(6) = 3.3$, $p = 0.016$). However, when we

separated the groups, neither group reached significance alone (Controls: $t(3) = 2.3$, $p = 0.103$; VS: $t(2) = 3.1$, $p = 0.092$). There was also significant learning when the animals had to choose between the 0 and -1 cue across groups, but not in either group individually ($t(6) = 3.9$, $p = 0.007$; Controls: $t(3) = 2.9$, $p = 0.062$; VS: $t(2) = 2.1$, $p = 0.164$). When the animals had to choose between the 0 and -2 cues there was learning across groups ($t(6) = 5.7$, $p = 0.001$). However, when we examined the groups separately we found that only the controls performed significantly better than chance (Controls: $t(3) = 3.5$, $p = 0.037$; VS: $t(2) = 4.1$, $p = 0.053$).

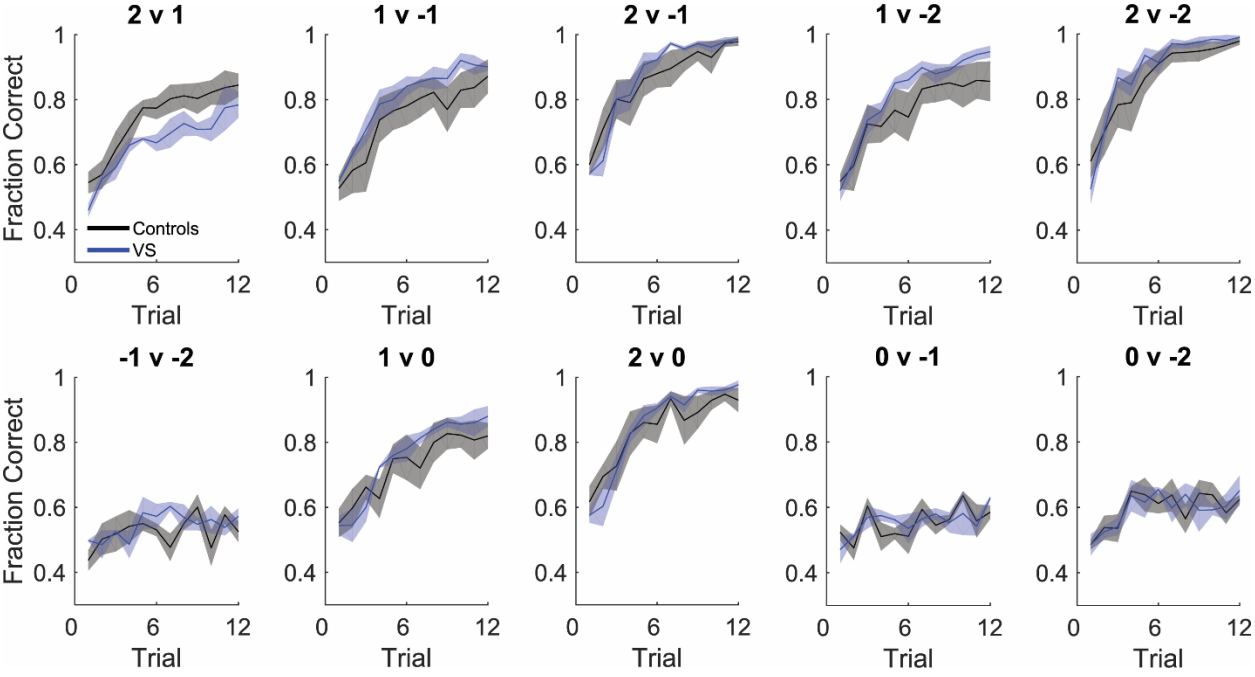


Figure 3. NtK Deterministic Reinforcement Learning Augmented by a Null Cue. Task 2 choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

TkS

In task 3, we introduced four cues with cue-outcome associations of +2, +1, -1 and -2. However, the cue-outcome associations were stochastic. Therefore, when the animals chose one of the options, they received the outcome associated with that option in 75% of the trials, and no outcome (i.e. no change in tokens) in 25% of the trials. We introduced this task because we have previously seen that monkeys with VS lesions learn poorly under stochastic schedules (10), and the VS may be more important for slow learning, which is more affected by trial-by-trial stochasticity (203). Both novel (Fig. 4) and familiar (See SI Appendix Fig. S3) blocks were 108 trials.

Performance was consistent with the previous tasks (Fig. 4). In the novel blocks, there was a difference in performance across conditions (Condition; $F(5, 38) = 315.0$, $p < 0.001$) and learning also differed across trials in the different conditions (Condition x Trial; $F(85, 135) = 7.4$, $p < 0.001$). In addition to these effects, and unlike the case for the tasks with deterministic outcomes, there was an overall effect of group (Group; $F(1, 70) = 15.2$, $p < 0.001$). When we examined differences between groups in each condition, we found that the 2 v 1 condition approached significance, but this did not survive correction for 6 comparisons (Group; $F(1,5) = 6.73$, $p = 0.049$). In this task monkeys showed learning when choosing between the -1 and -2 cues (All animals: $t(6) = 2.7$, $p = 0.037$). When we looked at the groups separately we found that only controls learned to choose the smaller loss, (Control: $t(3) = 3.3$, $p = 0.045$; VS: $t(2) = 0.7$, $p = 0.523$).

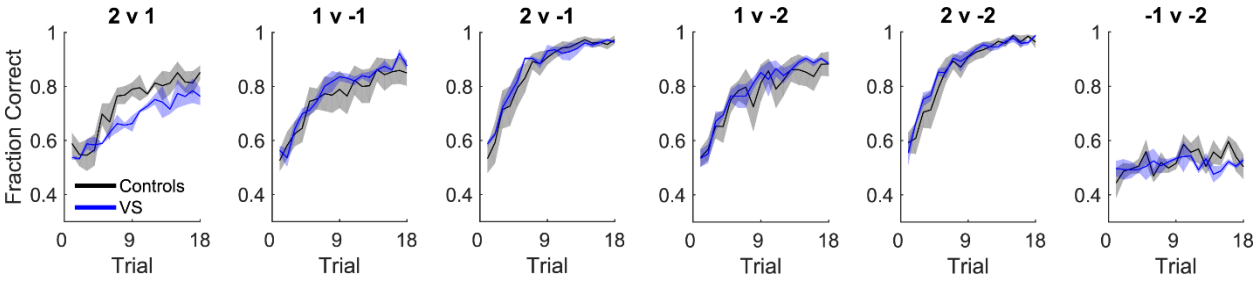


Figure 4. TkS Stochastic Reinforcement Learning. Task 3 choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

TkL

In task 4, we introduced four cues with cue-outcome associations of +2, +1, -1, -4. We added the larger loss cue to see if animals would learn to pick the smaller loss cue more effectively, when the difference between the two loss cues was larger. We also gave the monkeys an endowment of four tokens on the first trial after each cash-out. We did this to ensure that the animals had sufficient tokens to experience the large loss and to maintain motivation. Novel (Fig. 5) and Familiar (See SI Appendix Fig. S4) blocks were both composed of 108 trials, with 18 trials per condition.

Performance in novel blocks again showed a difference in performance across conditions (Fig. 5; Condition; $F(5,43) = 231.0$, $p < 0.001$) and a difference in performance across trials in different conditions (Condition x Trial; $F(85,313) = 4.6$, $p < 0.001$). There was also a main effect of group (Group; $F(1,31) = 30.7$, $p < 0.001$). None of the group effects in individual conditions survived multiple comparisons corrections. The monkeys were able to learn to choose the smaller of the two losses ($t(5) = 10.8$, $p < 0.001$). In addition, when we examined each group separately we found that both groups were able to learn to choose the smaller of the two losses (Control: $t(2) = 5.2$, $p = 0.035$), VS: $t(2) = 14.6$, $p = 0.004$).

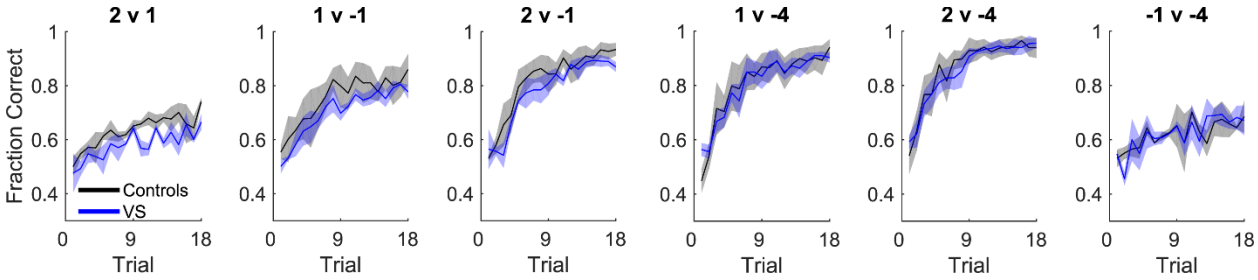


Figure 5. TkL Deterministic Reinforcement Learning with a Large Loss. Task 4 choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

3.2.2 Reinforcement Learning Models

Next, we fit RL models to the data in the novel blocks in all tasks. We fit two models which varied in the number of free parameters used to model the choice behavior. One model used one parameter for positive cues and one parameter for negative cues (VALENCE model), thus allowing learning rates to vary for positive vs. negative outcomes. The second model used one parameter for each cue (CUE model), allowing for different learning rates for each outcome. The VALENCE model fit behavior well in most conditions, particularly in the control group (See SI Appendix Fig. S5). However, the VALENCE model overpredicted performance in the 2 v 1 condition. This could be seen, for example, in task 3 which had stochastic outcomes, and in which there was a large discrepancy between behavior and model predictions in the 2 v 1 condition (See SI Appendix Fig. S5 B, D). This effect was strongest for the VS animals (See SI Appendix Fig. S5D) but could also be seen for the control animals (See SI Appendix Fig. S5B).

In the other conditions, however, the VALENCE model fit well. The CUE model did not show biases in any conditions in either group (See SI Appendix Fig. S5 A, C).

Averaged across tasks the VALENCE model overpredicted performance in the 2 v 1 condition in both groups (Fig. 6A, B; VALENCE Model vs. behavior, $F(1,5) = 167.9$, $p < 0.001$). The VALENCE model over-predicted behavior more for the VS group than the controls (Group x VALENCE Model vs. Behavior, $F(1, 5) = 19.4$, $p = 0.007$). The CUE model, on the other hand, did not differ from behavior in the 2 v 1 condition, across tasks (CUE Model vs. behavior, $F(1, 5) = 0.2$, $p = 0.582$), although the fit did differ by group (CUE model vs. Behavior x Group, $F(1, 5) = 14.2$, $p = 0.016$), with a closer fit between behavior and model in the VS group. We also used the Bayesian Information Criterion (BIC) to assess which model fit best in each session for each animal and task. In all animals in both groups, averaged across tasks, the CUE model was more frequently the best model than the VALENCE model. Across groups there was a preference for the CUE model over the VALENCE model ($t(6) = 2.84$, $p < 0.030$, 57% of sessions best fit by CUE model). This preference was not significant individually in the control ($t(3) = 1.6$, $p = 0.210$, 57%) or VS animals ($t(2) = 3.03$, $p = 0.094$, 57%).

Next, we compared the learning rate parameters between groups from the CUE model (Fig. 6C). We found that the parameters varied across cues (Fig. 6C;

$F(3, 15) = 45.6$, $p < 0.001$) and tasks ($F(3, 14) = 4.0$, $p = 0.030$). Learning rates were lower in the VS group (Group; $F(1, 5) = 9.75$, $p = 0.026$). The groups did not differ across cues (Group x cue, $F(3, 15) = 2.9$, $p = 0.070$) or tasks (Group x task, $F(3, 14) = 1.8$, $p = 0.193$). We also examined effects within the gain cues and within the loss cues, separately. There were group differences within the gain cues (Group, $F(1, 5) = 10.35$, $p = 0.024$) and these effects differed marginally across the two cues (Group x cue, $F(1, 5) = 9.3$, $p = 0.029$). There were no group differences in the loss cues (Group, $F(1, 5) = 2.1$, $p = 0.206$) but the groups did differ by cue (Group x cue, $F(1, 5) = 9.2$, $p = 0.029$). When we examined group differences in the individual tasks, we found a difference in groups across cues for Task 3 with stochastic outcomes (Fig. 6D, Group x cue, $F(3, 15) = 5.0$, $p = 0.014$) but no other group differences in the other tasks ($p > 0.05$). We also examined the inverse temperatures from the model fits. These differed by experiment ($F(3, 14) = 4.4$, $p = 0.021$). However, there were no differences across groups ($F(1, 5) = 0.0$, $p = 0.887$) and only a trend by group across experiments ($F(3, 14) = 1.5$, $p = 0.267$). None of the choice autocorrelation parameters (see methods) varied by group ($p > 0.05$).

Therefore, across tasks, the animals with VS lesions consistently had deficits in learning to discriminate between the two gain cues in the 2 v 1 condition, and this manifested as a significantly larger deficit relative to the VALENCE model

predictions in the 2 v 1 condition and a significant reduction in learning rates relative to controls specifically for the gain cues.

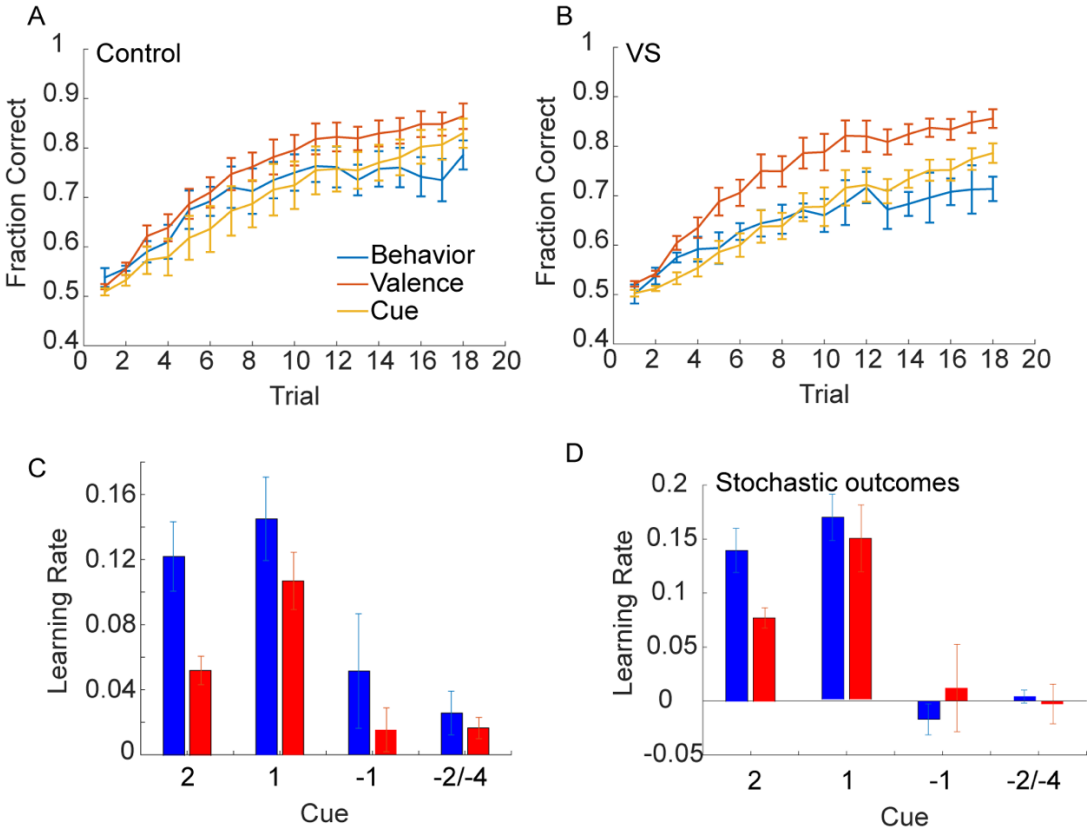


Figure 6. Best fitting reinforcement learning models. A. Overlay of behavior and predicted performance, averaged across experiments, for the 2 v 1 condition for the control animals. B. Same as panel A for the VS animals. C. Learning rate parameters averaged across tasks, extracted from the RL CUE model. Error bars are SEM across monkeys in each group. D. Average learning rates for the CUE model, cue parameters, in Task 3, with stochastic feedback. Control animals are blue and VS animals are red.

3.2.3 Aborted Trials & Reaction Times

We also examined aborted trials and reaction times across tasks (Fig. 7). It was sometimes the case that, when the two images were presented, the animals broke fixation and did not select either image in the novel (Fig. 7A) or familiar

conditions (Fig. 7B). The monkeys broke fixation more frequently when they had to choose between the -1 and -2 cues. This was true even though following an error we repeated the same condition. This differed by condition (Novel: $F(24,116) = 17.9$, $p < 0.001$; Familiar; $F(24, 116) = 9.92$, $p < 0.001$) and task (Novel: $F(3, 23) = 4653.0$, $p < 0.001$; $F(3, 16) = 155.9$, $p < 0.001$). There were, however, no interactions of task with other variables, and there was no effect of group (Novel: $F(1,5) = 2.26$, $p = 0.189$; Familiar: $F(1,5) = 0.85$, $p = 0.400$). In novel and familiar blocks, both groups aborted more trials in the -1 v -2 condition than any of the other conditions, when compared pair-wise (Controls: $p < 0.001$; VS: $p < 0.001$ for all pairs). The number of aborted trials was, however, larger in the familiar than novel tasks in the -1 v -2 condition (Novel vs. Familiar, $F(1,5) = 8.95$, $p = 0.031$).

Next, we examined reaction times. In both novel (Fig. 7C) and familiar (Fig. 7D) blocks, there were differences in reaction times across conditions (Novel: $F(24,116) = 11.4$, $p < 0.001$; Familiar: $F(24,115) = 15.2$, $p < 0.001$) and marginal differences across tasks (Novel: $F(3,14) = 3.4$, $p = 0.046$; Familiar: $F(3, 14) = 4.09$, $p = 0.029$. However, there were no differences between groups (Novel: $F(1,5) = 0.45$, $p = 0.534$; Familiar: $F1,5) = 0.25$, $p = 0.636$) and no higher-order interactions ($p > 0.05$). In both the novel and familiar blocks, animals in both groups were

slowest when choosing between the two loss cues relative to all other conditions (Controls: $p < 0.001$; VS: $p < 0.001$ for all pairs).

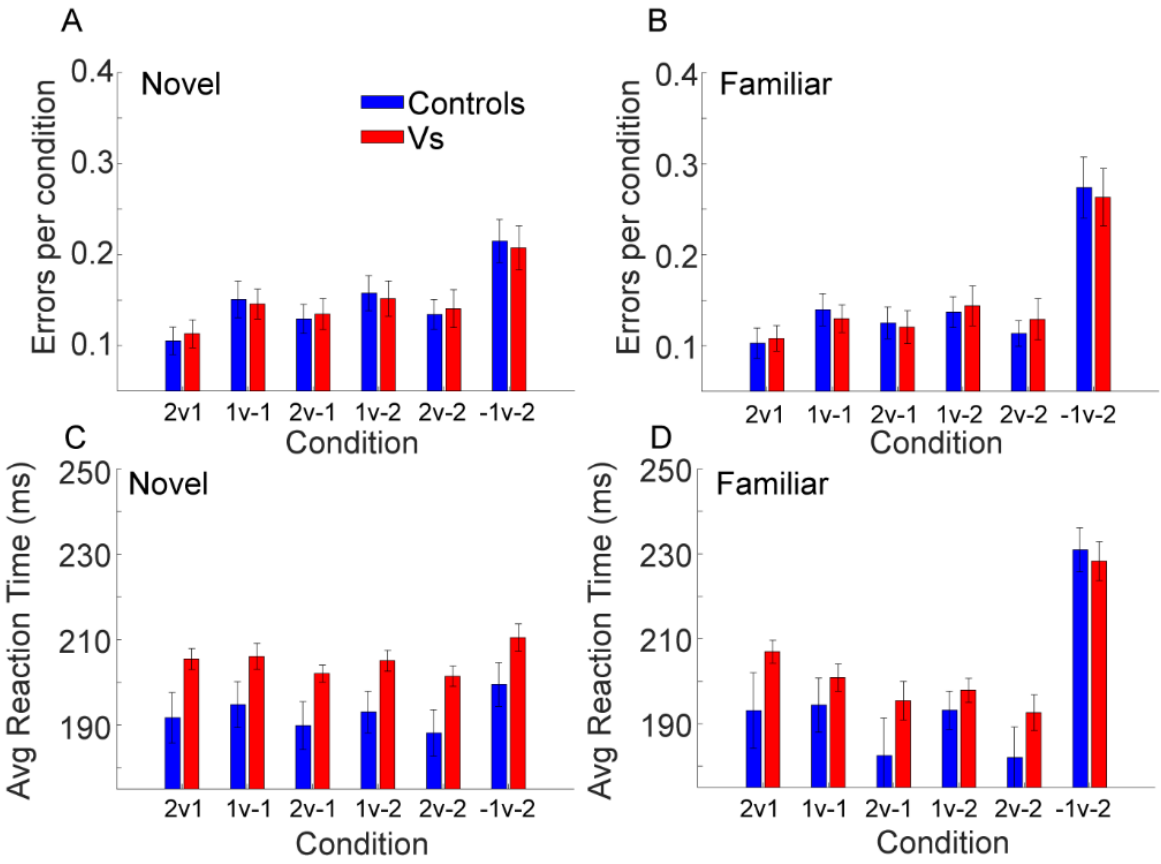


Figure 7. **Aborted trials and reaction times averaged across tasks.** Note that the data from Task 2 are averaged here except the conditions that included a Null cue (i.e. 0/1, 0/-1, etc). See Fig. S6 for all conditions of Task 2. In addition, the ANOVA model included all conditions, as they were nested under Task. A. Aborted trials in the novel conditions. Errors indicate the fraction of trials where the animals held initial fixation, but then failed to select one of the choice options. B. Aborted trials in the familiar condition. C. Reaction times in novel conditions. D. Reaction times in familiar conditions.

3.4 Discussion

We carried out four tasks in which we examined learning from gains and losses, using tokens as secondary reinforcers. We found that monkeys learned to make choices that increased their tokens and to avoid choices that decreased their tokens. When we examined group differences in learning novel cues, monkeys with VS lesions were impaired when the feedback was stochastic, and when the large loss choice had a value of -4. We also fit RL models to behavior and found a preference for the CUE model, which had a separate learning rate for each of the cues, relative to the VALENCE model which had one learning rate for positive outcomes and one for negative outcomes. When we compared learning rates from this model between groups, we found that animals with VS lesions had significantly reduced learning rates specifically for the gain cues. Furthermore, when we examined behavior relative to the VALENCE model, to see where it failed to account well for choices, we found that it specifically overpredicted performance in the 2 v 1 condition, and this overprediction was larger in the VS animals than the control animals. This was after optimizing learning rates in this model. Therefore, animals with VS lesions show specific deficits in learning to choose between secondarily reinforced rewarding options, with no apparent deficits in learning to choose between gain and loss cues, or between two loss cues.

Token based reward mechanisms have been used previously to motivate behavior in macaques (17). We found that monkeys learned effectively to choose options that increased their tokens and avoid options that decreased their tokens. In addition, aversive stimuli can affect behavior in multiple ways (193). Consistent with this we found that when monkeys had to choose between two loss options, they learned to choose the option leading to the smaller loss. They also aborted significantly more trials and had the longest reaction times when they had to choose between two losses. By these measures our monkeys found losing tokens to be aversive. We found effects of VS lesions on choice behavior, but not reaction times or aborted trial behavior. Therefore, these behaviors maybe be mediated by different systems, or the VS may contribute more to choice behavior than speed of response and avoidance.

Distinct circuitry underlying appetitive and aversive learning

Recent work has attempted to delineate separable neural circuits underlying appetitive and aversive learning. For example, Lammel et al. suggested that a circuit from the lateral habenula, through a subset of dopamine neurons that responded to aversive stimuli, to the IL/PL region of medial prefrontal cortex, was important for aversive learning (200). Distinct from this, another circuit from the rostral-medial tegmental nucleus, through a subset of dopamine neurons that responded to appetitive stimuli, to the VS, was important for appetitive learning.

However, subsequent anatomical work has not supported the suggestion that dopamine neurons with different projection targets have different inputs (204, 205). Other work has suggested that basolateral amygdala neurons that project to the VS are important for appetitive learning, and basolateral amygdala neurons that project to the central nucleus of the amygdala are important for aversive learning (197, 206). Circuitry connecting the amygdala to the dorsal anterior cingulate cortex has also been implicated in aversive learning (207). Inactivation of both ventrolateral prefrontal cortex, and orbitofrontal cortex has also been shown to increase sensitivity to punishment (208). And there is extensive work supporting the amygdala's role in learning threat of shock (209).

In addition to the circuit work in rodents and nonhuman primates, other work has directly examined learning in the context of winning money, or not losing money, which has similarities to our use of tokens. This work has also suggested that the VS, and dopamine modulation of VS activity, is important for learning to choose rewarding options (19). The same study suggested that the insular cortex was important for learning to avoid losing money, a finding supported by work in patients with insular cortex lesions (210). Notably, avoiding monetary losses has consistently been shown to be independent of dopamine (19, 211). Additional work has shown that aversive pruning, which is the process of eliminating choices that lead to future situations in which large punishments might

be experienced, engages the subgenual cingulate cortex (212). It has also been shown that microstimulation in a related subgenual cingulate region can bias choices away from aversive options, although not in the context of learning (213). Thus, the VS has often been implicated in appetitive learning. Aversive learning, on the other hand, has been linked to dorsal cingulate cortex, subgenual cingulate cortex, insular cortex and the central nucleus of the amygdala. Our data are consistent with the hypothesis that the VS plays a specific role in learning about gains, without contributing to learning about or choosing when losses are involved. In our behavioral data the deficits in learning about gains were specific to choosing between pairs of gains and did not manifest when a gain was paired with a loss. However, the RL model showed that learning rates were overall lower for gain cues.

The differences between the circuit work in rodents and the systems work in humans and monkeys that have identified different systems for aversive learning may in part be due to differences in the appetitive and aversive modalities used. Appetitive and aversive stimuli come in many forms, and these are processed in separable systems, at some level (193, 214). For example, nociceptive information relayed via the dorsal horn of the spinal cord is distinct from information about threats from conspecifics, which may arrive via the auditory or visual system, depending on the nature of the threats. In addition, the processing of token losses

would presumably involve different neural circuits from conditioned defensive responses to shock or loud noise. It is currently not clear where information about appetitive or aversive outcomes arising from different modalities is integrated, or if it is ever integrated. Thus, there may be no simple circuit that processes all appetitive or aversive information, independent of modality.

Although the VS is not often implicated in aversive learning, studies have shown that VS neurons respond to both rewarding and aversive stimuli (215). In addition, VS neurons respond to rewarding stimuli that have been subsequently negatively conditioned using injections of lithium chloride (216). Other work has shown that cues that have been negatively conditioned can lead to increased dopamine release in the VS shell, but decreased dopamine release in the VS core (217). In contrast to this, however, tail pinch has been shown to increase dopamine release in the dorsal striatum and the core of the VS (218, 219). Removal of tail pinch also leads to increased dopamine release in the VS shell (218), consistent with the hypothesis that pain relief can be rewarding (220). Further work has shown that oral infusions of quinine, which is aversive, leads to decreased dopamine concentration in the VS core, whereas oral infusion of sucrose leads to increased dopamine concentration (221). Therefore, the relationship between single neuron responses, dopamine concentration and appetitive and aversive

stimuli in the VS core and shell is complex and depends on the modality of the stimulus and perhaps anesthesia state.

Learning deficits in VS lesioned animals

In our study in the novel condition, we found differences between monkeys with VS lesions and unoperated controls in both task 3, in which outcomes were stochastic, and task 4 in which we used a large loss. We have previously found that animals with VS lesions learn more effectively when outcomes are deterministic, and have substantial deficits when outcomes are stochastic (10, 11, 201). The deficits are consistently largest when the monkeys with VS lesions have to learn to choose between two options that have the same reward magnitude, but differ in reward probability. This may be consistent with work showing that lesions of the VS affect dopamine coding of prediction errors for reward delays, but not reward magnitudes (26). Reward rate estimation, which is required for learning values in tasks with stochastic outcomes, requires estimates of time between rewards.

In the current study the two options always had different reward magnitudes, but the same reward probabilities. While the monkeys with VS lesions had deficits in these tasks, the effect was smaller than we observed in a series of tasks with stochastic outcomes. We have suggested that the amygdala, and also cortical systems (4, 158), learn in parallel with the VS. The amygdala, however, learns

with a higher learning rate than the VS (10, 203). Therefore, in monkeys with VS lesions the amygdala and anatomically related cortical systems may play a larger role in learning than in intact monkeys. The higher learning rate amygdala system is more susceptible to noise, because it rapidly updates value estimates following a non-rewarded choice, and values therefore tend to oscillate when feedback is stochastic (203). The VS updates values with a slower learning rate than the amygdala. When the VS is intact, learning is less affected by stochastic outcomes because the VS value estimates are updated less, after individual outcomes, thereby offsetting the rapid updating carried out by the amygdala. Thus, lesions of the VS lead to larger deficits when feedback is stochastic, and particularly large deficits when only reward probability, and not reward magnitude, can be used to optimize choices. It is likely that the cortex, and mediodorsal thalamus, also contribute to learning in these tasks (158, 222, 223). However, how this mono-synaptically connected circuit works together to mediate learning is a topic for future research.

We also found group differences in the familiar condition in all tasks. Except in task 1, however, the behavioral differences tended to be rather subtle. One possible explanation for the finding of subtle yet significant differences in the familiar conditions is that the controls often had near perfect performance in some conditions. Because accuracy is a bounded variable (and despite the fact that we used a transform to normality before running the ANOVAs), this near perfect

performance leads to very small variance, which leads to significant differences. Therefore, the subtle differences in choice accuracy were significant. In most conditions, performance was very high in the conditions that had at least 1 gain cue. Performance in the -1 v -2 condition, or in the 0 v loss conditions in task 2, never reached high levels, even after extensive experience.

In previous tasks, we also found that monkeys with VS lesions responded faster than controls (10, 11). In the current task, there were no group differences in reaction times, and there was a trend for the monkeys with VS lesions to respond more slowly than the controls. Thus, the presence of loss cues in the token tasks slowed the reaction times of the VS animals. Previously, we also found that much of the deficit in the VS animals, relative to controls, could be accounted for if reaction times were matched between groups (10). This followed because there was a speed accuracy trade-off, such that responding quickly led to less consistent choice of the best cue. Thus, the slowed reaction times in the current task may partially explain the accurate performance of the VS lesioned animals in several conditions.

RL model

An RL model with a separate learning rate for each cue (CUE model) best fit the data for both the control and VS groups in all tasks. For most of our tasks, the VALENCE model is the same as a model that would fit one learning rate for

positive reward prediction errors and one for negative reward prediction errors, because values start out at 0 and outcomes are deterministic. When we examined learning rates across experiments, the monkeys with VS lesions had reduced learning rates specifically for gain cues. When we examined performance of the models in each condition, to see where the VALENCE model failed to account for behavior, we found that it overpredicted performance in the 2 v 1 condition in both groups, but that this effect was larger in the monkeys with VS lesions relative to controls. Therefore, analysis of learning across experiments showed specific deficits in the animals with VS lesions in learning the values of gain cues, with no overall deficits in learning the values of loss cues.

As a final point, the monkeys in both groups also appeared to learn poorly in the -1 v -2 condition, although they did show statistically significant learning in all tasks. We did not find, however, that allowing for a different choice consistency parameter (i.e. inverse temperature) for loss choices improved the fit of the RL model. Both the CUE and VALENCE models used different learning rates for gain and loss cues and learning was slower in the loss conditions. In addition to the smaller learning rates for the loss cues, however, these cues were also chosen less often, and therefore their values were less frequently updated. For example, the +2 cue was frequently chosen in every pair it was part of, whereas the -2 cue was rarely chosen. Value updates only happen in the RL model when an option is

chosen and the outcome is experienced. Therefore, the decreased learning in the -1 v -2 condition follows both from decreased learning rates and less experience with the outcomes associated with those options.

3.4.1 Conclusion

We compared learning from gains and losses in animals with VS lesions and an unoperated control group. We found behavioral deficits in monkeys with VS lesions in two of the 4 tasks, when comparing choice accuracy. These deficits were consistently driven by trials in which animals had to choose between two cues that differed in positive reward magnitude. There were no deficits when animals had to choose between options, one of which was associated with a loss. We also fit RL models to the data, and found that learning rates were lower for gain cues in the VS animals relative to controls. Thus, lesions of the VS, in this task, specifically affected learning to choose between rewarding options, and had no effect on learning to avoid losses.

3.5 Supplemental data

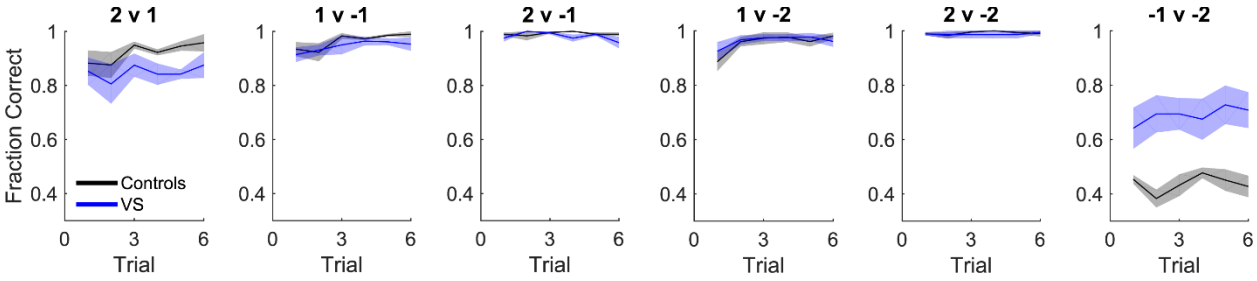


Fig. S1 Token task 1- Deterministic learning

In the familiar blocks (Fig. S1A), there were differences in performance across conditions despite the extensive experience the animals had with all cues (Condition; $F(5,34) = 182.0$, $p < 0.001$). There were also differences between the groups across conditions (Group x Condition; $F(5,26) = 14.8$, $p < 0.001$). To examine this effect, we tested each condition separately. (All condition effects are reported uncorrected, but we only state effects as significant that would survive Bonferroni correction for number of conditions.) The only condition that showed a significant difference between the groups was the -1 v -2 condition ($F(1,5) = 18.55$, $p = 0.008$). When we examined the average fraction correct in this condition we found that the monkeys did not do better than chance ($t(6) = 0.8$, $p =$

0.452). In addition, when we looked at both groups individually we found that neither group learned to choose the smaller loss at above chance levels (Control: t(3) = -2.9, p = 0.058), VS: t(2) = 3.2, p = 0.087). Overall, the VS animals performed slightly above chance, and the control animals performed slightly below chance, driving the group difference, but not leading to significant learning in either group.

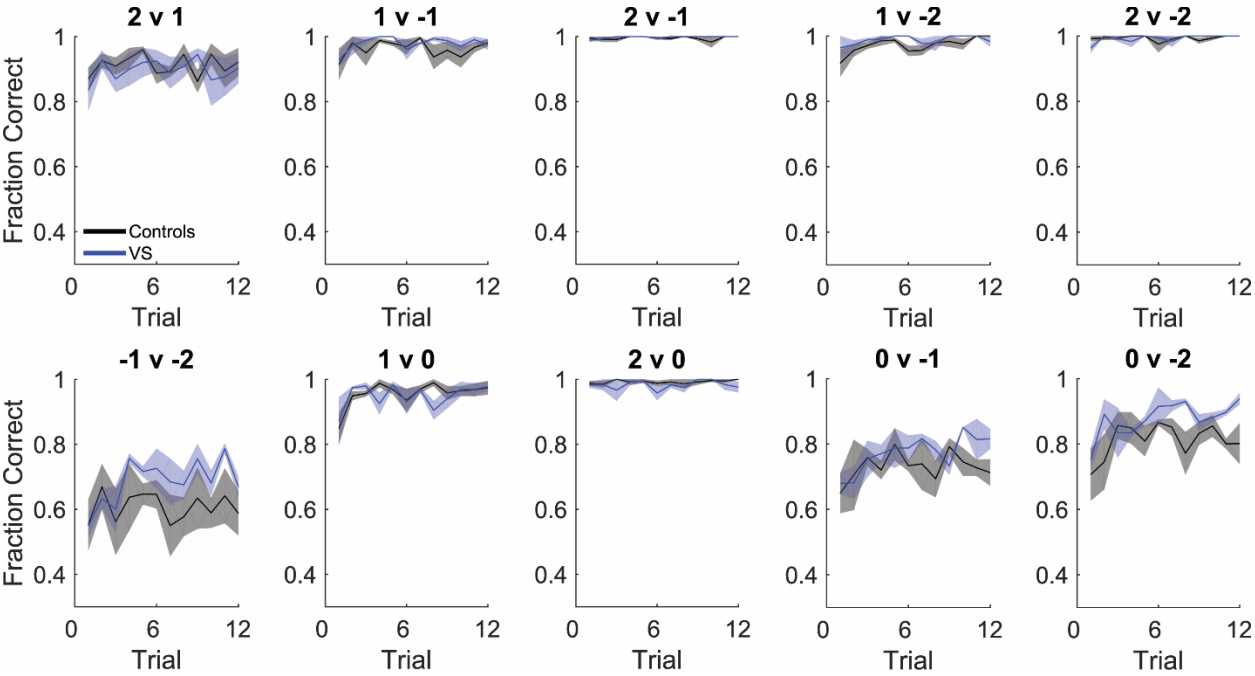


Fig. S2 Token task 2 - Deterministic learning with a null cue

In the familiar blocks (Fig. S2), there were differences across conditions (Condition; F(9,50) = 53.9, p < 0.001). There was also an effect of group (Group; F(1,20) = 7.8, p = 0.011), but no effect of group by condition (Group x Condition; F(9,41) = 0.8, p = 0.580). Because of the ceiling performance, some conditions

had low variance (e.g. 2 v 0, 2 v -1 and 2 v -2), which may have been driving the group differences. In the familiar blocks, the animals performance in the condition in which they had to choose between the two loss cues was above chance ($t(6) = 3.5$, $p = 0.012$). When we tested the groups separately we found that only the VS animals reached significance (Controls: $t(3) = 1.6$, $p = 0.198$; VS: ($t(2) = 6.6$, $p = 0.022$). All the animals and both groups chose between the 0 and -1 cue above chance (All animals: $t(6) = 11.5$, $p < 0.001$; Controls: $t(3) = 7.3$, $p = 0.005$; VS: $t(2) = 10.1$, $p = 0.009$). This was also the case for choosing between the 0 and -2 cue (All animals: $t(6) = 13.9$, $p < 0.001$; Controls: $t(3) = 8.5$, $p = 0.003$; VS: $t(2) = 19.5$, $p < 0.002$).

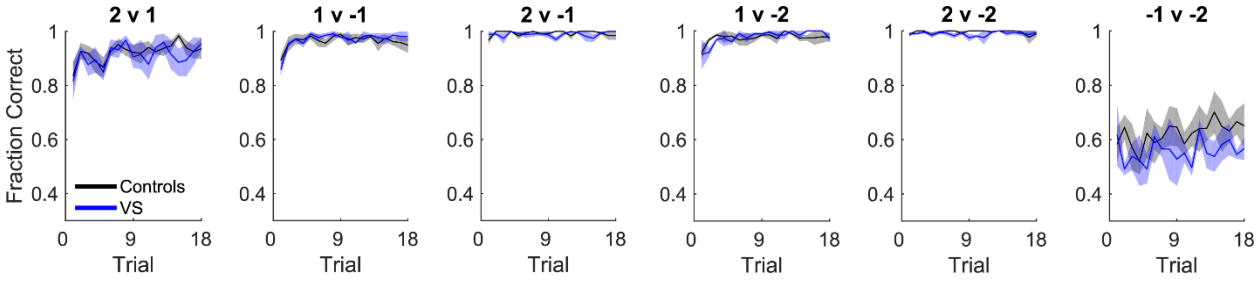


Fig. S3 Token task 3 - Stochastic learning

Performance in the familiar blocks was similar to performance in the other tasks. Consistent with the previous experiments, there was a difference in performance across conditions (Conditions; $F(5,38) = 539.0$, $p < 0.001$). There was also a difference between groups (Group; $F(1,32) = 8.2$, $p = 0.007$), but no difference in groups by condition (Group x Condition; $F(5,35) = 1.5$, $p = 0.249$).

The animals learned in the -1 v -2 condition ($t(6) = 2.7$, $p = 0.032$). However, when we examined each group separately, neither group reached significance alone (Controls: $t(3) = 2.4$, $p = 0.095$; VS: $t(2) = 2.9$, $p = 0.101$).

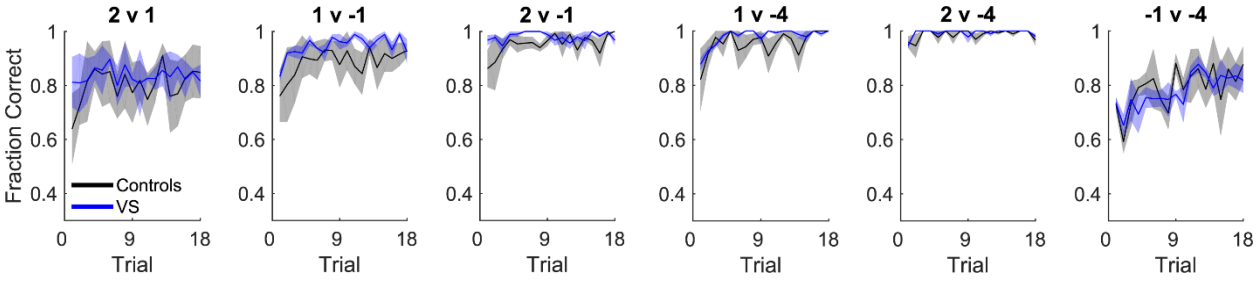


Fig. S4 Token task 4 - Deterministic learning with large loss

In the familiar condition there were differences in performance across conditions (Fig. 8B; Condition; $F(5,24) = 29.9$, $p < 0.001$). There was also an effect of group ($F(1,40) = 6.1$, $p = 0.017$), but no group by condition effect (Group x Condition; $F(5,17) = 0.2$, $p = 0.955$). Similar to the novel data the animals were able to pick the smaller of the two losses more often than chance ($t(5) = 13.9$, $p < 0.001$). When we examined the groups individually we found that both groups chose the smaller loss (Controls; $t(2) = 6.7$, $p = 0.022$, VS; $t(2) = 21.9$, $p = 0.002$).

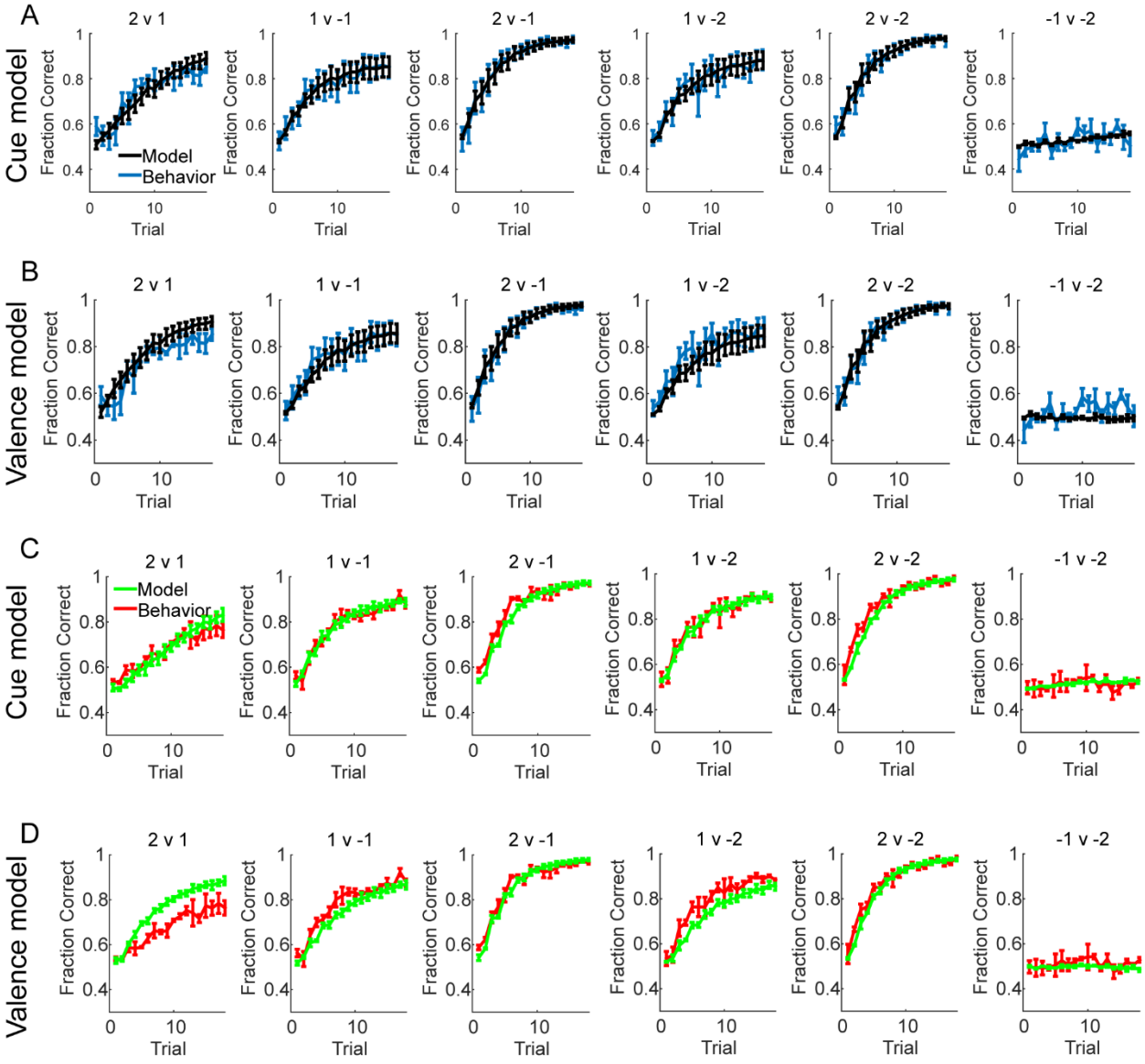


Fig. S5 Fits of CUE and VALENCE models overlaid on choice behavior for both groups for task 3, which used stochastic feedback. Error bars are +/- s.e.m. with N = number of animals.

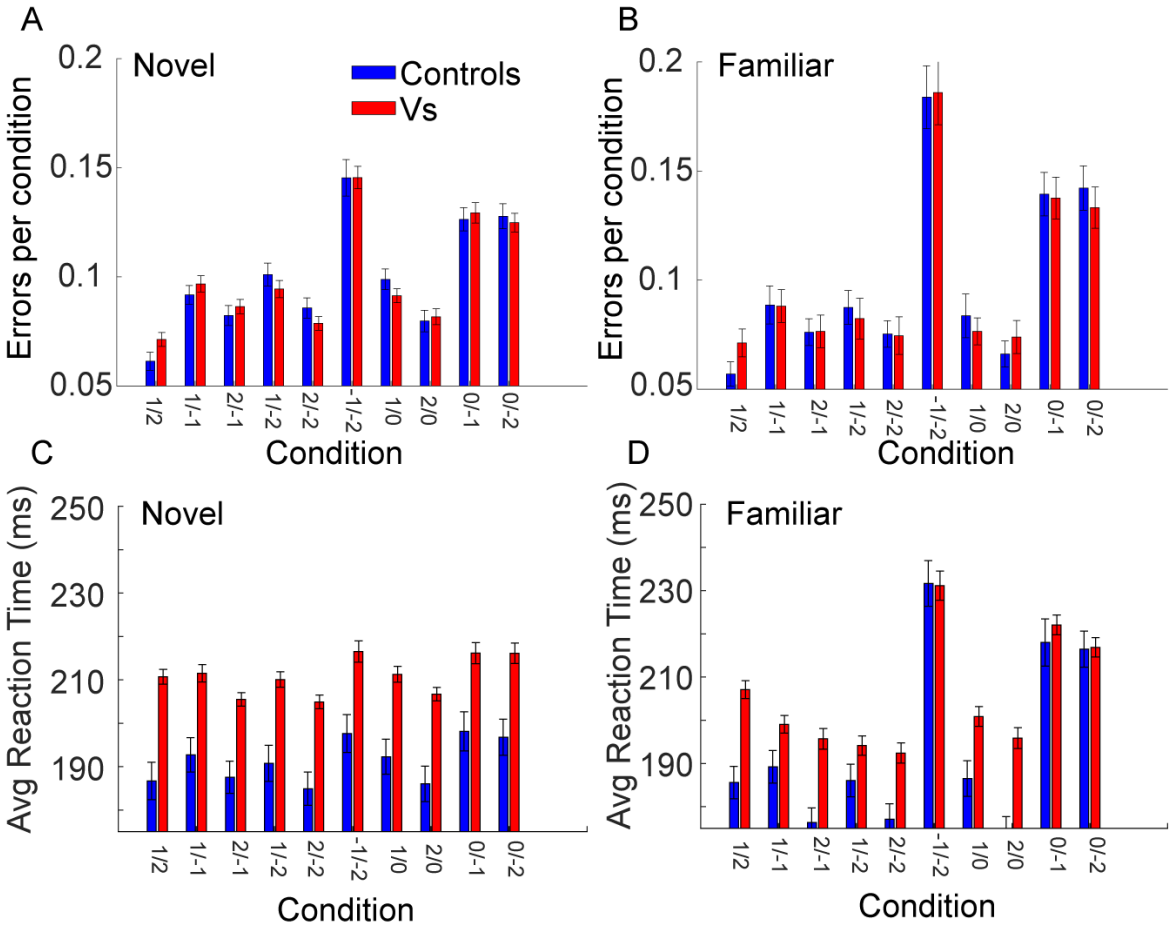


Fig. S6 Aborted trials and reaction times in the novel and familiar blocks of the Null token experiment. A. Aborted trials in each condition in the novel blocks. B. Same as A for familiar. C. Reaction times for the novel conditions. D. Same as C for the familiar conditions.

Chapter 4: The Amygdala's Role in Learning from Gains and Losses

4.1 Introduction

The natural environment is comprised of both appetitive and aversive stimuli and it is essential for survival that an agent learns to respond to these appropriately. Reinforcement learning (RL) is the process organisms use to navigate an ever changing external environment. More specifically, RL is the behavioral process of learning the value of actions or objects based on the consequences that follow the chosen action or object. Most work on RL has focused on learning in strictly appetitive environments. This work suggests that the striatum and the dopamine input to the striatum underlie RL (10, 19, 159). This work has led to general assumptions about RL that ignore the effects that complex learning environments have on an organism.

Different learning environments have different effects on psychological processes, such as motivation and emotion, which in turn can affect conditioning differently. For example, when we tested monkeys with ventral striatum (VS) lesions on a learning task in which monkeys could gain or lose tokens, we found

that the VS lesion animals only had deficits when choosing between two gain cues with different reward magnitudes (Taswell, C, et al, 2019). The monkeys with VS lesions did not have deficits when choosing between a gain and loss cue, or between two loss cues that varied in magnitude. If this task only contained the appetitive cues (gains), we would have missed this differential effect that learning from gains and losses simultaneously had on behavior. This result highlights the need to understand how a full spectrum learning environment, one that mimics the natural environment more closely where positive and negative outcomes are always possible, affects RL, and the ways in which neural systems underlie this process.

The fact that VS lesions only lead to deficits when monkeys chose between two gain cues suggests that another structure is responsible for the other components of this task. For numerous reasons the amygdala is uniquely suited to be this other structure. First, from an anatomical perspective the amygdala projects to the ventral striatum (175), and ventrolateral and orbital prefrontal cortex (176, 177), which positions it well to code value representations, specifically appetitive, and drive motivated behavior via these projections. Indeed, there is evidence supporting this view, showing that the amygdala plays an important role in visual stimulus based RL (10, 23, 51, 134-136).

Second, the amygdala also projects to the periaqueductal grey (140), these two systems are often implicated in emotional processing, in particular aversive emotions (Baxter & Murray, 2002), which is thought to affect aversive stimulus outcome associations. As with appetitive learning, there is evidence for the amygdala playing a role in aversive learning (Davis 1992, Cardinal, Parkinson et al. 2002, Namburi, Beyeler et al, 2015).

Finally, In addition to the evidence associating the amygdala with appetitive and aversive stimulus based RL, there is considerable evidence for the amygdala playing a crucial role in conditioned reinforcement. One of the possible conclusions for the minimal learning deficits of the VS lesioned animals in (Taswell, C, et al, 2019) is due to the tokens that were used as conditioned reinforcers. Conditioned reinforcers can aide in the learning process, so it is possible that another structure, perhaps the amygdala, was able to take over and learn the stimulus outcome pairing, which is why that study only found deficits when animals had to choose between two rewarding options. To assess the amygdala's role in learning from gains and losses, we tested monkeys with lesions to the amygdala (n = 4) on the same experiments used in (Taswell, C, et al, 2019).

4.2 Methods

4.2.1 Subjects

The subjects included 6 male and 2 female rhesus macaques with weights ranging from 6-11 kg. Three of the male monkeys and one female received bilateral excitotoxic lesions of the Amygdala. The remaining four monkeys served as unoperated controls (3 males and 1 female). One male from both the control and lesion animals was not able to complete all 4 tasks, and therefore task 4 only has 3 controls and 3 lesion monkeys (these were the same lesion monkeys used in chapter 2). For the duration of the study monkeys were placed on water control. On testing days monkeys earned their fluid from their performance on the task. Experimental procedures for all monkeys 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.

4.2.2 Surgery

See Chapter 2.2.2

4.2.3 Lesion assessment

See Chapter 2.2.3

4.2.4 Task

See Chapter 3.2.4

4.2.5 Image & Eye Tracking

See Chapter 3.2.5

4.2.6 ANOVA Models

To quantify differences between choice behavior in each group, we performed an arcsine transformation on the choice accuracy values from each session, as this transformation normalizes the data (152). We then carried out an N-way analysis of variance (ANOVAN). Monkey and session were included as random effects with session nested under monkey. All other factors were fixed effects.

4.3 Results

We compared monkeys with lesions to the amygdala (n = 4) against unoperated controls (n = 4) on a series of four tasks. In each task, trials involved a forced choice between two images. Selection of a particular image led to increases or decreases in tokens (Fig. 1A). The outcome of each trial following a choice was displayed on the monitor screen as a change in the number of tokens the animal had accumulated. Every four to seven trials, with the interval chosen randomly,

we cashed out the accumulated tokens. During cash-out the monkeys received one drop of juice for each token. The animals had to learn over trials to select the image from the pair that maximized their gains and minimized their losses. The animals could also not incur negative token counts, so when monkeys had no tokens and they chose a loss cue, there was no change in the tokens.

The tasks were run in a fixed sequence (Fig. 1A). Each task evaluated the choices during learning and performance on novel or familiar stimulus-outcome associations. In the novel blocks, the monkeys learned stimulus-reward associations for a novel set of images. In the familiar blocks, the monkeys chose between stimuli they had repeatedly sampled over the course of prior experimental sessions. The stimulus-outcome associations of these familiar choice options were fixed for the duration of the experiment. Novel and familiar blocks were randomly interleaved each day. The novel blocks allowed us to examine the rate at which cue-reward associations were learned, whereas the familiar blocks allowed us to examine asymptotic performance with overlearned cue-reward associations.

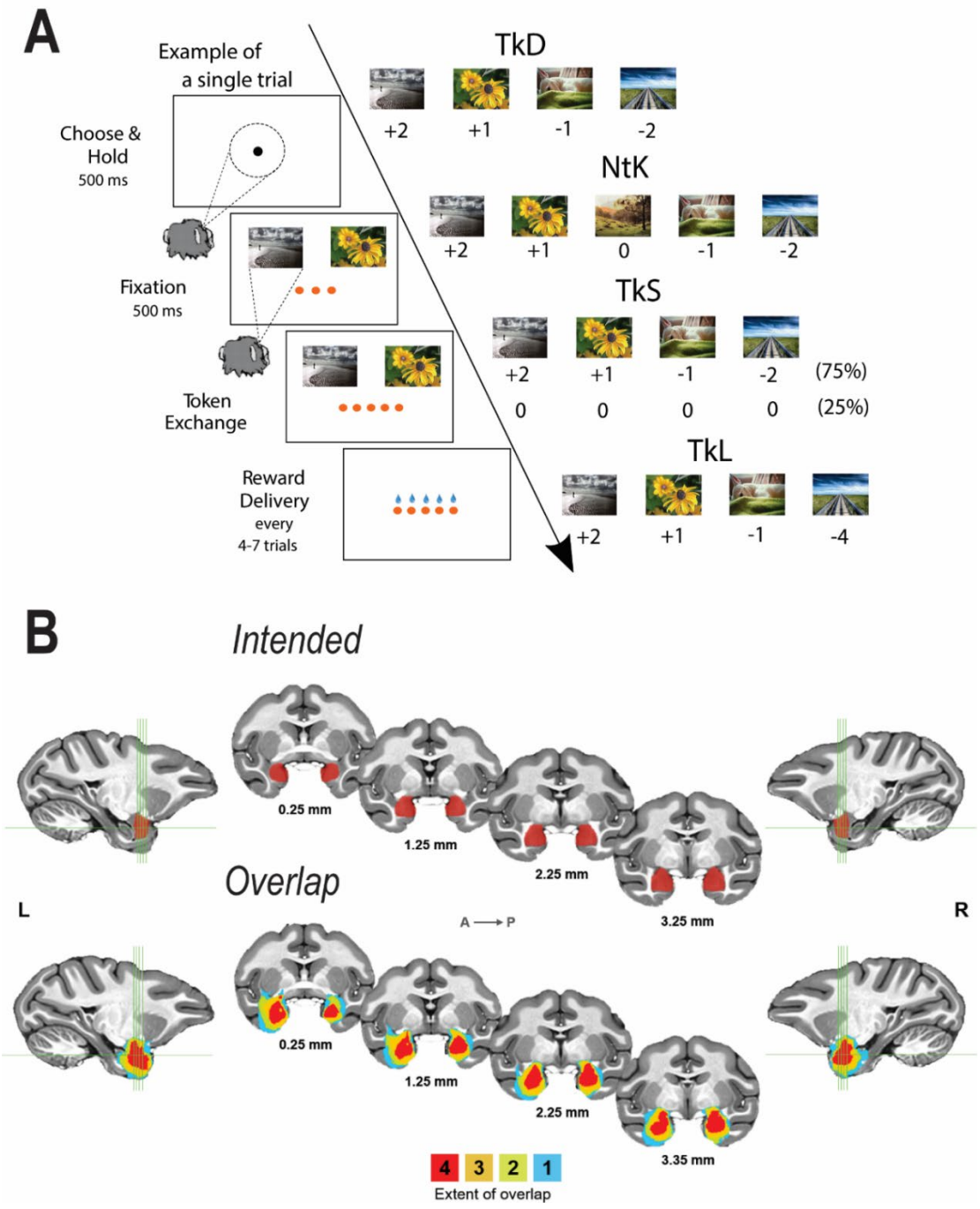


Figure 1. Tasks used and Lesion Map. A. Diagram of the trial structure used in all tasks. The specific reward magnitudes used in each task are shown. TkD, Task 1 in which deterministic reward magnitudes were +2, +1, -1 and -2. NtK, Task 2 in which we include a null token giving deterministic reward magnitudes of +2, +1, 0, -1 and -2. TkS, Task 3 in which feedback was stochastic with magnitudes of +2, +1, -1 and -2. TkL, Task 4 in which deterministic reward magnitudes, including a large loss, were +2, +1, -1 and -4. B. Lesion map of the 4 animals in the lesion group. Colors indicate number of animals that had lesion of corresponding extent.

4.3.1 Choice Behavior

TkD

We first evaluated the ability of the monkeys with or without lesions of the amygdala to learn deterministic stimulus-outcome associations. In this task, at the beginning of each novel block, the monkeys encountered four images they had not seen before. Each image was associated with a fixed, deterministic gain or loss of tokens (+2, +1, -1 or -2 tokens). Two of the images were presented as choice options on each trial. This resulted in six unique pairs of images, which we refer to as conditions. In each block, conditions were randomly interleaved over intervals of twelve trials until each condition occurred 18 times in novel (Fig. 2) blocks and 6 times in familiar (Fig. 3) blocks.

In novel blocks, the monkeys learned the stimulus-outcome associations efficiently. With experience, they were able to choose the better option of the pair on a high proportion of trials (Fig. 2). There were differences in performance across conditions (Condition; $F(5, 30) = 106$, $p < 0.001$) and differences in performance across trials in the different conditions (Condition x Trial; $F(85, 516) = 13.3$, $p < 0.001$). The monkeys performed best in the conditions in which there was a loss paired with the largest reward. For example, they most often picked the best cue when choosing between the +2 and -1 and +2 and -2 conditions. This effect was driven largely by the frequency with which they experienced the

outcomes associated with each cue and the differences in the values of the cues. The animals most frequently picked the +2 cue across all conditions, and therefore most frequently received feedback on its value. There were no overall differences between groups (Group; $F(1, 6) = 1.7$, $p = 0.244$) and no differences between the groups across conditions (Group x Condition; $F(5, 30) = 0.7$, $p = 0.64$). The monkeys did not perform well in the -1 v -2 condition, although across the groups there was a significant positive correlation between choice accuracy and trial, which indicates learning ($t(7) = 5.4$, $p < 0.001$). When we examined the groups individually, we found that controls learned to choose the smaller loss more often with experience (Control: $t(3) = 6.8$, $p = 0.006$), while the lesioned animals did not (Amygdala: $t(3) = 2.7$, $p = 0.059$).

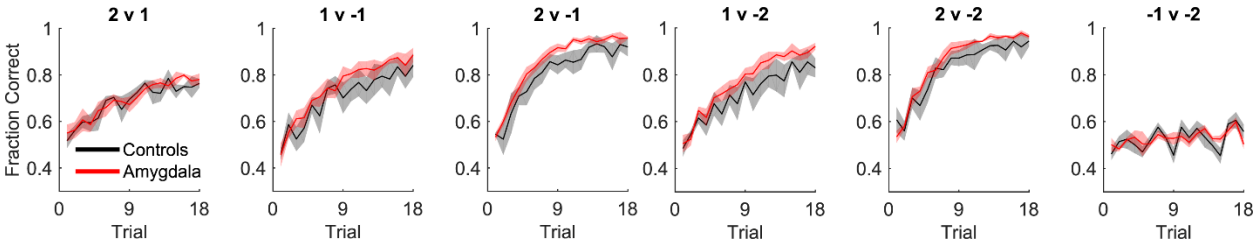


Figure 2. TkD Deterministic Reinforcement Learning of Stimulus-Outcome Associations. Task 1 novel choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

In the familiar blocks (Fig. 3), there were differences in performance across conditions despite the extensive experience the animals had with all cues

(Condition; $F(5,30) = 114.1$, $p < 0.001$). There were also differences between the groups across conditions (Group x Condition; $F(5,30) = 4.68$, $p = 0.003$). To examine this effect, we tested each condition separately. No condition showed a significant difference between the groups independently. The largest effect, however, was in the -1 v -2 condition ($F(1,6) = 8.9$, $p = 0.024$, uncorrected). When we examined the average fraction correct in this condition we found that the monkeys perform better than chance ($t(7) = 0.8$, $p = 0.443$). In addition, when we looked at both groups individually we found that neither group learned to choose the smaller loss at above chance levels (Control: $t(3) = -2.9$, $p = 0.059$), Amygdala: $t(3) = 2.8$, $p = 0.066$).

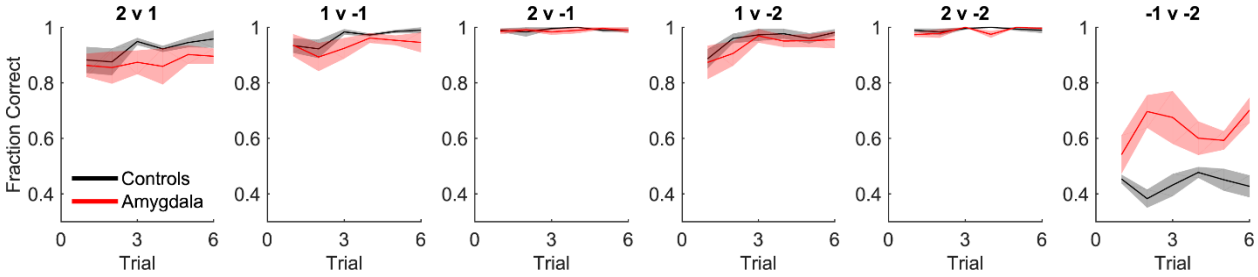


Figure 3. TkD Deterministic Reinforcement Learning of Stimulus-Outcome Associations. Task 1 familiar choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

NtK

In task 2, we used five cues in each block with cue-outcome mappings of +2, +1, 0, -1 and -2. This resulted in 10 pairs of cues and therefore 10 conditions. In task 2, both the novel (Fig. 4) and familiar (Fig. 5) blocks were composed of 120

trials, 12 per condition. Therefore, in the novel blocks the monkeys saw each pair of cues 12 times. Inclusion of the null cue allowed us to test two specific hypotheses. First, does the absolute difference between the value of the cues drive performance independent of the reward value associated with the cues? Second, can animals learn to select the null cue when it is paired with a loss cue?

In the novel blocks (Fig. 4), there was again a difference in performance across conditions (Condition; $F(9, 54) = 103.6$, $p < 0.001$) and also a difference in performance across trials in the different conditions (Condition x Trial; $F(99, 599) = 0.9$, $p < 0.001$). There were no overall differences between groups (Group; $F(1, 6) = 2.3$, $p = 0.17$) and no differences by condition (Group x condition; $F(9, 54) = 0.8$, $p = 0.6323$). There was, however, a difference between groups across trials (Group x trial; $F(11, 66) = 2.3$, $p = 0.0174$) and a three way interaction (Group x condition x trial; $F(99,599) = 1.4$, $p = 0.008$).

Similar to task 1, when we grouped all the animals together there was a significant correlation between trial and performance when the animals had to choose between the two loss cues ($t(7) = 4.3$, $p < 0.030$). However, when we separated the groups, only the lesion group reached significance (Controls: $t(3) = 2.6$, $p = 0.081$; Amygdala: $t(3) = 3.2$, $p = 0.049$). There was also significant learning when the animals had to choose between the 0 and -1 cue across groups ($t(7) = 6.1$, $p < 0.001$), and also in the amygdala animals when they were tested

individually (Amygdala: $t(3) = 13.2$, $p < 0.001$), but not the controls (Controls: $t(3) = 2.9$, $p = 0.062$). When the animals had to choose between the 0 and -2 cues there was learning across groups ($t(7) = 5.8$, $p < 0.001$). When we examined the groups separately we found that both groups performed significantly better than chance (Controls: $t(3) = 3.5$, $p = 0.037$; Amygdala: $t(3) = 10.5$, $p = 0.002$).

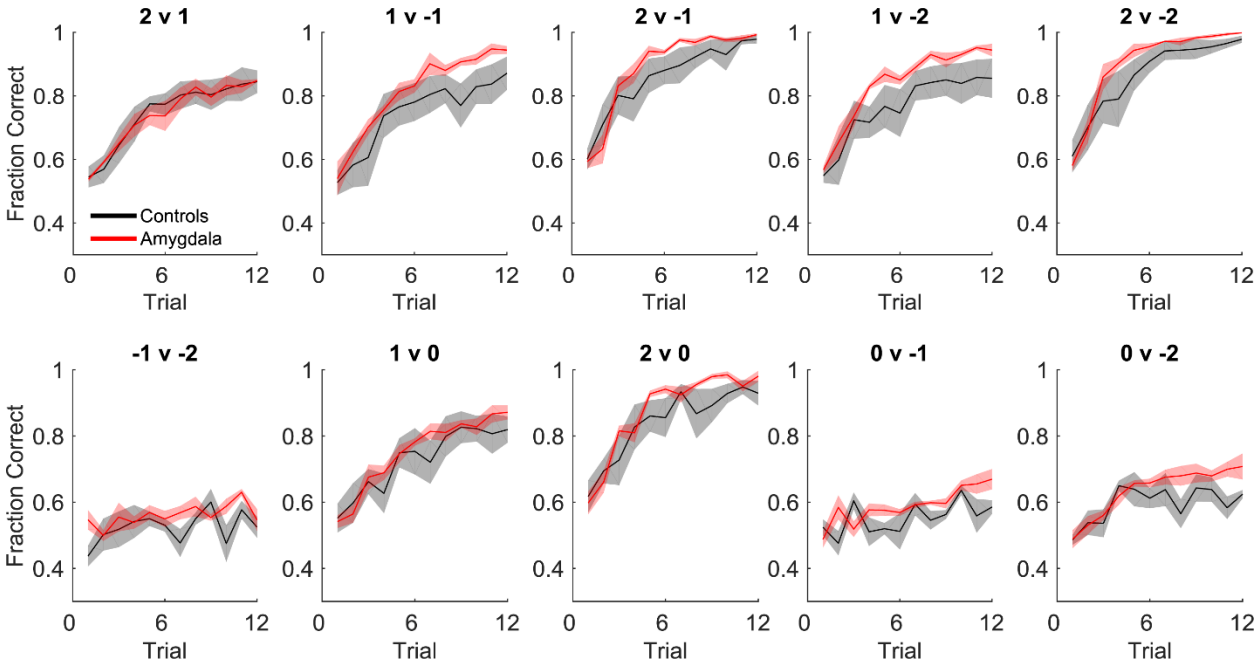


Figure 4. NtK Deterministic Reinforcement Learning Augmented by a Null Cue. Task 2 novel choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

In the familiar blocks (Fig. 5), there were differences across conditions (Condition; $F(9,54) = 38.1$, $p < 0.001$). There was no main effect of group (Group; $F(1,6) = 0.9$, $p = 0.383$), but there was an effect of group by condition (Group x

Condition; $F(9,54) = 2.2$, $p = 0.032$). When we analyzed the conditions separately, no condition showed a significant group effect on its own.

In the familiar blocks, the monkeys' performance in the condition in which they had to choose between the two loss cues was above chance ($t(7) = 3.2$, $p = 0.015$). When we tested the groups separately we found that neither group reached significance alone (Controls: $t(3) = 1.6$, $p = 0.198$; Amygdala: ($t(3) = 2.7$, $p = 0.069$). All the animals chose between the 0 and -1 cue above chance (All animals: $t(7) = 9.6$, $p < 0.001$). This was also true when we looked at each group separately (Controls: $t(3) = 7.2$, $p = 0.005$; Amygdala: $t(3) = 12.2$, $p < 0.001$). This was also the case for choosing between the 0 and -2 cue (All animals: $t(7) = 12.9$, $p < 0.001$; Controls: $t(3) = 8.6$, $p = 0.003$; Amygdala: $t(3) = 24.0$, $p < 0.001$).

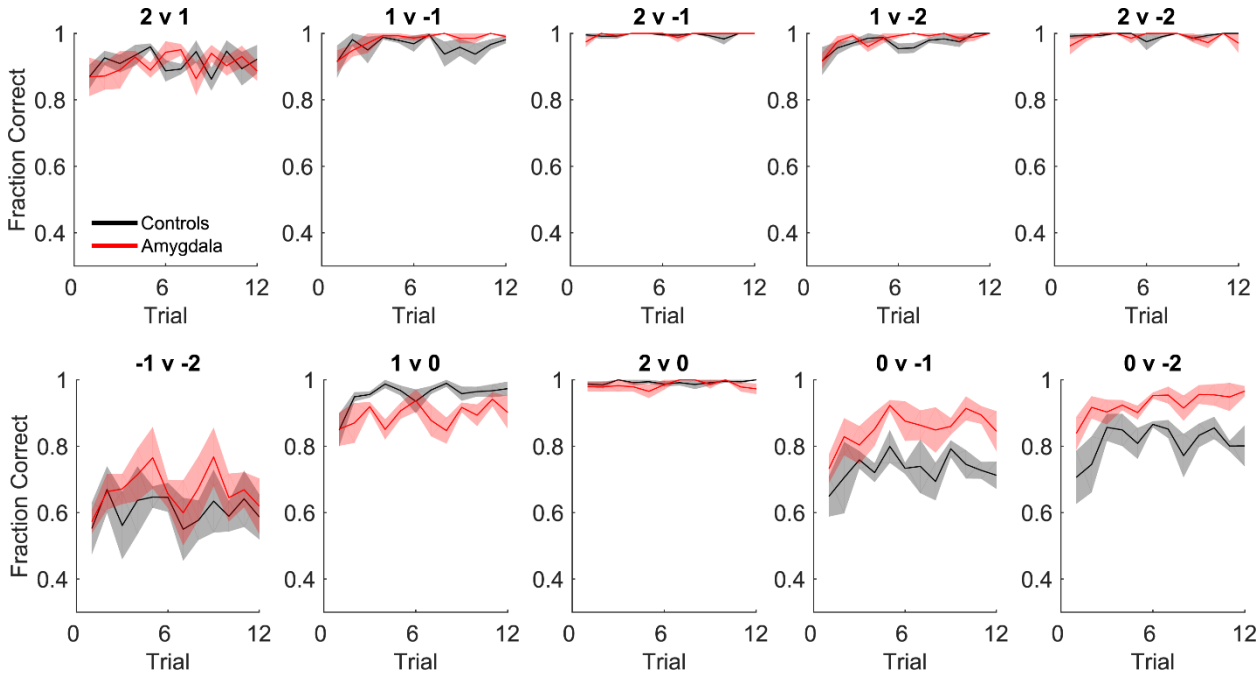


Figure 5. NtK Deterministic Reinforcement Learning Augmented by a Null Cue. Task 2 familiar choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials

TkS

In task 3, we introduced four cues with cue-outcome associations of +2, +1, -1 and -2. However, the cue-outcome associations were stochastic. Therefore, when the animals chose one of the options, they received the outcome associated with that option in 75% of the trials, and no outcome (i.e. no change in tokens) in 25% of the trials. Both novel (Fig. 6) and familiar (Fig. 7) blocks were 108 trials.

Performance was consistent with the previous tasks (Fig. 6). In the novel blocks, there was a difference in performance across conditions (Condition; F(5,

30) = 175.0, p < 0.001) and learning also differed across trials in the different conditions (Condition x Trial; F(85, 515) = 10.7, p < 0.001). There were no overall differences between groups (Group; F(1, 6) = 0.63, p = 0.4585) and no differences between groups across conditions (Group x Condition; F(5, 30) = 0.18, p = 0.969). In this task monkeys showed learning when choosing between the -1 and -2 cues (All animals: t(7) = 4.2, p = 0.003). When we looked at the groups separately we found that only controls learned to choose the smaller loss, (Control: t(3) = 3.3, p = 0.045; Amygdala: t(3) = 3.0, p = 0.055).

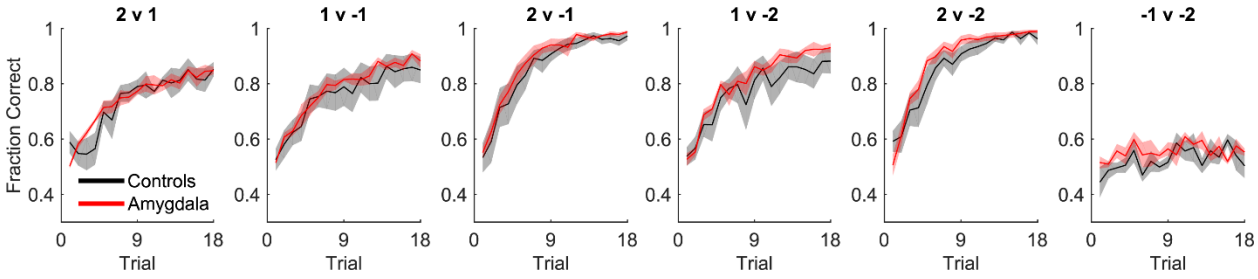


Figure 6. TkS Stochastic Reinforcement Learning. Task 3 novel choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

Performance in the familiar blocks (Fig. 7) was similar to performance in the other tasks. Consistent with the previous experiments, there was a difference in performance across conditions (Conditions; F(5,30) = 128.0, p < 0.001). There were no overall differences between groups (Group; F(1, 6) = 0.95, p = 0.352) and no differences between groups across conditions (Group x Condition; F(5, 30) =

0.53, p = 0.752). The animals learned in the -1 v -2 condition (t(7) = 4.4, p = 0.003). However, when we examined each group separately, only the lesion group performed better than chance (Controls: t(3) = 2.4, p = 0.095; Amygdala: t(3) = 4.3, p = 0.022).

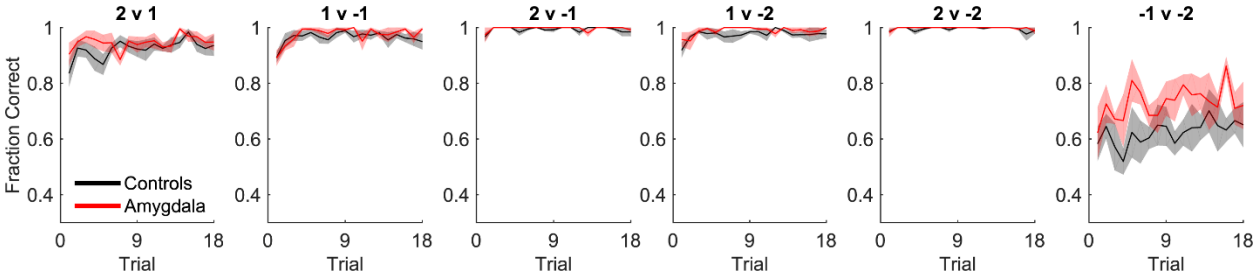


Figure 7. TkS Stochastic Reinforcement Learning. Task 3 familiar choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

TkL

In task 4, we introduced four cues with cue-outcome associations of +2, +1, -1, -4. We added the larger loss cue to see if animals would learn to pick the smaller loss cue more effectively, when the difference between the two loss cues was larger. We also gave the monkeys an endowment of four tokens on the first trial after each cash-out. We did this to ensure that the animals had sufficient tokens to experience the large loss and to maintain motivation. Novel (Fig. 8) and Familiar (Fig. 9) blocks were both composed of 108 trials, with 18 trials per condition.

Performance in novel blocks again showed a difference in performance across conditions (Fig. 8; Condition; $F(5,20) = 78.3$, $p < 0.001$) and a difference in performance across trials in different conditions (Condition x Trial; $F(85,340) = 3.6$, $p < 0.001$). There were no differences between groups (Group; $F(1,4) = 1.13$, $p < 0.3480$) and no difference between groups across conditions (Group x Condition; $F(5, 20) = 0.61$, $p = 0.693$). The monkeys were able to learn to choose the smaller of the two losses ($t(5) = 6.5$, $p = 0.001$). When we examined each group separately we found that only the control animals were able to learn to choose the smaller of the two losses (Control: $t(2) = 5.2$, $p = 0.035$), Amygdala: $t(2) = 3.5$, $p = 0.072$).

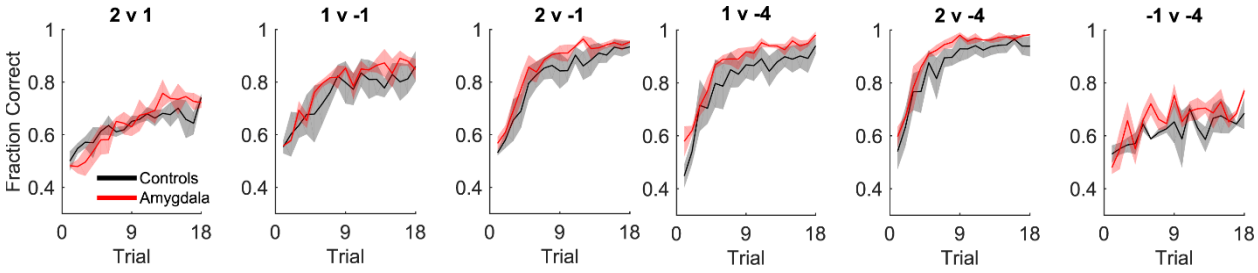


Figure 8. TkL Deterministic Reinforcement Learning with a Large Loss. Task 4 novel choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

In the familiar blocks there were differences in performance across conditions (Fig. 9; Condition; $F(5,20) = 11.8$, $p < 0.001$). There were no overall differences between groups (Group; $F(1, 4) = 0.7$, $p = 0.452$) and no differences

between groups across conditions (Group x Condition; $F(5, 20) = 0.56$, $p = 0.733$). Similar to the novel data the animals were able to pick the smaller of the two losses more often than chance ($t(5) = 11.3$, $p < 0.001$). When we examined the groups individually we found that both groups chose the smaller loss (Controls; $t(2) = 6.7$, $p = 0.022$, Amygdala; $t(2) = 19.9$, $p = 0.003$).

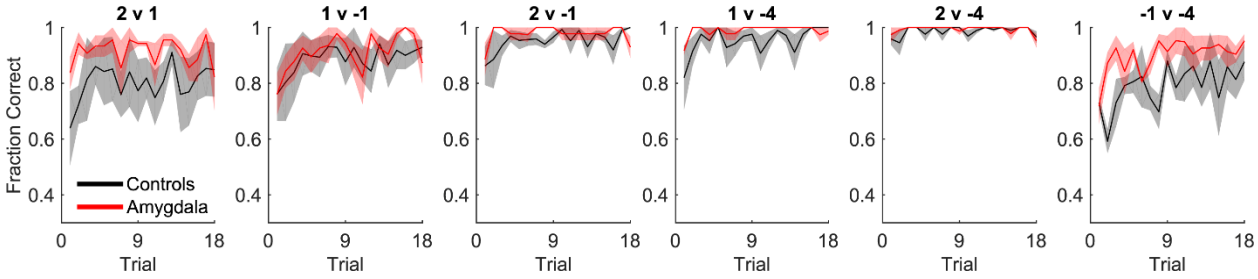


Figure 9. TkL Deterministic Reinforcement Learning with a Large Loss. Task 4 familiar choice behavior. Error bars are +/- s.e.m. with N = number of animals. Plots shows the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

4.3.2 Aborted trials & reaction times

To assess motivation we also examined aborted trials and reaction times. We only present stats and data from task 3 with stochastic outcomes for two reasons. One, the trend of the data was very similar throughout all experiments. Two, task 3 was the only task where we found a group difference.

It was sometimes the case that, when the two images were presented, the animals broke fixation and did not select either image. In both novel (Fig. 10A)

and familiar (Fig. 10B) blocks, the proportion of trials aborted differed by condition (Novel: $F(5,30) = 32.8$, $p < 0.001$; Familiar; $F(5, 30) = 36.8$, $p < 0.001$). Monkeys broke fixation more frequently when they had to choose between the -1 and -2/-4 cues. This was true even though following an error we repeated the same condition. There was no overall difference between groups (Novel: $F(1,6) = 2.92$, $p = 0.138$; Familiar: $F(1,6) = 2.95$, $p = 0.136$). However, the groups differed by condition (Group x condition; Novel: $F(5,30) = 3.58$, $p = 0.012$; Familiar: $F(5,30) = 4.58$, $p = 0.003$). When we examined the conditions individually it became clear that the driving force behind the difference between groups by condition was that lesion animals aborted a higher proportion of trials in the -1 v -2/-4 condition (Novel: $F(1,6) = 8.0$, $p = 0.03$; Familiar: $F(1,6) = 10.8$, $p = 0.017$). It is important to note that even though the -1 v -2/-4 condition was the biggest difference between the groups, the effects do not survive correction.

Next, we examined reaction times. In both novel (Fig. 10C) and familiar (Fig. 10D) blocks, there were differences in reaction times across conditions (Novel: $F(5,30) = 26.2$, $p < 0.001$; Familiar: $F(5,30) = 28.3$, $p < 0.001$). There was a tendency for the lesion animals to respond slower across all task and conditions, however, when we examined the stats we found no differences between groups (Novel: $F(1,6) = 2.02$, $p = 0.205$; Familiar: $F(1,6) = 1.35$, $p = 0.290$) and no higher-order interactions ($p > 0.05$).

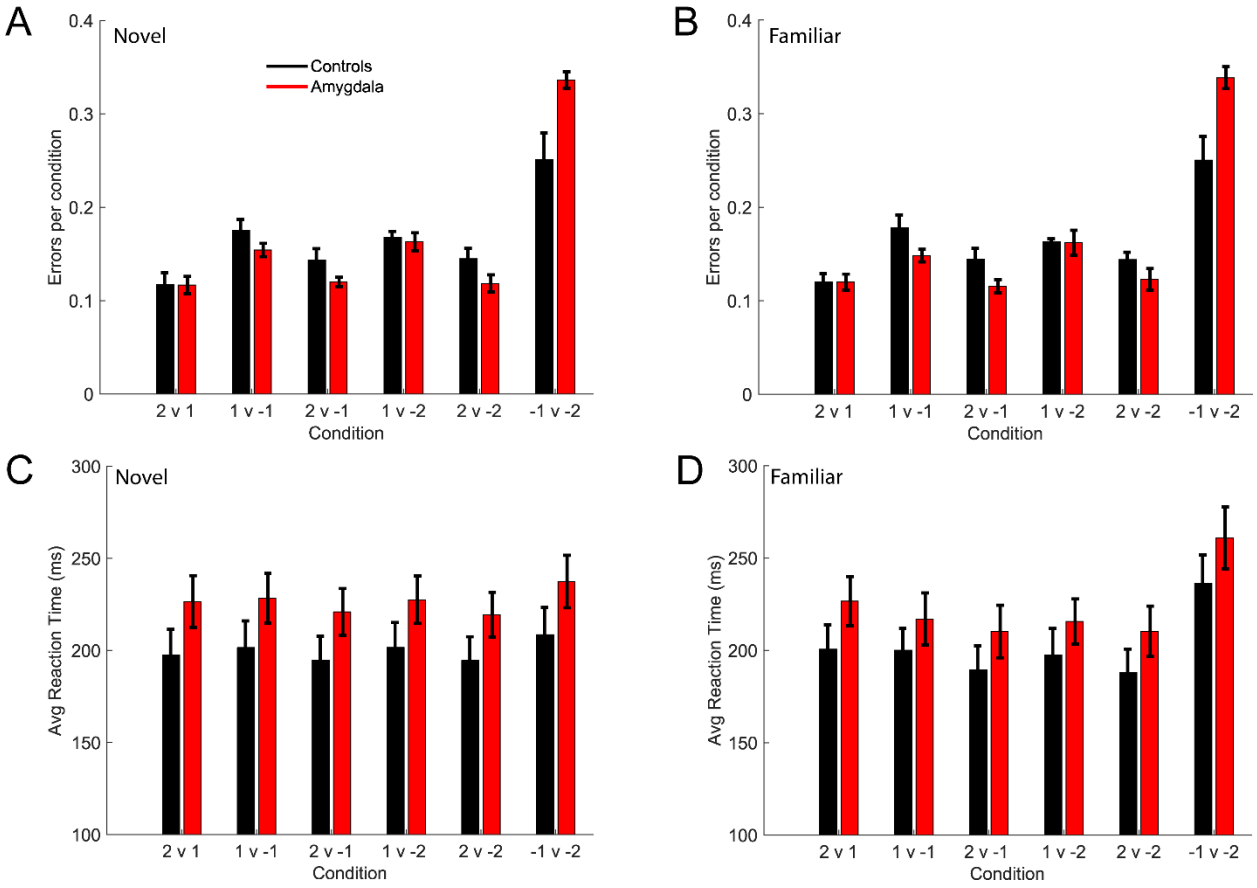


Figure 10. Aborted trials and reaction times for task 3. A. Aborted trials in the novel conditions. Errors indicate the fraction of trials where the animals held initial fixation, but then failed to select one of the choice options. B. Aborted trials in the familiar condition. C. Reaction times in novel conditions. D. Reaction times in familiar conditions.

4.4 Discussion

We carried out four tasks in which we examined learning from gains and losses, using tokens as conditioned reinforcers. Overall, we found that monkeys found gaining tokens to be reinforcing and losing tokens to be punishing. This was evident by the fact that monkeys learned to make choices that increased their tokens and learned to avoid choices that decreased their tokens. In addition to the

learning data, we found other behavioral factors that support the idea that monkeys found losing tokens to be aversive. When monkeys had to make choices between two loss cues, all monkeys aborted significantly more trials. This was the case despite the fact that aborted trials were repeated until they were successfully completed. We also found that all monkeys were significantly slower at making a choice when they had to choose between two losses.

When we examined group differences in learning, we found that monkeys with amygdala lesions had no learning deficits in any of the tasks. In fact, monkeys with lesions to the amygdala performed numerically better than controls in both the novel and familiar blocks in the null token task (NtK). When we analyzed what was driving these group differences, we found that lesion animals performed slightly better in conditions where a loss was paired with either a gain or neutral cue. One interpretation of these results is that the monkeys with amygdala lesions were more sensitive to negative feedback. We have previously found this to be the case in another bandit task where the negative feedback was neutral or no juice (224). Another study found that monkeys with amygdala lesion learned better than unoperated controls from correct trials that followed an error (22), and monkeys with amygdala lesions tend to reverse faster in classical behavioral reversal learning tasks (151).

At first glance our results seem surprising and appear to be counter-intuitive when compared to the current literature, much of which suggests an important role for the amygdala in appetitive (10, 23, 51, 134-136) and aversive conditioning (Davis 1992, Cardinal, Parkinson et al. 2002, Namburi, Beyeler et al, 2015). However, there are several unique things about this task which may explain our findings.

First, the current task is an instrumental conditioning paradigm and not a Pavlovian one. Although there is certainly some overlap between these two forms of conditioning (52), their distinction is quite important. Behaviorally these are separate process, due to the fact that they affect behavior differently. The fact that these forms of conditioning can be separated behaviorally likely means they are separate neural processes. The current literature supports this view, typically assigning a role for the amygdala in the formation of Pavlovian CS-US associations (51, 79) while instrumental conditioning is thought to be controlled by the striatal systems (115). The current experiments are more closely related to the latter form of conditioning.

Second, our task includes both appetitive and aversive cues. There is less evidence on how aversive cues affect instrumental conditioning when there are also appetitive cues. For example it is conceivable that the motivational value of a gain paired with a loss is much higher than a gain paired with a gain. The former

can motivate the animal via positive and negative reinforcement (ie in these conditions the animals are motivated to make the best choice because of the gain in tokens (positive reinforcement), but the animals are also motivated in these conditions to avoid losing tokens (negative reinforcement)). This is unique to the gain/loss conditions and it is possible that this is why we do not find any learning deficits in the monkeys with amygdala lesions. In support of this, we do find that all animals learn and maintain behavior the best in these conditions (gain/loss), but within and across experiments. In addition to this, we ran the same animals with lesions to the amygdala on a different learning task (What/Where) where the most aversive outcome was a lower probability of receiving juice (likely not truly aversive). When we compared the lesion animals to unoperated controls, we found that the lesion animals had learning deficits in both selecting rewarding stimuli (What) and selecting rewarding actions (Where) (Taswell, C, et al, 2020). The fact that this same group of lesion animals had deficits in one task but not the other, suggests that RL is not simply learning the stimulus-outcome relationship. If the amygdala is truly crucial for learning the stimulus-outcome relationship, one would expect deficits in the present study. The fact that we do not find deficits in the present study likely means that there is at least one other component to effective conditioning. We further support this theory with data from animals with lesions to the ventral striatum (VS). We ran the same animals with lesions to the VS on

both the ‘Tokens’ and ‘What/Where’ tasks. When compared to unoperated controls we found similar results to those discussed above for the amygdala animals. In the ‘What/Where task we found that VS animals had major deficits in the ‘What’ condition, which suggested that without the VS, monkeys have problems learning stimulus-outcome relationship. However, when we ran these same VS monkeys on the token experiments, much like amygdala animals, we found that the VS animals were able to learn stimulus-outcome relationships (128).

This idea that performance in an instrumental conditioning task is not solely based on learning the stimulus-outcome relationship is not new, although it is often overlooked. One theory suggests that instrumental behavior is under the influence of two systems. The first system’s role is to learn stimulus-consequence relationship, while the second system is concerned with acquiring the motivational value of the consequence (104). We propose that it is the latter system that is affected by the addition of aversive cues. In this theory aversive cues increase the motivational value of selecting the best option. This would explain how the same animals from the two lesion groups mentioned above have deficits in one task (What/Where) but not the other (Tokens). As one can imagine this separation of value and motivation is quite difficult to prove, most experiments are designed in ways in which it is impossible to separate the two. However, the few studies that have disassociated value from motivation have found that some brain areas, such

as the orbitofrontal cortex and anterior cingulate cortex only carry value signals, while areas such as the premotor cortex only carry motivation signals. And other brain areas, such as VS and cuneus carry both value and motivation signals (225-227). Thus, there is evidence for both a value and motivation system.

Finally, the token task used in the current study employs tokens as conditioned reinforcers. This is in contrast to just providing a primary reward, such as juice, water, or food for correct behavior. In the present study monkeys only receive juice (primary reinforcer) every 4-7 (randomly) trials for the tokens they have accumulated. It is unclear at present, but it is possible the use of tokens as conditioned reinforcers influences the value or motivation system in a way that primary reinforcers do not, and it is this aspect of the task that is responsible for the differences we find in the lesion animals' performance across the two tasks mentioned above.

4.4.1 Conclusion

The present study compared learning from gains and losses, using tokens as conditioned reinforcers, in monkeys with lesions to the amygdala compared to unoperated controls. Across 4 tasks we found that monkeys with amygdala lesions had no learning deficits, and in fact performed better than unoperated controls in the novel and familiar blocks of the null token experiment (NtK). The monkeys

with lesions to the amygdala appeared to be more sensitive to negative feedback, and this is what drove their better performance.

Chapter 5: What behavior can tell us about reinforcement learning

5.1 Introduction

One of the most successful RL theories is the temporal-difference reinforcement learning theory (TD), which suggests that midbrain dopamine codes the temporal difference error from RL (27), which is then used to learn actions that maximize reward and minimize punishment (1, 2). In essence RL theories assume behaving organisms are optimal agents in the computer science/optimal control sense. To understand and make proper conclusions about the RL literature, we need to understand and clarify what it means to learn, and the behavioral components that make up the RL process. From a behavioral perspective there are a couple of things that are problematic for this theory. First, what does it mean to learn something, and how do we measure learning? Is it the case that once an organism learns something their behavior always displays it (ie they are optimal agents)? There is considerable evidence that organisms do not simply behave based on what they know, but instead behave based on internal and external motivational processes. It will become clear throughout this section that the distinction between learning and conditioning is an important one. Behavior is

often controlled by the latter (28, 29). In essence theories of optimality do not properly account for motivation.

It has long been known that there are a number of environmental factors that affect the rate of responding in operant conditioning, which in turn affects performance. For the purposes of this chapter we will focus on two of these factors, reinforcement rate and reinforcement structure. While the former is concerned with the average per trial (or per time interval) rate of reinforcement, the latter is concerned with the reinforcement schedule. There is no understating the importance of understanding how different reinforcement schedules predict different slopes and rates of behavior (28, 29). However, despite the substantial amount of evidence presented in 'Schedules of Reinforcement' (29), which is regarded as one of the most important works in the science of understanding both human and nonhuman behavior, these important environmental factors are often overlooked and not accounted for in the reinforcement learning (RL) literature. Not accounting for how these environmental factors affect behavior can and often does lead to false or overstated conclusions in the RL literature. It is beyond the scope of this chapter to examine all the different types of reinforcement schedules. We instead focus on two, fixed ratios (deterministic) and variable ratios (stochastic). These are two of the more well-known schedules of reinforcement, yet the different affects they have on operant conditioning are often ignored. In

this chapter we take a more traditional behavioral approach to the RL problem, believing that the conflicts in results and differences in theories are products of task designs and misinterpretations of the results they suggest.

Formally there are at least two components to learning: acquisition of the stimulus/behavior-consequence relationship, and the maintenance of that relationship. In traditional learning theories the former is concerned with learning (there is a cap on how fast an organism can learn an association), while the latter has been assigned to motivation. Specifically, once an association is learned an organism can exploit it as much or as little as it wants depending on its motivational level. It is not trivial to dissociate these two components, and it will become clear later, that without dissociating these one can attribute behavioral results to the wrong process. In most cases an experimenter must judge learning from the behavior of the organism, which can be a mix of motivation and learning processes. However, this behavior is not fixed. Many things can alter this behavior, such as reward rate, reward schedule, and reward magnitude (30). The fact that environmental contingencies have such a profound and stable effect (even across species) on behavior suggests, at the very least, that behavior is not simply a display of what one knows. This means that an experimenter needs to design the proper experiments to get at their question (ie most studies suggest they are studying learning, but they are not studying learning in isolation). To know what

aspect of behavior one is studying, one needs to systemically manipulate environmental contingences. This is a prerequisite before one can make any conclusions about the behavior they are viewing.

This is a tough problem to deal with because when it comes to behavior, separating learning rates from motivational factors is not trivial. To understand and separate the environmental effects of learning tasks, we must run a series of tasks, where each task varies by just one parameter. This is how we can separate the motivational effects of the environment from learning ability, thus, make strong statements about learning. In the present study we attempt to account for these problems by conducting a post hoc analysis on the behavioral data presented in chapters 3 and 4.

5.2 Methods

5.2.1 Subjects

The subjects included the unoperated control monkeys from chapters three and four (same control monkeys in both chapters). Subjects also included the monkeys with VS lesion from chapter 3 and monkeys with amygdala lesions from chapter 4.

5.2.2 Task

We conducted a post hoc analysis on the behavioral data in chapters 3 and 4 for three of the token experiments. For reasons stated in the introduction we only use data from 3 of the token experiments, the first (TkD), the third (TkS) and the fourth (TkL). While these three experiments only vary slightly from one another the second token experiment (NtK) varied the most. More importantly it has a different task parameter, which without major assumptions (which at its core this chapter is trying to point out is the problem as to why we are having problems identifying what these neuro structures are doing) makes it difficult to compare to the other task. While TkD, TkS, and TkL have 4 cues and blocks that consist of 108 trials, which means monkeys saw each condition eighteen times, NtK had 5 cues and a block length of a 120 trials, which means monkeys saw each condition twelve times. It is the block length and difference in the number of cues that excluded NtK from this analysis. It should be noted that we did not try to include NtK in this analysis, simply because from a theoretical stand point, it does not make sense. One of the points that the earlier chapters make is that making assumptions about tasks that have different environmental parameters is can lead to problems with interpretation especially when it is not clear how those parameters affect behavior.

5.2.3 Reinforcement learning models

We fit a large set of models that varied in the number of parameters they used to model the conditions. All models were built around a Rescorla-Wagner, or stateless RL value update equation given by:

(1) $$v_i(t+1) = v_i(t) + \alpha_j\big(R - v_i(t)\big).$$

These values were then passed through a soft-max function to give choice probabilities for the pair presented in each trial:

(2) $$d_1(t) = (1 + e^{\beta_k(v_2(t) - v_1(t))})^{-1}, d_2(k) = 1 - d_1(k).$$

The variable v_i is the value estimate for option i, R is the change in the number of tokens that followed the choice in trial t, and α_j is the condition dependent learning rate parameter, for condition j. In addition, we also used, for some models, condition dependent values of the choice consistency parameter, β_k. We then maximized the likelihood of the animal's choices, D, given the parameters, using as a cost function:

(3) $$f\big(D\big|\alpha_j, \beta_k\big) = \prod_t [d_1(k)c_1(k) + d_2(k)c_2(k)].$$

Where $c_1(k)$ was an indicator variable that took on a value of 1 if option 1 was chosen and zero otherwise, and $c_2(k)$ took on a value of 1 if option 2 was chosen and 0 otherwise.

5.2.4 ANOVA models

To quantify differences between choice behavior in each group, we performed an arcsine transformation on the choice accuracy values, as this transformation normalizes the data (152). We then carried out an N-way analysis of variance (ANOVAN). All single condition choice analyses are presented in their uncorrected form. Monkey and session were included as random effects with session nested under monkey. All other factors were fixed effects. Within group analysis was done the same way, the group factor was just dropped.

5.3 Results

We conducted a post hoc analysis on the behavioral data in chapters 3 and 4 for three of the token experiments. For reasons stated in the introduction we only use data from 3 of the token experiments, the first (TkD), the third (TkS) and the fourth (TkL).

5.3.1 Choice Behavior

TkD v TkS v TkL

We started by comparing the novel data for the first (TkD), third (TkS) and fourth (TkL) token experiments (Fig. 1). When we compared the novel behavior in

these experiments across groups, we found that performance differed across experiment (Experiment; $F(2,14) = 20.9$, $p < 0.001$). We also found difference in performance across conditions (Experiment x Condition; $F(10,70) = 27.9$, $p < 0.001$), trials (Experiment x Trial; $F(34,238) = 2.1$, $p < 0.001$), and in trials in conditions (Experiment x Condition x Trial; $F(170,1190) = 1.6$, $p < 0.001$) across experiments. We found no groups effects, which suggest that all groups modulated their behavior similarly across experiments. Next, to get a better idea of what conditions varied across the three experiments, we analyzed each condition separately. All effects are presented in their uncorrected form, but we only consider effects to be significant if they correct. When we looked at each condition separately across the groups, we found that performance in the 2 v 1 condition varied across experiments (Experiment; $F(2,14) = 26.9$, $p < 0.001$), and trials by experiment (Experiment x Trial; $F(34,238) = 2.1$, $p = 0.001$). There were no group effects in the 2 v 1 condition, but there was a trend for groups to differ by trial (Group x Trial; $F(34,132) = 1.7) = 0.018$), although this effect does not survive correction. The 1 v -1 condition did not survive correction, this was the only condition where performance did not vary by experiment (Experiment; $F(2,14) = 6.7$, $p = 0.009$). There was, however, a trend for groups to differ in this condition across experiments (Group x Experiment; $F(4,14) = 3.2$, $p = 0.047$), which seems to be the result of VS animals not modulating their behavior the same

way as the controls and amygdala animals in this condition. This was also the case for the 2 v -1 condition, which varied across experiments (Experiment; $F(2,14) = 32.9$, $p < 0.001$), and had a trend for groups' performance to differ by experiment (Group x Experiment; $F(4,14) = 3.3$, $p = 0.043$). Performance in the 2 v -2/-4 condition only varied across experiments (Experiment; $F(4,14) = 16.1$, $p < 0.001$). Performance in the 1 v -2/-4, and -1 v-2/-4 conditions varied across experiments (1 v -2/-4: Experiment; $F(2,14) = 22.9$, $p < 0.001$; -1 v -2/-4: Experiment; $F(4,14) = 57.7$, $p < 0.001$), and trials by experiments (1 v -2/-4: Experiment x Trial; $F(34,238) = 1.8$, $p = 0.008$; -1 v -2/-4: Experiment x Trial; $F(34,238) = 2.3$, $p < 0.001$).

Controls

To see to what extent each group performed differently across the experiments, we completed separate within group analysis for each group across experiments. For the control animals we found that performance varied across experiments (Experiment; $F(2,4) = 8.1$, $p = 0.04$). We also found that this difference in performance across experiments, differed by condition (Experiment x Condition; $F(10,20) = 6.3$ $p < 0.001$) and trials (Experiment x Trial; $F(34,68) = 1.6$, $p = 0.048$). The three way interaction of experiment by condition by trial was not significant (Experiment x Condition x Trial; $F(170,340) = 1.1$, $p = 0.232$).

VS

For the VS monkeys we found a main effect of experiment (Experiment; $F(2,4) = 7.0$, $p = 0.049$) and an interaction of experiment by condition (Experiment x Condition; $F(10,20) = 11.3$, $p < 0.001$). VS animals did not perform differently across trials in different experiments (Experiment x Trial; $F(34,68) = 0.54$, $p = 0.973$), but did perform differently across trials in different conditions by experiment (Experiment x Condition x Trial; $F(170,340) = 1.3$, $p = 0.013$).

Amygdala

For the amygdala monkeys we found that performance did not differ by experiment (Experiment; $F(2,4) = 5.1$, $p = 0.078$), but there was an interaction of experiment by condition (Experiment x Condition; $F(10,20) = 16.4$, $p < 0.001$). The amygdala animals' performance did not differ across trials by experiment (Experiment x Trial; $F(34,68) = 1.0$, $p = 0.477$) and there was no three way interaction of experiment by condition by trial (Experiment x Condition x Trial; $F(170,340) = 1.0$, $p = 0.431$).

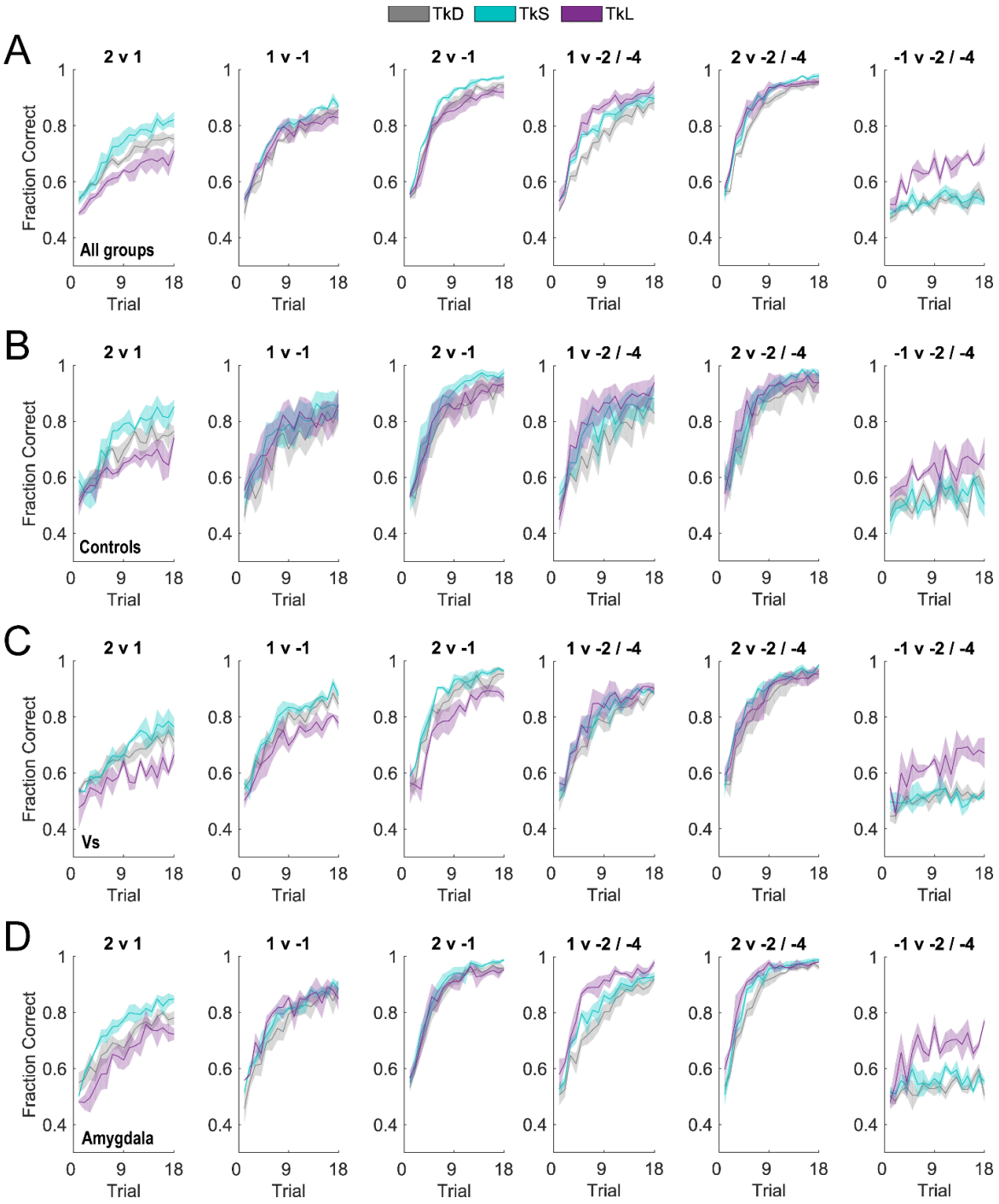


Figure 2. Reinforcement Learning of Stimulus-Outcome Associations across the 3 tasks. A. Within group performance averaged across groups for each of the 3 task. Error bars are +/- s.e.m. with N = number of groups. B. Average performance for the control animals across tasks. C. Average performance for the VS animals across tasks. D. Average performance for amygdala animals across tasks. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

The analysis above shows that for all groups performance varies with experiment. However, because there are 3 experiments with slight variations in their reward environment, these analyses are not very informative in regards to what is driving these differences in performance across the experiments. In the following sections we complete pairwise comparisons of each experiment.

TkD v TkS

To get a better idea of what was driving this difference in performance across experiments, we first compared performance in the first token experiment (TkD) against the performance in the third token experiment (TkS). These two experiments were the most similar, and only had two differences. The first difference is reward schedule, TkD had a deterministic reward schedule, while TkS had a stochastic reward schedule. The second difference is a result of the first and has to do with the reward rate, TkD had a higher reward rate when compared to TkS. When compared to un-operated controls we found no differences in performance in the TkD experiment for VS (chapter 3) or amygdala monkeys (chapter 4). For the TkS experiment we did find that VS monkeys performed worse than controls (chapter 3), while there was no difference between the controls and amygdala monkeys (chapter 4).

When we examined the performance for all three groups across experiments, we found that all groups performed better in TkS (Experiment; $F(1,8) = 33$, $p <$

0.001). This was also true on a condition (Experiment x Condition; $F(5,40) = 3.9$, $p = 0.005$) and trial (Experiment x Trial; $F(17,136) = 1.9$, $p = 0.018$) basis. The three way interaction, however, was not significant (Experiment x Condition x Trial; $F(85,680) = 0.9$, $p = 0.571$). Again we found no group effects, which provides further support that all groups performed better in the TkS experiment (to varying degrees). To get a better idea of what conditions drove this better performance in the TkS experiment, we analyzed each condition separately across experiments and groups. When we looked at each condition separately we found that monkeys performed better in every condition in the TkS experiment when compared to the TkD experiment, with the exception of the -1 v -2 condition. In the 2 v 1 condition there was a main effect of experiment (Experiment; $F(1,8) = 12.5$, $p = 0.008$) and an experiment by trial interaction (Experiment x Trial; $F(17,136) = 2.3$, $p = 0.004$). For conditions 2-5 we found main effects of experiment (1 v -1: Experiment; $F(1,8) = 15.5$, $p = 0.004$; 2 v -1: Experiment; $F(1,8) = 33.9$, $p < 0.001$; 1 v -2: Experiment; $F(1,8) = 13.9$, $p = 0.006$; 2 v -2: Experiment; $F(1,8) = 19$, $p = 0.002$). The experiment by trial interaction was not significant in any of these conditions. There were also no group effects in any of the conditions.

Controls

Next, we performed within group analyses to see to what extent each group contributed to the overall better performance in the TkS experiment. For control monkeys we found that they performed better in the TkS experiment (Experiment; $F(1,3) = 13.5$, $p = 0.035$). In addition to this effect, we also found that control monkeys performed better in TkS across conditions (Experiment x Condition; $F(5,15) = 4.5$, $p = 0.010$) and trials (Experiment x Trial; $F(17,51) = 2.0$, $p = 0.026$). No higher order interactions reached significance. Following this, we looked at each condition separately to see what conditions were driving this better performance in the TkS experiment. We found that control monkeys performed slightly better in the TkS experiment in the 2 v 1 (Experiment; $F(1,3) = 5.7$, $p = 0.097$), 1 v -1 (Experiment; $F(1,3) = 14.1$, $p = 0.033$), 2 v -1 (Experiment; $F(1,3) = 21.8$, $p = 0.019$), 1 v -2 (Experiment; $F(1,3) = 5.8$, $p = 0.094$) and 2 v -2 (Experiment; $F(1,3) = 12.7$, $p = 0.038$) conditions. We found no trend for differences across experiments for the -1 v -2 (Experiment; $F(1,3) = 0.04$, $p = 0.641$). No trial by experiment interaction reached significance alone, but there was a trend for the control animals performing better in TkS on a trial by trial basis for the 2 v 1 (Experiment x Trial; $F(17,51) = 2$, $p = 0.029$) condition. None of these effects survive correction individually, thus the better performance in TkS is due to control monkeys performing slightly better in each of the conditions above.

VS

When we looked at the plot for VS monkeys' performance in TkD and TkS, it appeared that the VS monkeys also performed overall better in TkS. However, that was not the case, when we competed our within group analysis for VS monkeys, we found that performance in TkD and TkS did not differ (Experiment; $F(1,3) = 6.8$, $p = 0.120$). In addition, there was no trial by experiment interaction (Experiment x Trial; $F(17,34) = 0.4$, $p = 0.978$), no condition by experiment interaction (Experiment x Condition; $F(5,10) = 0.6$, $p = 0.674$), and no three way interaction of experiment by condition by trial (Experiment x Condition x Trial; $F(85,170) = 0.9$, $p = 0.734$).

Amygdala

We found that amygdala monkeys followed the same trend as the control animals, and performed overall better in the TkS experiment (Experiment; $F(1,3) = 17.3$, $p = 0.025$). No higher order interactions involving experiment reached significance. After examining each condition independently, we found that amygdala monkeys performed slightly better in the TkS experiment in every condition. The conditions with the biggest difference between the two experiments were the 2 v 1 (Experiment; $F(1,3) = 7.8$, $p = 0.068$), the 2 v -1 (Experiment; $F(1,3) = 14.5$, $p = 0.032$), the 2 v -2 (Experiment; $F(1,3) = 10.6$, $p = 0.047$), and the -1 v -2 (Experiment; $F(1,3) = 7.8$, $p = 0.068$).

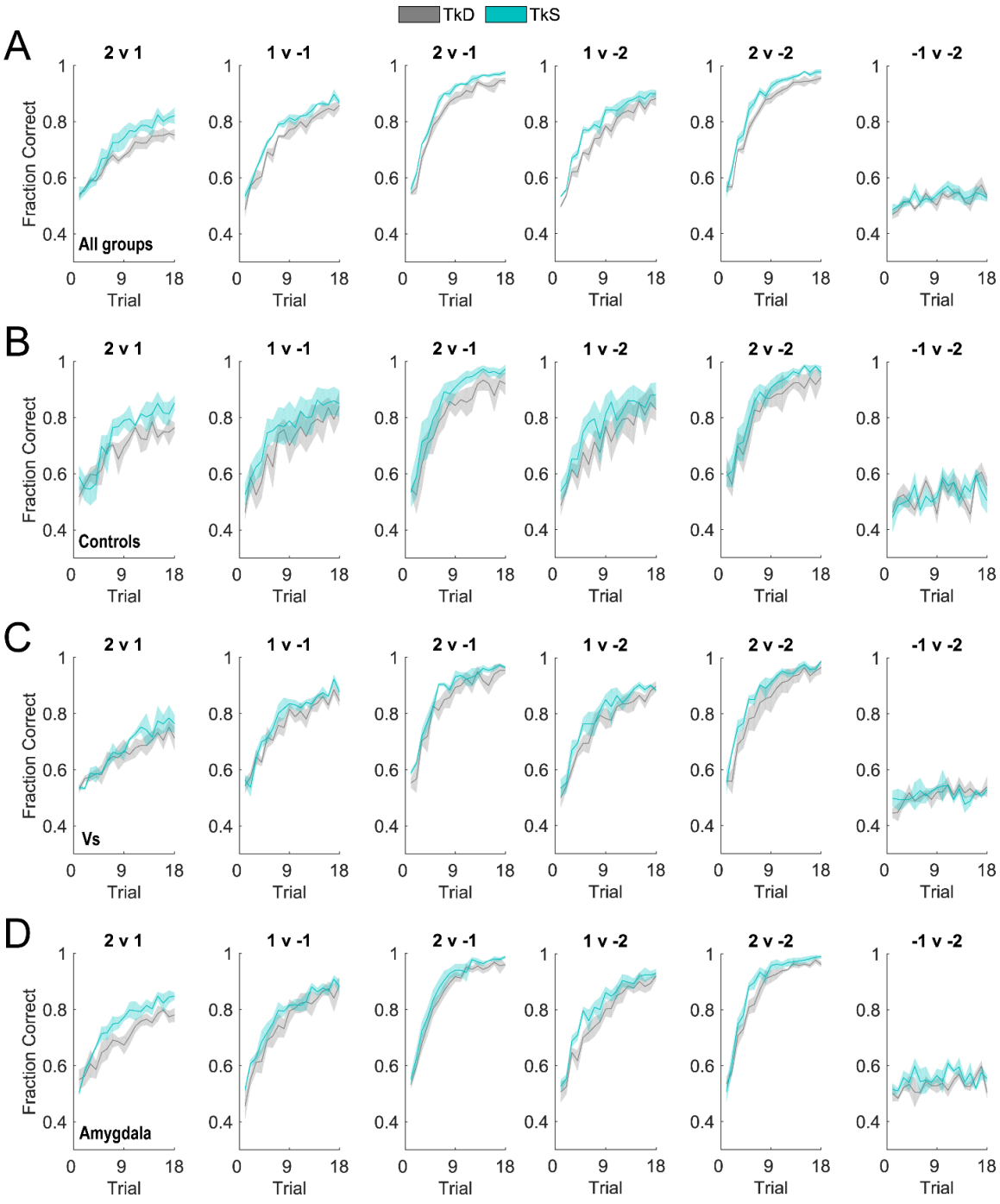


Figure 3. Reinforcement Learning of Stimulus-Outcome Associations across the TkD & TkS tasks. A. Within group performance averaged across groups for each of the tasks. Error bars are +/- s.e.m. with N = number of groups. B. Average performance for the control animals across tasks. C. Average performance for the VS animals across task. D. Average performance for amygdala animals across task. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

This analysis revealed that as a whole, all animals from all groups trended toward better performance in TkS when compared to TkD, with the only difference being the level of performance modulation. While all groups performed better, they did not do this equally. Thus the difference between the groups lies mainly in their performance modulation between the two experiments. Both the control and amygdala monkeys perform significantly better in the TkS, while the VS animals only performed slightly better. In chapter three when we compared the control monkeys to the VS monkeys we found that the control monkeys performed better than VS monkeys in the TkS experiment. This point taken with the fact that we found no difference between the groups in the TkD experiment, suggest that the effect we found in chapter three, in the TkS experiment, is due to control monkeys improving their choice behavior more than VS monkeys did under stochastic schedules.

In chapter four when we compared controls to the amygdala monkeys we found no difference in performance across the groups, which follows because the amygdala monkeys improve their choice behavior as well as control animals under the stochastic schedule of TkS. It should be noted that if these differences in performance across experiments is in fact due to learning and not some other component, like motivation, animals should not

be able to perform better in the TkS experiment. It should be harder to learn a stochastic cue-reward association when compared to a deterministic cue-reward association.

TkS v TkL

Next, because TkD was the first experiment and TkS was the third experiment, we postulated that this could be a training effect. So we compared TkS, the third experiment, to TkL, the fourth experiment (Fig. 3). These two experiments differed in three ways. The first way was the reward schedule, TkL had a deterministic reward schedule, while TkS had a stochastic reward schedule. The second way these two experiments differed is by reward rate, in TkL we gave monkeys an endowment of 4 tokens at the beginning of the session, and after each cash-out during the session. The final way these experiments differed was with the value of the cues. In the TkL experiment we replaced the -2 cue with a cue worth -4.

Similar to the comparison above, we first examined the effect of experiment on performance for all monkeys. We found that performance across the two experiments did not differ (Experiment; $F(1,6) = 3.9$, $p = 0.095$). However, we did find that across the two experiments performance did differ across conditions (Experiment x Condition; $F(5,30) = 39.7$, $p < 0.001$), trials (Experiment x Trial; $F(17,102) = 2.0$, $p = 0.016$) and by trials in conditions (Experiment x Condition x

Trial; $F(85,510) = 2.2$, $p < 0.001$). Again, we did not find any group effects, which indicates that all groups modulated their behavior in the same direction. When we examined the individual conditions across groups, we found that monkeys performed better in TkS in the 2 v 1 condition (Experiment; $F(1,6) = 45.2$,$p < 0.001$; Experiment x Trial; $F(17,102) = 2.8$, $p < 0.001$). We also found a trend for groups to differ in this condition across trials (Group x Trial; $F(34,126) = 1.6$, $p = 0.025$). Performance across the experiments did not differ in the 1 v -1 condition (Experiment; $F(1,6) = 5.6$, $p = 0.055$). In the 2 v -1 condition we found that monkeys performed overall better in the TkS experiment (Experiment; $F(1,6) = 53.16$, $p < 0.001$). The experiment by trial interaction for this condition did not survive correction. The next 3 conditions have a value difference in favor of the TkL experiment. Despite this value difference we find no differences (after correction) in performance across experiments in the in the 1 v -2/-4 condition (Experiment; $F(1,6) = 6.8$, $p = 0.04$). This was also the case for the 2 v -2/-4 condition (Experiment; $F(1,6) = 2.4$, $p = 0.174$). However, in this condition, we did find an experiment by trial interaction (Experiment x Trial; $F(17,102) = 2.3$, $p = 0.004$). Monkeys performed overall better in the TkL experiment in the -1 v -2/-4 condition (Experiment; $F(1,6) = 56.7$, $p < 0.001$) condition. This difference also differed across trials (Experiment x Trial; $F(17,102) = 2.6$,$p = 0.002$).

Controls

When we looked at the groups separately across tasks, we found no overall difference in performance across experiments (Experiment; $F(1,2) = 1$, $p = 0.423$) for control monkeys. There was, however, a difference in performance across conditions (Experiment x Condition; $F(5,10) = 6.9$, $p = 0.005$), and in performance across trials (Experiment x Trial; $F(17,34) = 2.0$, $p = 0.038$), but no three way interaction of experiment by condition by trial (Experiment x Condition x Trial; $F(85,170) = 1.2$, $p = 0.169$). When we looked at each condition separately, we found that control monkeys trended toward better performance in the TkS experiment in the 2 v 1 (Experiment x Trial; $F(17,34) = 2.6$, $p = 0.009$), and in the 2 v -1 (Experiment; $F(1,2) = 9.3$, $p = 0.092$) conditions. We found a trend for control animals to perform better in TkL in the 2 v -2/-4 (Experiment x Trial; $F(17,34) = 1.8$, $p = 0.078$), and in the -1 v -2/-4 (Experiment; $F(1,2) = 11$, $p = 0.080$). We found no trend for differences across experiments in any other condition.

VS

When we examined the VS animals' performance across the experiments, we found no main effect of experiment (Experiment; $F(1,2) = 7.2$, $p = 0.115$). However, we did find that VS monkeys' performance across conditions (Experiment x Condition; $F(5,10) = 24.4$, $p < 0.001$), trials (Experiment x Trial; $F(17,34) = 5.9$, $p < 0.001$), and trials by conditions (Experiment x Condition x

Trial; $F(85,170) = 1.8$, $p < 0.001$) differed by experiment. When we looked at each condition independently, we found that VS monkeys trended toward better performance better in the TkS experiment in the 2 v 1 (Experiment; $F(1,2) = 35.5$, $p = 0.027$), 1 v -1 (Experiment; $F(1,2) = 81.1$, $p = 0.012$), 2 v -1 (Experiment; $F(1,2) = 109.7$, $p = 0.009$). In addition to these trends for main effects of experiment, the 2 v 1 and 1 v-1 trended toward being better on a trial by trial bias (2 v 1: Experiment x Trial; $F(17,34) = 2.4$, $p = 0.016$; 1 v -1: Experiment x Trial; $F(17,34) = 2.2$, $p = 0.024$). We found a trend for VS animals to perform both overall and on a trial by trial basis better in TkL in the -1 v -2/-4 condition (Experiment; $F(1,2) = 19.0$, $p = 0.049$; Experiment x Trial; $F(17,34) = 1.9$, $p = 0.048$). There were, no other effects or trends in any other condition.

Amygdala

For the amygdala animals there was no overall differences across experiments (Experiment; $F(1,2) = 0.02$, $p = 0.889$), or in trials across experiments (Experiment x Trial; $F(17,34) = 0.7$, $p = 0.789$). However, we did find that performance in conditions differed across experiments (Experiment x Condition; $F(5,10) = 25.2$, $p < 0.001$). The three way interaction of experiment by condition by trial was not significant (Experiment x Condition x Trial; $F(85,170) = 1.2$, $p =$

0.191). When we looked at each condition separately, we found that amygdala monkeys performed better in the TkS experiment in the 2 v 1 (Experiment; $F(1,2) = 134.4$, $p = 0.007$), and trended toward performing better in the TkL experiment in the -1 v -2/-4 (Experiment; $F(1,2) = 50.9$, $p = 0.019$). It looked like there was at least a trend for amygdala animals performing better in the TkL experiment in the 1 v -2/-4 condition, but this was not the case (Experiment; $F(1,2) = 4.8$, $p = 0.161$). There were, no other effects or trends in any other condition.

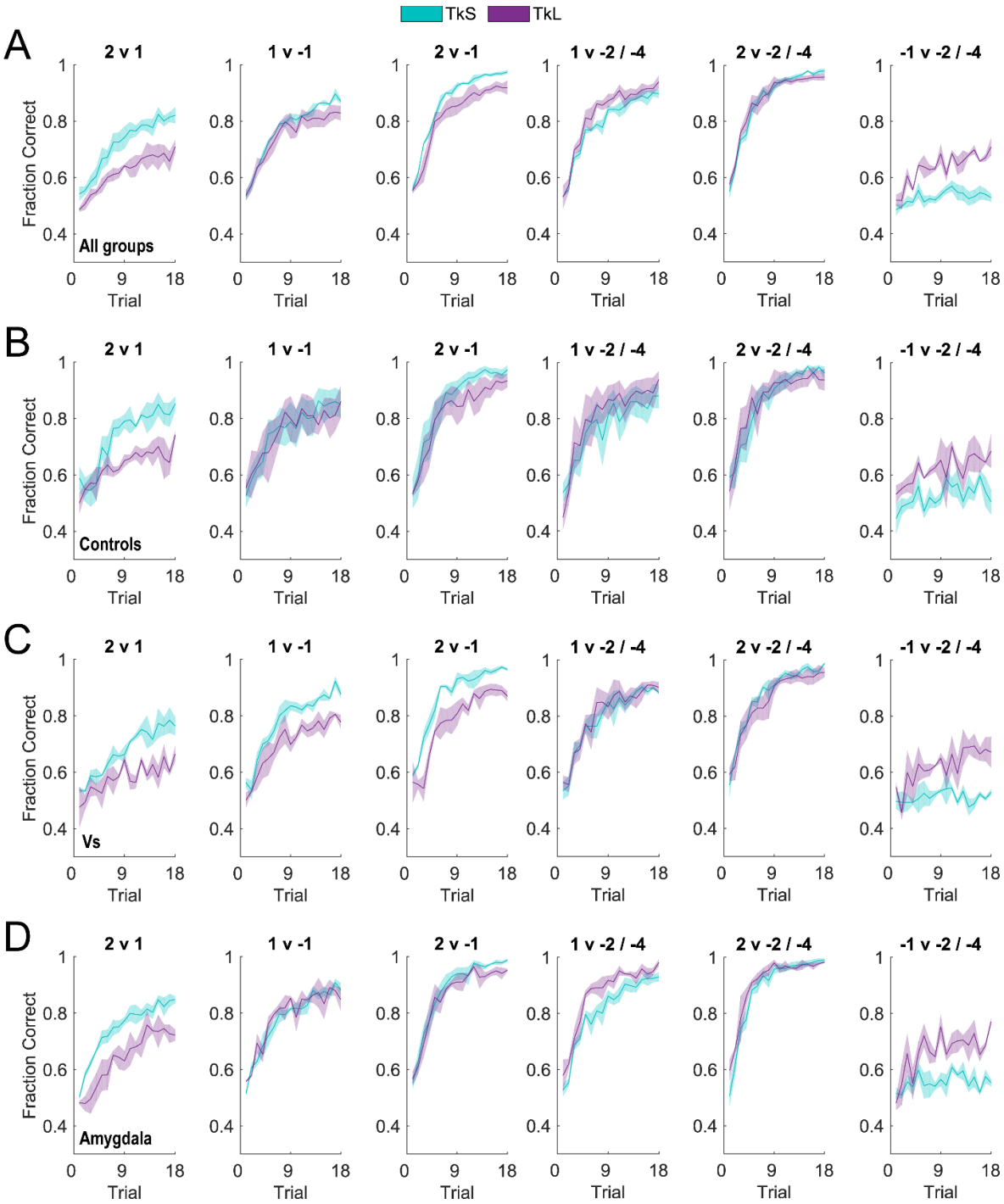


Figure 4. Reinforcement Learning of Stimulus-Outcome Associations across the TkS & TkL tasks. A. Within group performance averaged across groups for each of the tasks. Error bars are +/- s.e.m. with N = number of groups. B. Average performance for the control animals across tasks. C. Average performance for the VS animals across task. D. Average performance for amygdala animals across task. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

This analysis revealed several things. First, despite the order of experiments and the value difference between the two experiments, monkeys don't perform overall better in the TkL experiment. In fact we find that monkeys performed better in the TkS experiment in the 2 v 1 and 2 v -1 conditions. This suggest that the difference we found in the TkD v TkS comparison was not a result of training, but instead of a more basic behavior principal that underlies stochastic versus deterministic reward schedules. Second, this analysis reveals that cue value does indeed matter, its just not the only component animals take into account. This can be seen by the fact that monkeys trended toward performing better in the TkS experiment in the three conditions without the bigger value difference. In the conditions with the bigger value difference, we do find that monkeys trended toward better performance in the TkL experiment. The important thing to remember about these experiments is that the cues are the same for each condition throughout a block, thus these condition effects within and across experiments indicate that performance is not simply based cue acquisition. Again, the difference between the groups lies in their level of behavior modulation. In particular, the VS animals seem to discount more in the conditions without the value difference, when compared to control and amygdala animals. This is consistent with the results we find in chapter three and chapter four. In chapter three when we compared control monkeys to VS monkeys, we found that VS

monkeys had performance deficits. In chapter four when we compared control monkeys to amygdala monkeys we found no deficits.

TkD v TkL

Finally, we compared performance in the first experiment (TkD) to the performance in the fourth experiment (TkL). Both of these experiments were on deterministic reward schedules, with TkL having a bigger value difference between the cues and a much richer reward environment (because of the endowment).

When we compared these two experiments across groups we found that monkeys performed overall better in the TkL experiment (Experiment; $F(1,6) = 22.5$, $p = 0.003$). We found that this effect of experiment differed by condition (Experiment x Condition; $F(5,30) = 29.4$, $p < 0.001$), trial (Experiment x Trial; $F(17,102) = 1.9$, $p = 0.019$), and trials by conditions (Experiment x Condition x Trial; $F(85,510) = 1.7$, $p < 0.001$). In addition to these effects, for the first time in any of the task comparisons, we found that the performance difference across experiment differed by group (Group x Experiment; $F(2,6) = 7.8$, $p = 0.019$). No other group effects were significant. When we analyzed the conditions individually, we found that monkeys performed better in the TkD experiment in the 2 v 1 condition (Experiment; $F(1,6) = 18.1$, $p = 0.005$). There were no higher order interactions or group effects. When we looked at the 1 v -1 condition across experiments, we found no difference in performance (Experiment; $F(1,6) = 0.01$, p

= 0.944). Again, there were no higher order interactions or group effects. For the 2 v -1 condition there was a trend for monkeys to perform better in the TkD experiment, however this effect did not survive correction (Experiment; $F(1,6) = 6.0$, $p = 0.049$). As with the previous conditions there were no higher order interactions or group effects. In the next three conditions where there was a value difference between the two experiments. We found that monkeys performed better in TkL conditions (1 v -2/-4: Experiment; $F(1,6) = 85.9$, $p < 0.001$; 2 v -2/-4: Experiment; $F(1,6) = 24.1$, $p = 0.002$; -1 v -2/-4: Experiment; $F(1,6) = 54.5$, $p < 0.001$). In addition to these main effects of experiment, we found experiment by trials interactions for the 1 v – 2/-4 (Experiment x Trial; $F(17,102) = 2.4$, $p = 0.003$), -1 v -2/-4 (Experiment x Trial; $F(17,102) = 3.3$, $p < 0.001$), but not the 2 v -2/-4 (Experiment x Trial; $F(17,102) = 1.6$, $p = 0.068$).

Controls

In our within group analysis we found an overall trend for better performance in the TkL experiment for control animals (Experiment; $F(1,2) = 15.9$, $p = 0.057$). However, this effect was not significant, because the better performance in TkL that drove this trend is based off much better performance in the conditions with the value difference. Thus, we found an condition by experiment interaction (Experiment x Condition; $F(5,10) = 6.9$, $p = 0.005$). There was no trial by experiment interaction (Experiment x Trial; $F(17,34) = 1.4$, $p =$

0.202), nor was there the three way interaction with condition (Experiment x Condition x Trial; $F(85,170) = 1.1$, $p = 0.318$). Next, we examined each condition separately. It looked like there was a trend for control animals to perform better in the TkD experiment in the 2 v 1 condition, but this was not the case (Experiment; $F(1,2) = 1.8$, $p = 0.313$; Experiment x Trial; $F(17,34) = 0.9$, $p = 0.570$). We found trends for control monkeys performing better in TkL for the conditions with the value difference (1 v -2/-4: Experiment; $F(1,2) = 42.55$, $p = 0.023$; 2 v -2/-4: Experiment; $F(1,2) = 9.7$, $p = 0.089$; -1 v -2/-4: Experiment; $F(1,2) = 13.7$, $p = 0.066$). The interactions with trials was less significant for all of these conditions.

VS

For the VS monkeys we found no effect of experiment (Experiment; $F(1,2) = 4.3$, $p = 0.175$). There was, however, a difference in conditions across experiments (Experiment x Condition; $F(5,10) = 11.8$, $p < 0.001$), but not in trials across experiments (Experiment x Trial; $F(17,34) = 0.6$, $p = 0.873$). The three way interaction of experiment by condition by trial was also not significant (Experiment x Condition x Trial; $F(85,170) = 0.077$). When we examined each condition separately, we found that there were trends for VS monkeys performing better in the TkD experiment in the 2 v 1 (Experiment; $F(1,2) = 72.0$, $p = 0.014$), and the 2 v -1 (Experiment; $F(1,2) = 16.5$, $p = 0.056$) conditions. We found that VS monkeys performed better in the TkD experiment in the 1 v -1 condition (Experiment; F(1,2)

= 179.7, p =0.005). We found trends for VS monkeys performing better in the TkL experiment in the 1 v -2/-4 (Experiment; F(1,2) = 9.6, p = 0.090) and in the -1 v -2/-4 (Experiment; F(1,2) = 8.6, p = 0.099) conditions. The experiment by trial interaction was not significant for any of these conditions.

Amygdala

We found no overall difference in performance across experiments (Experiment; F(1,2) = 10.2, p = 0.086) for the amygdala animals. We did find that performance in conditions varied by experiment (Experiment x Condition; F(5,10) = 14.9, p < 0.001), but not trials across experiments (Experiment x Trial; F(17,34) = 1.1, p = 0.356). The three way interaction was also not significant (Experiment x Condition x Trial; F(85,170) = 1.1, p = 0.248). When we examined each condition separately, we found that amygdala monkeys performed better in the TkD experiment in the 2 v 1 condition (Experiment; F(1,2) = 148.9, p = 0.006). There was a trend for amygdala animals to perform better in the TkL experiment in the 1 v -2/-4 (Experiment; F(1,2) = 45.6, p = 0.021), and in the 2 v -2/-4 (Experiment; F(1,2) = 72.2, p = 0.014) conditions. Amygdala animals performed significantly better in the TkL experiment in the -1 v -2/-4 (Experiment; F(1,2) = 445.0, p = 0.002) condition. The experiment by trial interaction was not significant for any of these conditions.

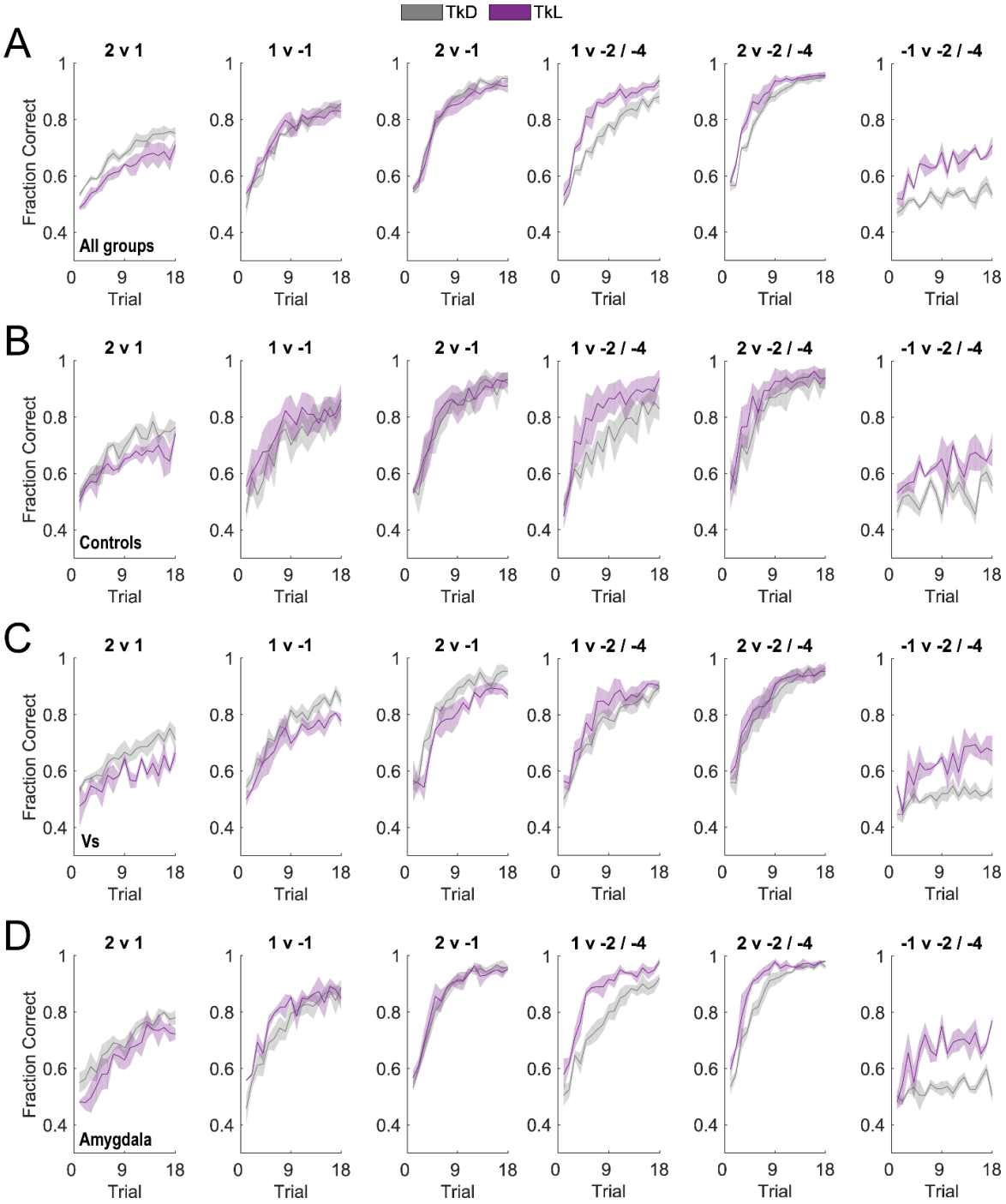


Figure 5. Reinforcement Learning of Stimulus-Outcome Associations across the TkD & TkL tasks. A. Within group performance averaged across groups for each of the tasks. Error bars are +/- s.e.m. with N = number of groups. B. Average performance for the control animals across tasks. C. Average performance for the VS animals across task. D. Average performance for amygdala animals across task. Error bars are +/- s.e.m. with N = number of animals. Plots show the fraction of times monkeys chose the higher value option averaged across novel blocks for each group. Numbers at the top of each plot indicate the condition, which corresponds to the cues which were shown in those trials.

Despite the order of experiments, and the value difference between the two experiments, monkeys still perform better in some of the TkD conditions. The TkD and TkL experiments were both on deterministic reward schedules, so why do monkeys still perform better in some TkD conditions? The only other difference between these two experiments, besides the value difference, is that TkL has a much richer reward environment, due to the endowment we give the monkeys. We posit that in leaner reward environments (when everything else is held constant) monkeys become greedier in the leaner environment. Each of their choices becomes more valuable in the leaner reward environment. This can be illustrated with the 2 v 1 condition, which seems to be the condition that modulates the most across experiments. The difference between the groups in this comparison, again, seems to be that VS animals discount their choice behavior much more in the richer reward environment (TkL), when compared to the other two groups, which is shown by the group by experiment effect we found.

5.3.2 Cash-out

Fundamentally these three tasks only differ in a few ways. We posit that performance across tasks varies with one or both of the following parameters of the task environment. The two parameters we will discuss are reward rate and reward schedule. The former is concerned with the average rate of reward, the latter is

concerned with the reward schedule. The tasks in this analysis either employ a fixed (deterministic) or variable (stochastic) schedule of reinforcement. In each of these tasks, the conditions within a block of 108 trials were presented pseudo-randomly. The animals saw each condition twice, once on the left and once on the right, every twelve trials, before seeing any condition a third time. This information allowed us to find the maximum number of tokens monkeys could get every twelve trials, if they chose the higher value option in each of these trials. From this we could estimate a maximum per trial token rate. Since we cashed out monkeys' accumulated tokens every four to seven trials randomly, we could multiply this per trial token rate by the average number of trials in a cash out period (5.5), to get an average estimate of the maximum number of tokens monkeys could have per cash out (Fig. 6A). Next, we found the average number of tokens monkeys had at cash outs (Fig.6B), and divided this value by the average maximum number of tokens monkeys could have at cash out, to get the average proportion (of maximum tokens) of tokens monkeys had per cash out (Fig. 6C).

When we looked at the average proportion of tokens monkeys had per cash out across tasks (Fig. 6D), we found that these proportions differed by experiment (Experiment; $F(2,14.5) = 317.6$, $p < 0.001$). We also found that this difference in proportions by experiment, differed by group (Group x Experiment; $F(4,14.5) = 3.4$, $p = 0.0350$). Next, to see what was driving this group effect, we looked at

each group separately across the three experiments. We found that the proportion of tokens differed across experiments for all three groups (Controls: $F(2,4) = 89.8$, $P < 0.001$; VS: $F(2,4) = 108.0$, $p < 0.001$; Amygdala: $F(2,4) = 80.2$, $p < 0.001$).

TkD v TkS

To get a better idea of how the proportions of tokens varied by tasks we completed our pairwise task comparisons. We started by comparing TkD to TkS, these two experiments have the same max cashout, except in TkS choices are only reinforced 75% of the time. Thus in TkD, the max tokens monkeys could get every 12 trials was 14, which came out to 1.17 tokens per trial. While for TkS, the max tokens monkeys could get every 12 trials was 10.5 (14 * 0.75), which came out to 0.875 tokens per trial. On average a cash-out occurs every 5 trials, so when we multiplied the per trial rate for each experiment by 5, we got a max average cash-out of 5.83 for TkD, and 4.38 for TkS. When we compared TkD to TkS, we found that that monkeys earned a larger proportions of their max possible cash-out in the TkS experiment (Experiment; $F(1,8) = 165.9$, $p < 0.001$). There were no group effects. When we looked at each group separately across these two experiments, we found that control and amygdala animals earned a larger proportion of their max possible cash-out in the TkS experiment (Controls: Experiment; $F(1,3) = 135.4$, $p = 0.001$; Amygdala: Experiment; $F(1,3) = 57.8$, $p = 0.005$). There was a trend for VS animals to earn a larger proportion of their max

possible cash-out in TkS, this effect did not survive correction (VS: Experiment; $F(1,2) = 19.4$, $p = 0.048$). This follows because in this task comparison, monkeys performed overall better in TkS, but the groups performed better to varying degrees. Control and amygdala animals performed significantly better in TkS when compared to TkD, while VS animals only performed slightly better.

TkS v TkL

Next, we compared TkS to TkL. In the TkL experiment monkeys received an endowment of 4 tokens on the first trial after a cash-out. This resulted in a max of 26 tokens every 12 trials. It should be noted that this is the max number of tokens monkeys could earn in TkL every 12 trials, this value assumes monkeys received three endowments of 4 tokens, totaling 12 tokens from endowments in this 12 trial period. Thus, this max value assumes tokens were cashed-out every 4 trials, which is the minimum number of trials before a cash-out occurred. Cash-outs don't always occur on the fourth trial, but none the less, 26 tokens is the max number of tokens monkeys could earn for every 12 trials, which comes out to a avg per trial rate of 2.16. When we compared TkS to TkL, we found that monkeys earned a larger proportion of their max possible tokens in TkS (Experiment; $F(1,6) = 854.8$, $p < 0.001$). We found no group effects in this comparison. When we looked at each group individually, we found that all groups earned a higher proportion of the max possible tokens in TkS (Controls: Experiment; $F(1,2) =$

235.2, p = 0.004; VS: Experiment; F(1,2) = 250.9, p = 0.003; Amygdala: Experiment; F(1,2) = 476.8, p = 0.002).

TkD v TkL

In the final task comparison TkD v TkL, we found that monkeys earned a higher proportion of tokens in the TkD task (Experiment; F(1,6) = 141.2, p < 0.001). There were no group differences in this task comparison. When we looked at each group individually, we found that only VS animals earned a higher proportion of token in TkD (VS: Experiment; F(1,2) = 442.7, p = 0.002). There was a trend for control and amygdala animals to earn a higher proportion of tokens in TkD, but neither group survived correction (Controls: Experiment; F(1,2) = 17.4, p = 0.053; Amygdala: Experiment; F(1,2) = 31.1, p = 0.031). These effects also follow the trend of the behavioral data, VS animals discount more in the TkL experiment when compared to TkD, in particular across the three conditions without the value difference, when compared to the other two groups.

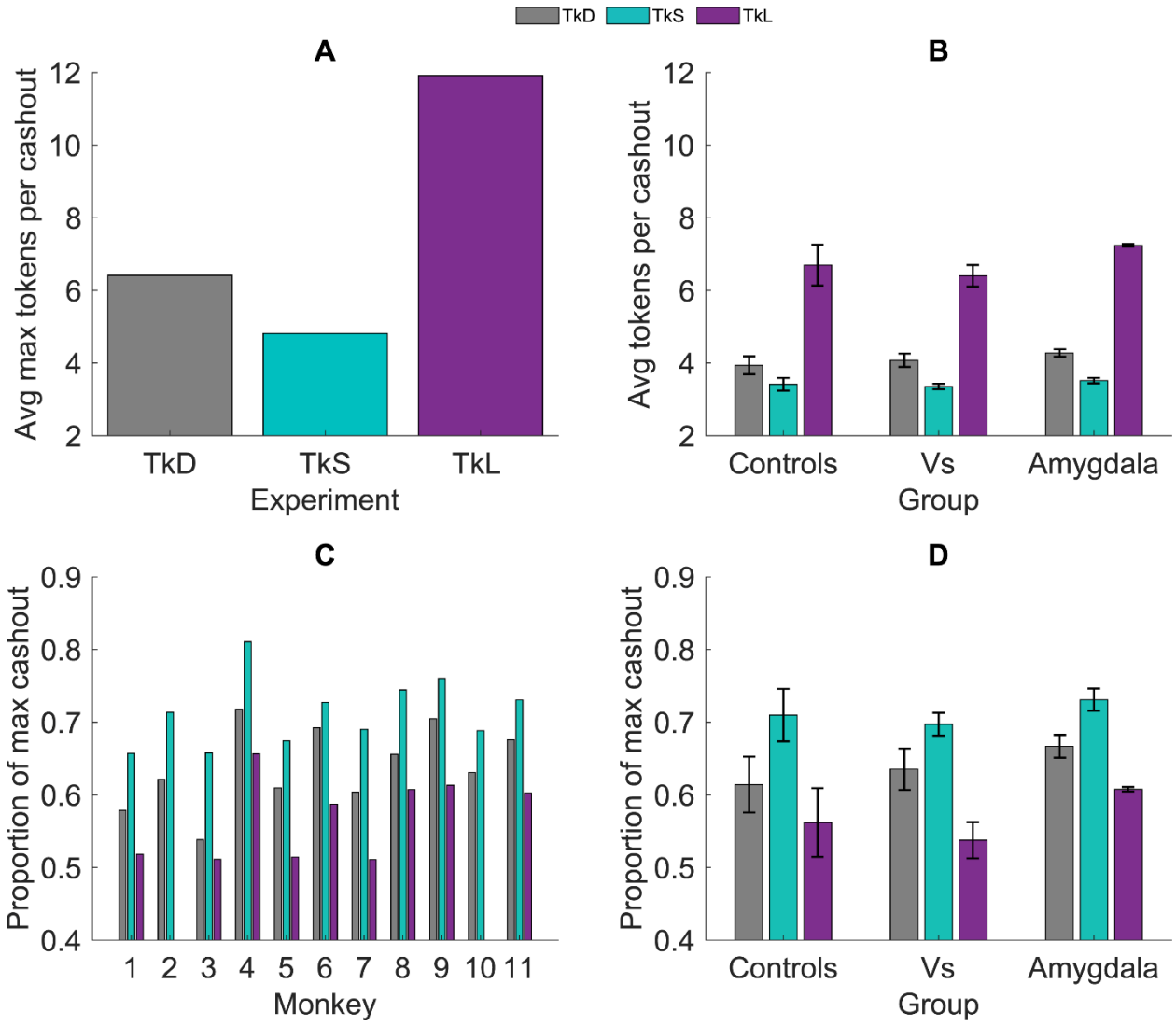


Figure 6. Tokens at Cashout across the 3 tasks. A. Max number of tokens monkeys could have on average per cash-out across tasks. B. Average number of tokens monkeys had per cashout across tasks. C. Average proportion of tokens monkeys had on cash-outs for each monkey across tasks (Controls = monkeys 1-4, VS = monkeys 5-7 & Amygdala = monkeys 8-11). D. Average proportion of tokens monkeys had on cash-outs averaged within group. C & D were derived by dividing the number of tokens monkeys had by the max tokens monkeys could have (B/A). Error bars in B & D are +/- s.e.m. with N = number of animals.

5.3.3 Reinforcement learning model

We fit a large set of RL models (see supplemental material) to the novel choice data presented in chapter 3 and chapter 4. All models were built around the Rescorla-Wagner or stateless RL equation, in which the models varied in the

number of free parameters used to model the choice behavior. Following this we used the Bayesian Information Criterion (BIC) and Akaike information criterion (AIC) to assess which model fit best in each session, for each animal and experiment. There was some variation in model selection across groups and tasks, but overall, model 11 was selected the most by both the AIC (Fig. 7A – 7C) and BIC (Fig 7D – 7F). Model 11 consisted of one learning rate for each cue, one inverse temperature parameter across all conditions, one choice autocorrelation parameter, and one decay parameter.

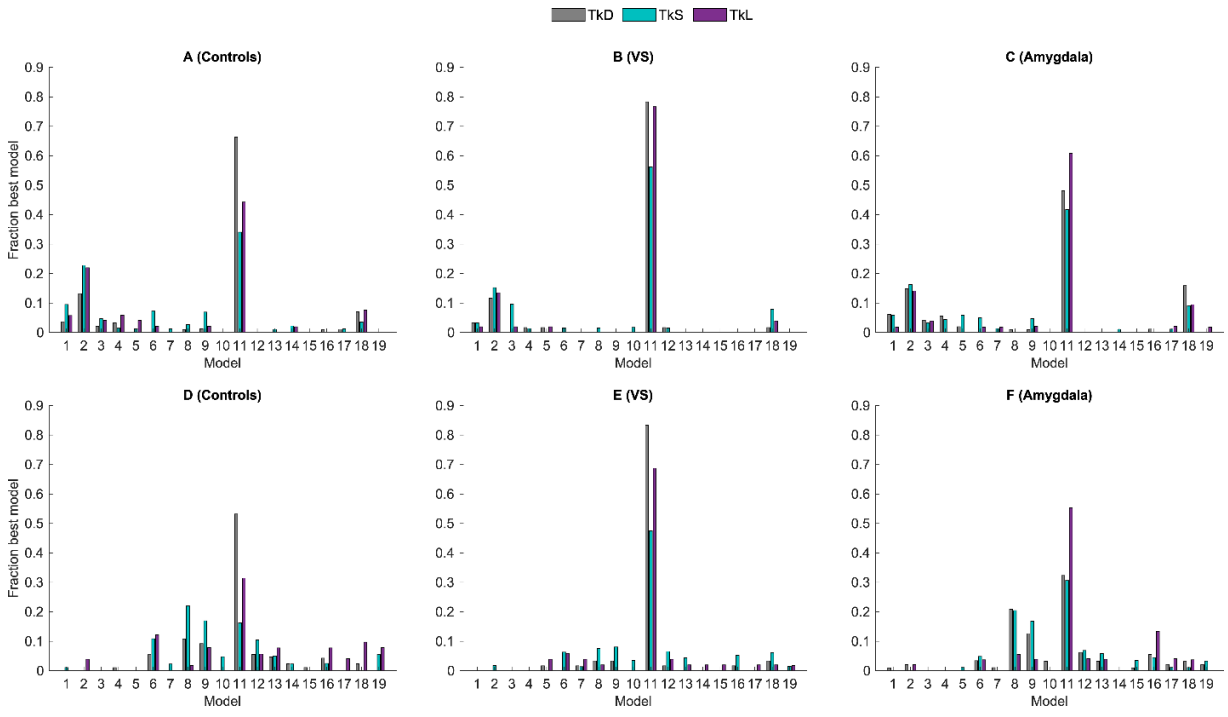


Figure 7. **RL model selection across the three tasks.** Top row (A-C) is AIC model selection for each group across the three experiments. Bottom row is the same as the top row but for BIC model selection.

We start each comparison section by presenting a table summarizing the anova results for model 11's cue learning rate parameters across experiments for both our across and within group analysis. Our across group analysis started with an anova over cue learning rate parameters for each monkey across the three groups. This anova had main effects of 'Group', 'Monkey', 'Cue', and 'Experiment'. A full model was run, so this initial model had two and three way interactions of all the factors listed above. The interactions of interest are 'Group x Cue', 'Group x Experiment', 'Cue x Experiment', and 'Group x Cue x Experiment'. The three way interaction was never significant, so we stick to describing the main effects and the two way interactions. Next, we separated the learning rates for cues by valance (gains & losses) and ran an anova with the same effects listed above. Following this, if we found that learning rates from cues of the same valance differed from one another or across experiments, we ran an anova on each cue separately for that particular valance. In the table and in the analysis all effects are presented in their uncorrected form, but we only bold (in table), and consider an effect significant in the written analysis, if it survived correction. Finally, we performed within group analysis in the same way described above, except the 'Group' factors are dropped.

TkD v TkS v TkL

Monkeys	Cues Included	Group	Cue	Experiment	Group x Cue	Group x Experiment	Cue x Experiment
All	+2,+1,-1,-2/-4	**p = 0.037**	**p < 0.001**	**p = 0.014**	**p = 0.045**	p = 0.435	**p = 0.001**
All	+2, +1	**p = 0.005**	**p < 0.001**	**p < 0.001**	p = 0.963	p = 0.413	p = 0.224
All	+2	**p = 0.019**	n/a	**p = 0.003**	n/a	p = 0.649	n/a
All	+1	**p = 0.013**	n/a	**p = 0.001**	n/a	p = 0.435	n/a
All	-1,-2/-4	p = 0.689	p = 0.465	p = 0.713	p = 0.731	p = 0.622	p = 0.734
Controls	+2,+1,-1,-2/-4	n/a	**p = 0.001**	p = 0.418	n/a	n/a	**p = 0.039**
Controls	+2, +1	n/a	p = 0.087	p = 0.208	n/a	n/a	p = 0.148
Controls	-1,-2/-4	n/a	p = 0.928	p = 0.482	n/a	n/a	p = 0.457
VS	+2,+1,-1,-2/-4	n/a	**p = 0.004**	p = 0.358	n/a	n/a	p = 0.064
VS	+2, +1	n/a	**p = 0.005**	**p = 0.024**	n/a	n/a	p = 0.073
VS	+2	n/a	n/a	p = 0.018	n/a	n/a	n/a
VS	+1	n/a	n/a	p = 0.032	n/a	n/a	n/a
VS	-1,-2/-4	n/a	p = 0.543	p = 0.603	n/a	n/a	p = 0.858
Amygdala	+2,+1,-1,-2/-4	n/a	**p = 0.006**	p = 0.111	n/a	n/a	p = 0.577
Amygdala	+2, +1	n/a	p = 0.057	p = 0.028	n/a	n/a	p = 0.601
Amygdala	-1,-2/-4	n/a	p = 0.698	p = 0.594	n/a	n/a	p = 0.728

Table 1. Across and within group ANOVA results for cue learning rate parameters across TkD, TkS, and TkL.

When we compared learning rate parameters for model 11 across all three experiments, we found that learning rate parameters varied across cues (Cue; $F(3,24) = 83.1$, $p < 0.001$), experiments (Experiment; $F(2,14) = 5.8$, $p = 0.0150$), and learning cues by experiments (Experiment x Cue; $F(6,43) = 4.7$, $p = 0.001$).

We found that learning rates for cues differed by group (Group; $F(2,8) = 5.2$, $p = 0.037$). We also found that this group effect was driven by particular cues (Group x Cue; $F(6,24) = 2.6$, $p = 0.045$). No other interactions were significant. Next, we examined learning rates for the gain and loss cues separately. For the gain cues, we found differences across cues (Cue; $F(1,8) = 45.1$, $p < 0.001$) and experiments (Experiment; $F(2,14) = 17.6$, $p < 0.001$). We also found that the groups differed across gain cues (Group; $F(2,8) = 11.1$, $p = 0.005$). No other interactions were significant. Since we found differences in the gain cues, we looked at the learning rate for each gain cue separately across experiments. For both the +2 and +1 cue, we found that their learning rates differed by experiment (+2: Experiment; $F(2,14) = 8.9$, $p = 0.003$; +1: Experiment; $F(2,14) = 11.1$, $p = 0.001$). We also found a trend for the learning rates to differ by group (+2: Group; $F(2,8) = 6.7$, $p = 0.02$; +1: Group; $F(2,8) = 7.7$, $p = 0.013$), but individually they did not survive correction. When we examined just the loss cues, we found no differences across cues (Cue; $F(1,8) = 0.6$, $p = 0.465$), or experiments (Experiment; $F(2,14) = 0.35$, $p = 0.713$). There were no group differences for the loss cues (Group; $F(2,8) = 0.4$. $p = 0.689$), and no other interactions.

Controls

For the control animals we found that learning rate parameters varied across cues (Cue; $F(3,6) = 21.8$, $p = 0.001$). There was no main effect of experiment (Experiment; $F(2,4) = 1.1$, $p = 0.418$), but learning rates for cues did differ by experiment (Experiment x Cue; $F(6,12) = 3.2$, $p = 0.039$). When we looked at learning rates for gain and loss cues separately, we found that learning rates for gain cues did not differ by cue (Cue; $F(1,4) = 9.9$, $p = 0.087$), experiment (Experiment; $F(2,4) = 2.4$, $p = 0.208$), or learning rates by experiment (Experiment x Learning rate; $F(2,4) = 3.2$, $p = 0.148$). Learning rates for loss cues did not differ by cue (Cue; $F(1,4) = 0.01$, $p = 0.928$), experiment (Experiment; $F(2,4) = 0.9$, $p = 0.482$), or learning rate by experiment (Experiment x Cue; $F(2,4) = 0.9$, $p = 0.457$).

VS

For the VS monkeys, we found that learning rate parameters differed across cues (Cue; $F(3,6) = 13.7$, $p = 0.004$). There was no main effect of experiment (Experiment; $F(2,4) = 1.3$, $p = 0.358$), or cue by experiments interaction (Experiments x Cue; $F(6,12) = 2.7$, $p = 0.064$). When we looked at the learning rate parameters for gain and loss cues separately, we found that the learning rates for the gain cues did differ by experiment (Experiment; $F(2,4) = 10.8$, $p = 0.024$), cue (Cue; $F(1,2) = 204.9$, $p = 0.005$), but the interaction of learning rates and experiment was not significant (Experiment x Learning rate; $F(2,4) = 5.4$, $p = 0.073$). To better characterize the individual effects of both gain cues, we analyzed

learning rates for these cues separately. We found that learning rates for both the +2 (Experiment; $F(2,4) = 12.8$, $p < 0.018$) and +1 cue (Experiment; $F(2,4) = 9.1$, $p = 0.032$) trended toward being different across experiments (but they do not survive correction). The learning rates for the loss cues did not differ by experiment (Experiment; $F(2,4) = 0.6$, $p = 0.603$), cue (Cue; $F(1,2) = 0.5$, $p = 0.543$), or learning rates by experiment (Experiment x Cue; $F(2,4) = 0.16$, $p = 0.858$).

Amygdala

For amygdala animals, we found that learning rates differed by cue (Cue; $F(3,6) = 29.2$, $p < 0.001$). There was no main effect of experiment (Experiment; $F(2,4) = 4.0$, $p = 0.111$), and no interaction of experiment and learning rate (Experiment x Cue; $F(6,12) = 0.8$, $p = 0.577$). When we looked at learning rates for gain and loss cues separately, it became clear that these effects were primary driven by the gain cues. There was a trend for learning rates to differ across gain cues (Cue; $F(1,2) = 16.0$, $p = 0.057$), and experiments (Experiment; $F(2,4) = 9.9$, p 0.028). Again, there was no interaction of learning rate by experiment (Experiment x Cue; $F(2,4) = 0.6$, $p = 0.601$). Learning rates for loss cues did not differ across experiments (Experiment; $F(2,4) = 0.6$, $p = 0.594$). Learning rates for the loss cues did not differ from one another (Cue; $F(1,2) = 0.2$, $p = 0.698$), or across experiments (Experiment x Cue; $F(2,4) = 0.3$, $p = 0.728$).

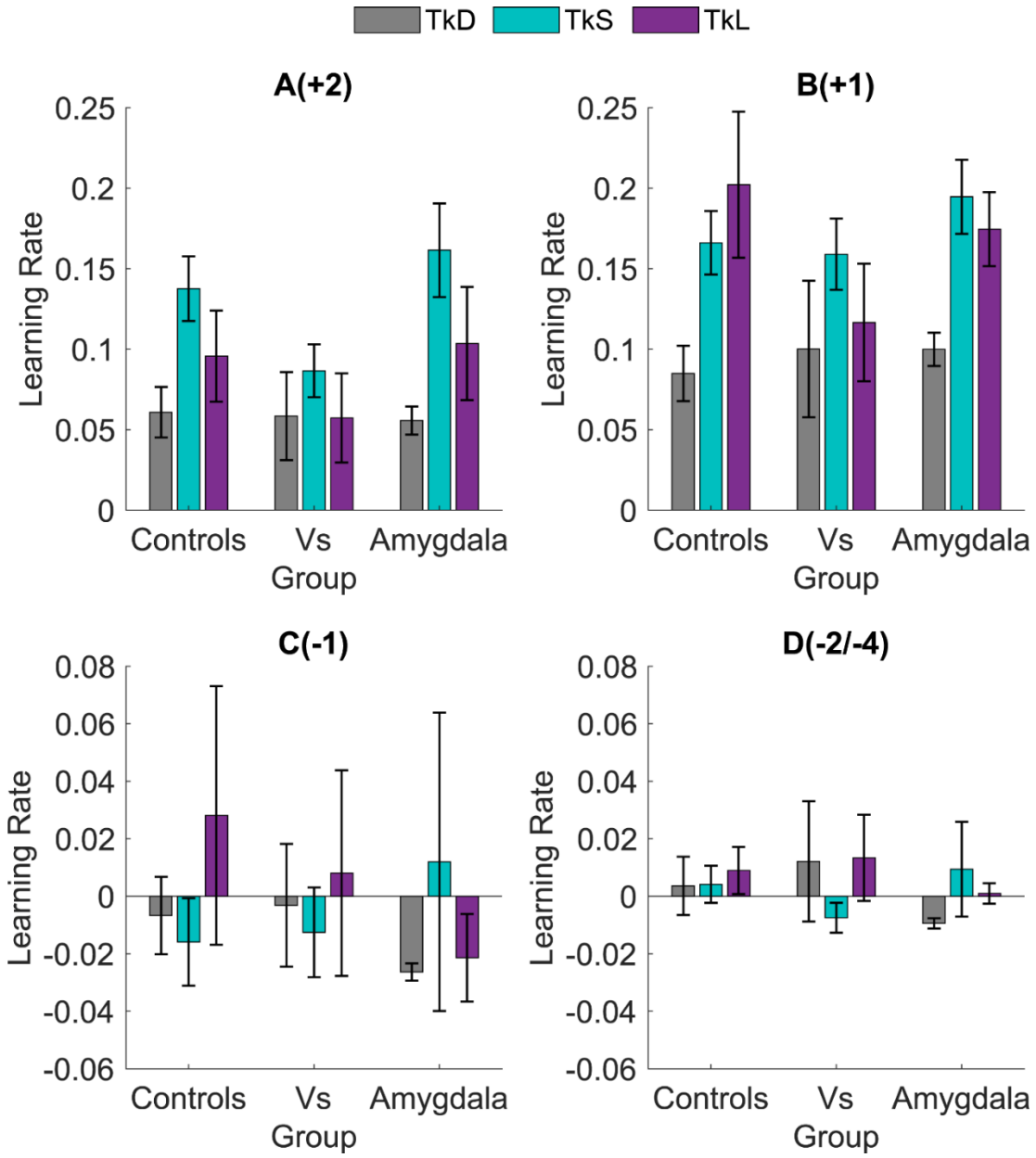


Figure 8. Learning rate parameters for each cue across the three tasks**.** A. Learning rate parameters for the +2 cue across the three tasks. B. Learning rate parameters for the +1 cue across the three tasks. C. Learning rate parameters for the -1 cue across the three tasks. D. Learning rate parameters for the -2/-4 cue across the three tasks.

TkD v TkS

Monkeys	Cues Included	Group	Cue	Experiment	Group x Cue	Group x Experiment	Cue x Experiment
All	+2,+1,-1,-2/-4	p = 0.227	**p < 0.001**	**p = 0.006**	p = 0.249	p = 0.297	**p < 0.001**
All	+2, +1	**p = 0.018**	**p = 0.007**	**p < 0.001**	p = 0.701	p = 0.169	p = 0.644
All	+2	**p = 0.046**	n/a	**p = 0.001**	n/a	p = 0.266	n/a
All	+1	p = 0.125	n/a	**p < 0.001**	n/a	p = 0.863	n/a
All	-1,-2/-4	p = 0.998	p = 0.234	p = 0.923	p = 0.829	p = 0.497	p = 0.619
Controls	+2,+1,-1,-2/-4	n/a	**p < 0.001**	p = 0.061	n/a	n/a	**p = 0.006**
Controls	+2, +1	n/a	p = 0.122	**p = 0.004**	n/a	n/a	p = 0.904
Controls	+2,	n/a	n/a	**p = 0.007**	n/a	n/a	n/a
Controls	+1	n/a	n/a	p = 0.047	n/a	n/a	n/a
Controls	-1,-2/-4	n/a	p = 0.031	p = 0.821	n/a	n/a	p = 0.570
VS	+2,+1,-1,-2/-4	n/a	**p < 0.001**	p = 0.34	n/a	n/a	p = 0.062
VS	+2, +1	n/a	**p = 0.010**	p = 0.071	n/a	n/a	p = 0.105
VS	+2	n/a	n/a	p = 0.078	n/a	n/a	n/a
VS	+1	n/a	n/a	p = 0.074	n/a	n/a	n/a
VS	-1,-2/-4	n/a	p = 0.499	p = 0.582	n/a	n/a	p = 0.553
Amygdala	+2,+1,-1,-2/-4	n/a	**p < 0.001**	p = 0.073	n/a	n/a	p = 0.156
Amygdala	+2, +1	n/a	p = 0.217	**p = 0.007**	n/a	n/a	p = 0.848
Amygdala	+2,	n/a	n/a	p = 0.069	n/a	n/a	n/a
Amygdala	+1	n/a	n/a	p = 0.017	n/a	n/a	n/a
Amygdala	-1,-2/-4	n/a	p = 0.731	p = 0.459	n/a	n/a	p = 0.607

Table 2. Across and within group ANOVA results for cue learning rate parameters across TkD and TkS.

When we compared the learning rate parameters for TkD and TkS, we found that learning rates differed by cue (Cue; $F(3,24) = 78.4$, $p < 0.001$), experiment (Experiment; $F(1,8) = 13.5$, $p = 0.006$), and learning rates by experiment

(Experiment x Cue; $F(3,24) = 10.5$, $p < 0.001$). Across all the learning rates, we found no group effects. This supports the behavioral results above showing that all groups performed better in the TkS experiment, thus overall learning rates were higher for all groups in TkS as compared to TkD. When we separated the learning rates for gain and loss cues, we found that the learning rates for gain cues differed by cue (Cue; $F(1,8) = 12.9$, $p = 0.007$) and the learning rates for gain cues were higher in TkS (Experiment; $F(1,8) = 96.1$, $p < 0.001$). There was no interaction of learning rates by experiment (Experiment x Cue; $F(1,8) = 0.23$). We did, however, find a main effect of group across learning rates for gain cues (Group; $F(2,8) = 6.8$, $p = 0.018$). We did not find that learning rate parameters for gain cues differed by group across experiments (Group x Experiment; $F(2,8) = 2.2$, $p = 0.170$), which means all groups trended toward the same direction (higher learning rates in TkS). Since the learning rates for gain cues differed across cues and experiments, we looked at learning rates for each gain cue separately. When we separated the learning rates for gain cues, it became clear that the learning rate for the +2 cue was driving this group effect. For the +2 cue, we found that learning rates were higher in the TkS (Experiment; $F(1,8) = 22.6$, $p = 0.001$). In addition to this effect, there was a trend for control and amygdala animals to have higher learning rates for the +2 cue (Group; $F(2,8) = 0.046$) when compared to VS animals (although this group effect does not survive correction). When we looked at the learning

rates for the +1 cue, we found that learning rates for the +1 cue were higher in TkS (Experiment; $F(1,8) = 40.4$, $p < 0.001$). The groups did not differ when it came to this learning rate (Group; $F(2,8) = 2.7$, $p = 0.125$). For loss cues, we found that learning rates did not differ by cue (Cue; $F(1,8) = 1.6$, $p = 0.234$) or by experiment (Experiment; $F(1,8) = 0.01$, $p = 0.923$).

Controls

Next, we completed our within group analysis of the learning rates in TkD compared to learning rates in TkS. For control animals, we found that learning rates differed by cue (Cue; $F(3,9) = 92.9$, $p < 0.001$). We did not find a main effect of experiment (Experiment; $F(1,3) = 8.5$, $p = 0.061$), but we did find that learning rates differed across experiments (Experiment x Cue; $F(3,9) = 8.3$, $p = 0.006$). When we separated the learning rates for gain and loss cues, we found that learning rates for gain cues were higher in TkS (Experiment; $F(1,3) = 63.0$, $p = 0.004$). The learning rates for gain cues did not differ from one another (Cue; $F(1,3) = 4.5$, $p = 0.122$), or by experiment (Experiment x Cue; $F(1,3) = 0.02$, $p = 0.904$). When we separated the learning rates for gain cues, we found that learning rates for the +2 cue was driving the higher learning rates for gain cues in TkS (Experiment; $F(1,3) = 42.9$, $p = 0.007$). Learning rates for the +1 cue trended toward being higher in TkS (Experiment; $F(1,3) = 10.5$, $p = 0.047$), but did not survive correction. We found no difference across tasks in learning rates for loss cues (Experiment; $F(1,3)$

= 0.1, p = 0.821), and no difference in learning rates between loss cues (Cue; F(1,3) = 14.5, p = 0.031).

VS

For VS animals we found that learning rates for cues differed from one another (Cue; F(3,6) = 40.3, p < 0.001). We found no main effect of experiment (Experiment; F(1,2) = 1.5, p = 0.340), or learning rates across experiments (Experiment x Cue; F(3,6) = 4.3, p = 0.062). When we looked at just the learning rates for gain cues, we found a trend for learning rates for gain cues to be higher in TkS (Experiment; F(1,2) = 12.6, p = 0.071), but did not reach significance. Learning rates for gain cues, also, did not differ across experiments (Experiment x Cue; F(1,2) = 8.0, p = 0.105). We did find that learning rates for gain cues differed from one another (Cue; F(1,2) = 97.0, p = 0.010). When we separated learning rates for gain cues, we found that VS monkeys trended toward having higher learning rates for both the +2 (Experiment; F(1,2) = 11.3, p = 0.078), and +1 (Experiment; F(1,2) = 12.0, p = 0.074) cue in TkS, but neither reached significance. The learning rates for loss cues did not differ from one another (Cue; F(1,2) = 0.67, p = 0.499), by experiment (Experiment; F(1,2) = 0.4, p = 0.582), or learning rates across experiments (Experiment x Cue: F(1,2) = 0.5, p = 0.553).

Amygdala

For amygdala animals, we found that learning rates for cues differed from one another (Cue; $F(3,9) = 20.3$, $p < 0.001$). We did not find that learning rates differed by (Experiment; $F(1,3) = 7.3$, $p = 0.073$), or across (Experiment x Cue; $F(3,9) = 2.2$, $p = 0.156$) experiments. When we separated learning rates for gain and loss cues, we found that learning rates for gain cues were higher in TkS (Experiment; $F(1,3) = 43.9$, $p = 0.007$). Learning rates for gain cues did not differ from one another (Cue; $F(1,3) = 2.3$, $p = 0.217$) or across experiments (Experiment x Cue; $F(1,3) = 0.04$, $p = 0.848$). When we separated learning rates for the gain cues, we found that learning rates for both the +2 and +1 cues trended toward being higher in TkS (+2: Experiment; $F(1,3) = 7.7$, $p = 0.069$; +1: Experiment; $F(1,3) = 22.5$, $p = 0.017$), but neither was significant alone. Learning rates for loss cues did not differ by cue (Cue; $F(1,3) = 0.1$, $p = 0.731$), or by experiment (Experiment; $F(1,3) = 0.7$, $p = 0.459$).

Using RL modeling, we were able to better show how the groups differed between their performance in TkD versus their performance in TkS. While we did not find a group effect across conditions, when we analyzed the choice behavior, this was the case because all groups performed better in the TkS experiment, and the condition where the control an amygdala animals performed better than the VS animals was spread across a couple conditions. The learning rates for our RL model show that control and amygdala animals valued the gain cues more than the

VS animals, this was shown by bigger increases in the learning rates for gain cues by the control and amygdala animals when compared to VS animals. This difference in value for the gain cues in the TkS experiment was mostly driven by the +2 cue.

TkS v TkL

Monkeys	Cues Included	Group	Cue	Experiment	Group x Cue	Group x Experiment	Cue x Experiment
All	+2,+1,-1,-2/-4	p = 0.078	**p < 0.001**	p = 0.556	p = 0.148	p = 0.381	p = 0.154
All	+2, +1	**p = 0.015**	**p = 0.001**	p = 0.107	p = 0.934	p = 0.516	p = 0.212
All	+2	p = 0.039	n/a	p = 0.072	n/a	p = 0.770	n/a
All	+1	p = 0.058	n/a	p = 0.75	n/a	p = 0.286	n/a
All	-1,-2/-4	p = 0.717	p = 0.914	p = 0.587	p = 0.741	p = 0.448	p = 0.739
Controls	+2,+1,-1,-2/-4	n/a	**p = 0.004**	p = 0.633	n/a	n/a	p = 0.131
Controls	+2, +1	n/a	p = 0.157	p = 0.954	n/a	n/a	p = 0.046
Controls	-1,-2/-4	n/a	p = 0.977	p = 0.308	n/a	n/a	p = 0.359
VS	+2,+1,-1,-2/-4	n/a	**p < 0.001**	p = 0.512	n/a	n/a	p = 0.211
VS	+2, +1	n/a	**p = 0.011**	p = 0.075	n/a	n/a	p = 0.361
VS	+2	n/a	n/a	p = 0.042	n/a	n/a	n/a
VS	+1	n/a	n/a	p = 0.112	n/a	n/a	n/a
VS	-1,-2/-4	n/a	p = 0.931	p = 0.503	n/a	n/a	p = 0.995
Amygdala	+2,+1,-1,-2/-4	n/a	**p = 0.001**	p = 0.313	n/a	n/a	p = 0.764
Amygdala	+2, +1	n/a	p = 0.109	p = 0.145	n/a	n/a	p = 0.535
Amygdala	-1,-2/-4	n/a	p = 0.895	p = 0.613	n/a	n/a	p = 0.610

Table 3. Across and within group ANOVA results for cue learning rate parameters for TkS and TkL.

When we compared the learning rates from TkS to TkL, we found that learning rates differed by cue (Cue; $F(3,23) = 56.6$, $p < 0.001$), but not experiment (Experiment; $F(1,6) = 0.4$, $p = 0.556$). There were no group effects or other interactions. When we separated learning rates for gain and loss cues, we found that learning rates for gain cues differed by cue (Cue; $F(1,8) = 23.2$, $p = 0.001$) but not experiment (Experiment; $F(1,6) = 3.6$, $p = 0.107$). We did, however, find a main effect of group (Group; $F(2,8) = 7.5$, $p = 0.015$). To further investigate this group effect, we separated the learning rates for the gain cues, we found that neither cue alone had learning rates that differed across experiment (+2: Experiment; $F(1,6) = 4.8$, $p = 0.072$; +1: Experiment; $F(1,6) = 0.11$, $p = 0.75$). There was, however, a trend for a group difference for both gain cues (+2: Group; $F(2,7) = 5.1$, $p = 0.04$; +1: Group; $F(2,7) = 4.2$, $p = 0.058$), but neither was significant alone. Learning rates for loss cues did not differ by cue (Cue; $F(1,8) = 0.01$, $p = 0.914$) or experiment (Experiment; $F(1,6) = 0.33$, $p = 0.587$).

Controls

In our within group analysis we found that learning rates for control animals differed by cue (Cue; $F(3,6) = 14.0$, $p = 0.004$). Learning rates did not differ by (Experiment; $F(1,2) = 0.3$, $p = 0.633$), or across experiments (Experiment x Cue; $F(3,6) = 2.8$, $p = 0.131$). When we separated learning rates for the gain and loss cues, we found that learning rates for gain cues did not differed by cue (Cue; $F(1,2)$

= 4.9, p = 0.157), or experiment (Experiment; F(1,2) = 0, p = 0.954). There was a trend for learning rates for gain cues to differ across experiments (Experiment x Cue; F(1,2) = 13.5 p = 0.046), but this effect did not survive correction. Learning rates for loss cues did not differ by cue (Cue; F(1,2) = 0, p = 0.977), or experiment (Experiment; F(1,2) = 1.8, p = 0.308).

VS

For VS animals, we found that learning rates across cues differed (Cue; F(3,6) = 26.8, p < 0.001), but not by experiment (Experiment; F(1,2) = 0.6, p = 0.512). When we separated learning rates for gain and loss cues, we found that learning rates for gain cues differed by cue (Cue; F(1,2) = 88.9, p = 0.011). Learning rates for gain cues did not differ by (Experiment; F(1,2) 11.7 = 0.075), or across experiments (Experiment x Cue; F(1,2) = 1.4, p = 0.361). The interaction of learning rate by experiment was not significant (Experiment x Cue; F(1,228) = 0.8, p = 0.382). Since there was a trend for the learning rates for gain cues to differ by experiment, we separated learning rates for the gain cues, and found that this trend was driven by the +2 (Experiment; F(1,2) = 22.0, p = 0.042). Learning rates for loss cues did not differ by cue (Cue; F(1,2) = 0.01, p = 0.931), or by experiment (Experiment; F(1,2) = 0.7, p = 0.503).

Amygdala

For the amygdala animals, we found that learning rates differed by cue (Cue; $F(3,6) = 21.3$, $p = 0.001$). Learning rates did not differ by (Experiment; $F(1,2) = 1.7$, $p = 0.313$), or across (Experiment x Cue; $F(3,6) = 2.1$, $p = 0.764$) experiments. When we separated the learning rates for gain and loss cues, we found that neither differed by cue (Gains: Cue; $F(1,2) = 7.6$, $p = 0.109$; Losses: Cue; $F(1,2) = 0.02$, $p = 0.895$), or by experiment (Gains: Experiment; $F(1,2) = 5.4$, $p = 0.145$; Losses: Experiment; $F(1,2) = 0.3$, $p = 0.613$). There was no interaction of learning rates by experiment for gain or loss cues.

This analysis revealed that despite the value difference in TkL, monkeys did not value the cues more in this experiment. The difference between the groups lies in how the groups valued the gain cues in both experiments. Control and amygdala animals valued the gain cues more in these two experiments, when compared to VS animals.

TkD v TkL

Monkeys	Cues Included	Group	Cue	Experiment	Group x Cue	Group x Experiment	Cue x Experiment
All	+2,+1,-1,-2/-4	p = 0.064	**p < 0.001**	p = 0.097	p = 0.075	p = 0.607	p < 0.089
All	+2, +1	**p = 0.017**	**p < 0.001**	p = 0.069	p = 0.654	p = 0.498	p = 0.123
All	+2	p = 0.145	n/a	p = 0.206	n/a	p = 0.594	n/a
All	+1	**p = 0.012**	n/a	p = 0.058	n/a	p = 0.495	n/a
All	-1,-2/-4	p = 0.453	p = 0.500	p = 0.351	p = 0.425	p = 0.885	p = 0.473
Controls	+2,+1,-1,-2/-4	n/a	**p = 0.005**	p = 0.395	n/a	n/a	p = 0.279
Controls	+2, +1	n/a	**p = 0.025**	p = 0.314	n/a	n/a	p = 0.224
Controls	+2,	n/a	n/a	p = 0.443	n/a	n/a	n/a
Controls	+1	n/a	n/a	p = 0.275	n/a	n/a	n/a
Controls	-1,-2/-4	n/a	p = 0.684	p = 0.618	n/a	n/a	p = 0.563
VS	+2,+1,-1,-2/-4	n/a	p = 0.188	p = 0.145	n/a	n/a	p = 0.551
Amygdala	+2,+1,-1,-2/-4	n/a	**p < 0.001**	p = 0.086	n/a	n/a	p = 0.537
Amygdala	+2, +1	n/a	p = 0.127	p = 0.203	n/a	n/a	p = 0.677
Amygdala	-1,-2/-4	n/a	p = 0.145	p = 0.392	n/a	n/a	p = 0.781

Table 4. Across and within group ANOVA results for cue learning rate parameters across TkD and TkL.

When we compared learning rates from TkD to TkL, we found that learning rates differed by cue (Cue; $F(3,23) = 36.5$, $p < 0.001$), but not experiment (Experiment; $F(1,6) = 3.8$, $p = 0.097$). There was a trend for the control and amygdala animals to have higher learning rates across the experiments as compared to VS animals (Group; $F(2,5) = 4.8$, $p = 0.064$), and for the learning rates to differ by group across experiments (Group x Cue; $F(6,23) = 2.2$, $p =$

0.075). When we separated learning rates for gain an loss cues, for gain cues, we found that learning rates differed by cue (Cue; $F(1,8) = 30.4$, $p < 0.001$), group (Group; $F(2,6) = 8.5$, $p = 0.017$). Learning rates did not differ by (Experiment; $F(1,6) = 4.8$, $p = 0.069$), or across experiments (Experiment x Cue; $F(1,6) = 3.2$, $p = 0.123$). When we separated the learning rates for gain cues, we found that the group effect across learning rates for gain cues was driven by the +1 cue (Group; $F(2,5) = 11.7$, $p = 0.012$). The learning rates for the +1 cue did not differ by experiment (Experiment; $F(1,6) = 5.5$, $p = 0.058$), or by, group across experiment (Group x Experiment; $F(2,6) = 0.8$, $p = 0.495$). There were no effects for the +2 cue. For loss learning rates, we found no difference across cues (Cue; $F(1,8) = 0.5$, $p = 0.500$), experiment (Experiment; $F(1,6) = 1.0$, $p = 0.351$), or any other interactions.

Controls

When we looked at learning rates for just the control animals, we found that learning rates differed by cue (Cue; $F(3,6) = 12.8$, $p = 0.005$). Learning rates did not differ by (Experiment; $F(1,2) = 1.2$, $p = 0.395$), or across experiments (Experiment x Cue; $F(3,6) = 0.9$, $p = 0.279$). Learning rates for just gain cues differed by cue (Cue; $F(1,2) = 37.8$, $p = 0.025$). They did not, however, differ by (Experiment; $F(1,2) = 1.8$, $p = 0.314$), or across experiments (Experiment x Cue; $F(1,2) = 3.0$, $p = 0.224$). Learning rates for loss cues did not differ across cues

(Cue; $F(1,2) = 0.2$, $p = 0.684$), or by experiment (Experiment; $F(1,2) = 0.3$, $p = 0.618$).

VS

For VS animals, we found that learning rates did not differ across cues (Cue; $F(3,6) = 2.2$, $p = 0.188$), or by experiment (Experiment; $F(1,2) = 5.4$, $p = 0.145$). Learning rates for gain cues differed by cue (Cue; $F(1,2) = 139.9$, $p = 0.007$). Learning rates for gain cues did not differ by (Experiment; $F(1,2) = 1.2$, $p = 0.385$), or across experiments (Experiment x Cue; $F(1,2) = 9.9$, $p = 0.087$). Learning rates for loss cues did not differ across cues (Cue; $F(1,2) = 0.4$, $p = 0.594$), or by experiment (Experiment; $F(1,2) = 3.1$, $p = 0.221$).

Amygdala

Learning rates for amygdala animals differed across cues (Cue; $F(3,6) = 24.9$, $p < 0.001$). Learning rates did not differ by (Experiment; $F(1,2) = 10.1$, $p = 0.086$), or across experiments (Experiment x Cue; $F(3,6) = 0.8$, $p = 0.537$). Learning rates for gain cues did not differ by cue (Cue; $F(1,2) = 6.4$, $p = 0.127$), or by experiment (Experiment; $F(1,2) = 3.5$, $p = 0.203$). Learning rates for loss cues did not differ by cue (Cue; $F(1,2) = 5.4$, $p = 0.145$), or by experiment (Experiment; $F(1,2) = 1.2$, $p = 0.392$). We did, however, find a trend for the learning rate for the -1 v -2/-4 condition to be higher in TkL (Experiment; $F(1,2) = 18.4$, $p = 0.050$), but this effect did not survive correction.

This analysis revealed that despite value difference between the two experiments, monkeys did not value the cues differently. However, as can been seen in Fig.7, control and amygdala animals value the gain cues more in TkL, when compared to VS animals.

5.4 Discussion

We carried out a post hoc analysis for three of the experiments (TkD, TkS, TkL) presented in chapter 3 and chapter 4. We completed within and between group analysis of performance across the three experiments and found that for all three groups, performance varied by experiment. All three groups modulated their behavior in the same directions (to varying degrees) across the three experiments, and this level of behavior modulation is why we find deficits in some experiments but not others. When compared to control animals, we only found that monkeys with lesions to the VS had deficits in the last two experiments (TkS & TkL). Our within group analysis revealed that these deficits were the result of VS animals not modulating their choice behavior as well as control and amygdala animals.

We were able to fit RL models to this data, and further characterize what was driving these different levels of performance across experiments. We found that VS animals had lower learning rates for gain cues in TkS and TkL, when compared to control and amygdala monkeys. In the context of this analysis, these

cue specific learning rates can be seen as reflections of the value animals hold for said cue in its environmental setting. Thus the deficits we find in TkS and TkL result from VS animals not valuing the gain cues as much as the other two groups did in these two experiments.

Our present analysis of comparing performance across experiments revealed that several things affect performance. First, we show that reward schedules have differential effects on learning. This can be seen in our comparison of TkD to TkS. Overall all animals performed better in TkS, with the only difference between the groups being that control and amygdala animals performed significantly better in TkS when compared to their performance in TkD. This effect is at odds with current theories of RL, if these differences in performance across experiments are in fact due to learning, animals should not be able to perform better in the TkS experiment. It should be harder to learn a stochastic cue-reward association when compared to a deterministic cue-reward association. Instead, what these analyses reveal is a behavioral phenomenon that has long been known in operant conditioning: stochastic schedules of reinforcement maintain higher rates of behavior, when compared to deterministic schedules of reinforcement (28, 29). The behavioral theory behind this effect is that in the real world most reinforcement is stochastic, thus organisms have developed mechanisms that make them more sensitive to stochastic reward schedules. Recent work in dopamine

responses provide support for this theory, (228) found amplified dopamine responses in monkeys receiving rare rewards when compared to monkeys receiving constant rewards.

Our analyses indicate that the experimental effects we find in chapter three showing that VS animals had learning deficits in TkS is not due to VS animals not being able to establish stimulus-outcome relationships, but instead due to VS animals not upping (lacking motivation to perform better) their choice behavior as well as control animals under stochastic schedules. This is an important distinction because without our within group analysis across experiments, it would seem that VS animals just have problems learning cue-reward associations in stochastic environments. Instead our analysis gives us a more in-depth look and helps reveal what role the VS actually plays in RL. Based on this analysis it seems that the VS plays a role in making cues more valuable from a motivational aspect in stochastic learning environments. This result along with the work showing that Pavlovian-instrumental transfer (PIT) is at least partially mediated by the nucleus accumbens (13), provide strong support for the VS playing a motivational role in RL, and is not responsible for the actual learning of the stimulus-outcome relationship.

Furthermore this theory holds for the results found in chapter four, in which monkeys with amygdala lesions had no performance deficits in any of the token experiments. Evidence suggest that this form of PIT is mediated partially by the

amygdala, and partially by the nucleus accumbens. In one study rats were trained to associate a light-noise compound stimulus with water. Following this half of the rats received excitotoxic lesions of the basolateral amygdala. Next both groups received intra-accumbens amphetamine infusions of d-amphetamine and began the test phase. In the test phase two novel levers were available. Neither lever produced water, but one did produce the conditioned reinforcer of the light-noise compound. The authors found that the amphetamine infusions increased responding on the lever that produced the conditioned reinforcer, and no change in responding on the lever that had no consequence for both the sham-controls and lesion animals (53). Thus, the lesion animals responded like the control animals (intra-accumbens amphetamine infusions of d-amphetamine produced amplified responding).

This result should be viewed in contrast to a study by (13). In this study experimenters examined the effects of cytotoxic lesions of the nucleus accumbens in rats across two instrumental conditioning experiments. When experimenters compared rats with lesions of the nucleus accumbens to sham-controls, they found that instrumental responding of lever pressing and chain pulling for food reinforcers was mildly suppressed in the lesion animals. However, this reduction in responding was not due to lesion animals having trouble learning the instrumental contingency, but instead due to a reduction in motivation. In a second

experiment, d-amphetamine was administered into both the sham-controls and lesion rats, the authors found that the normally increased responding found when d-amphetamine is administered was significantly reduced in lesion animals. These results suggest that the nucleus accumbens' role in instrumental conditioning is to provide excitatory motivational effects of appetitively conditioned Pavlovian signals, instead of holding the value that is attached to instrumental outcomes. The fact that infusions of d-amphetamine to the accumbens made rats with amygdala lesions respond like controls, but not rats with nucleus accumbens lesions provide further support for the hypothesis that the nucleus accumbens plays this excitatory motivational role.

5.4.1 Conclusion

Using our method of testing the monkeys on several different learning tasks we were able to replicate the inconsistent results in the RL literature and narrow down what role the VS plays in RL. Based on our findings the VS seems to play a motivational role in appetitive learning conditions. Specifically, our results suggest that the VS is important for Pavlovian motivational value for appetitive cues.

5.5 Supplemental Material

5.5.1 Reinforcement learning models

Model 1(M1)-

M1 had nine total parameters. Three Learning rate parameters, one for positive feedback, one for neutral feedback, and one for negative feedback. In addition to the three learning rate parameters, there were six inverse temperature parameters, one for each condition.

Model 2(M2)-

M2 had ten total parameters. Four Learning rates (one for each cue) & six inverse temperature parameters (one for each condition).

Model 3(M3)-

M3 had twelve total parameters. six learning rates (one for each condition) & six inverse temperature parameters (one for each condition).

Model 4(M4)-

M4 had nine parameters. Three learning rates (one for each type of Trial (gain v gain, loss v loss & gain v loss)) & six inverse temperature parameters (one for each condition).

Model 5(M5)-

M5 had six parameters. Three Learning rates (one for each type of trial) & three inverse temperature parameters (one for each type of trial).

Model 6(M6)-

M6 had four parameters. Three Learning rate parameters, one for positive feedback, one for neutral feedback, and one for negative feedback. In addition to the three learning rate parameters, there were was one inverse temperature parameter.

Model 7(M7)-

M7 had five parameters. Two learning rates, one for positive feedback and one for negative feedback. In addition to the learning rate parameters three were three inverse temperature parameters (one for each type of trial).

Model 8(M8)-

M8 had three parameters. Two learning rates, one for positive feedback and one for negative feedback & one inverse temperature parameter.

Model 9(M9)-

M9 had four Parameters. Two Learning rates, one for positive feedback and one for negative feedback. This model also had two inverse temperature parameters (one for gain/gain trial type & one for everything else)

Model 10(M10)-

M10 had four parameters. One Learning rate & three inverse temperature parameters (one for each trial type).

Model 11(M11)-

M11 had five parameters. One learning rate parameter for each cue & one inverse temperature parameter.

Model 12(M12)-

M12 had two parameters. One learning rate & one inverse temperature parameter.

Model 13(M13)-

M13 had three parameters. One learning rate & two inverse temperature parameters (one for loss/loss trial type & one for everything else).

Model 14(M14)-

M14 had four parameters. Two learning rates, one for positive feedback and one for negative feedback & two inverse temperature parameters (one for loss/loss trial type & one for everything else).

Model 15(M15)-

M15 had three parameters. One learning rate & two inverse temperature parameters (one for gain/gain trial type & one for everything else).

Model 16(M16)-

M16 had three parameters. Two learning rates, one for gain/gain trial type and one for everything else. This model also had one inverse temperature parameter.

Model 17(M17) –

M17 had four parameters. Two learning rates, one for gain/gain trial type and one for everything else. This model also had two inverse temperature parameter, one for gain/gain trial type and one for everything else.

Model 18(M18)-

M18 had seven parameters. Six learning rates, one for each condition. This model also had one inverse temperature parameter.

Model 19(M19)-

M19 had four parameters. Three learning rates, one for each trial type. This model also had one inverse temperature parameter.

Chapter 6: Conclusions

6.1 Discussion

In this thesis, we investigated the role of the VS and amygdala in reinforcement learning (RL). There was little doubt that these two areas played a role in RL. This was evident by sheer amount and the diversity of research implicating these two areas in RL. However, the question was, what role do they play? We hypothesized that the roles of these two areas might be overstated in current RL theories and this could be seen with the amount of inconsistent literature concerning these two areas. In general current theories of RL suggest that the VS is responsible for operant conditioning (10, 11), while the amygdala is responsible for Pavlovian conditioning (4, 21, 229). The problem with this view is that the literature does not support it for the VS (12-14), or amygdala (20, 21).

We submitted that a large part of these inconsistences was due to three possible reasons. One, the difference in learning environments (task design). Learning is a complicated process with many moving parts, many of the experiments that provide support for these two structures and their associated roles use similar learning environments. When these learning environments are

changed, they are usually accompanied by results that are inconsistent with current theories.

Two, definitions and a failure to account for well researched behavioral findings. The foundation for all of the current theories on learning was built by the extensive behavioral work done by behaviorist, yet as the latency increases since these fields clearly separated, important behavioral distinctions have gotten blurred. This is okay with some higher level concepts that were not well worked out, but this is not the case for the well worked out basic concepts. Much like rules in mathematics, these concepts are based on one another, so it is difficult to pick and choose which behavioral laws to acknowledge. As we stated in the introduction, this has led to misinterpretations of results and lead to overdrawn conclusions.

Three, most of the strong RL evidence comes from physiology experiments (mostly studying Pavlovian conditioning). This is important to note because in these complex learning tasks it is possible that one could be mistaking a factor that correlates with value but is not in fact value (for example motivation). This risk is heightened when behavior and the underlying processes that evoke it are not well understood (ie the type of conditioning in control of the behavior being studied).

To account for these reasons, we investigated the role of the VS and amygdala using monkeys with lesions to one of these areas and studied them on a

series of learning tasks that differed in a few environmental parameters. This method controlled for all the reasons listed above, and allowed us to make stronger conclusions about the role of these two areas in RL.

In chapter two we examined the role of the amygdala on object-based versus action-based learning. In chapter three using a series of four tasks, we investigated the role of the VS in learning from gains and losses. In chapter four we used the same series of tasks and examined the role of the amygdala in learning from gains and losses. And finally, in chapter five we conducted a post hoc analysis of the VS and amygdala data from chapters three and four.

Results recap

In chapter two we found that lesions of the amygdala led to deficits in consistently choosing the more frequently rewarded options. We found these deficits in both the What condition, when animals had to learn to choose the best visual object, and in the Where condition, when the animals had to learn to choose the best action. The deficits in choice accuracy were due to amygdala monkeys switching after a non-rewarded outcome, which led to decreased performance due to the stochastic schedules. We did not find deficits in reversal accuracy; thus, monkeys with amygdala lesions were able to reverse their choice-outcome mappings, in both conditions, as well as controls. Inferring reversals in choice-outcome mappings may, therefore, be more dependent on other brain areas,

including cortex. Overall, this suggests that the amygdala is important for consistently choosing a rewarded option.

In chapter three we carried out four tasks in which we examined the role of the VS in learning from gains and losses. We found learning deficits in monkeys with VS lesions in two of the four tasks. These deficits were consistently driven by trials in which animals had to choose between two cues that differed in positive reward magnitude. There were no deficits when animals had to choose between options, one of which was associated with a loss. We also fit RL models to the data, and found that learning rates were lower for gain cues in the VS animals relative to controls. Thus, lesions of the VS, in this series of tasks, specifically affected learning to choose between rewarding options, and had no effect on learning to avoid losses.

In chapter four using the same tasks from chapter three we looked at learning from gains and losses in animals with amygdala lesions. When we examined group differences in learning, we found that monkeys with amygdala lesions had no learning deficits in any of the tasks. In fact, monkeys with lesions to the amygdala performed numerically better than controls in both the novel and familiar blocks in the null token task (NtK). When we analyzed what was driving these group differences, we found that lesion animals performed slightly better in conditions where a loss was paired with either a gain or neutral cue. One

interpretation of these results is that the monkeys with amygdala lesions were more sensitive to negative feedback.

In chapter five we carried out a post hoc analysis for three of the experiments (TkD, TkS, TkL) presented in chapter three and four. We completed within and across group analysis of performance across the three experiments. We found that for all three groups, performance varied by experiment. All three groups modulated their behavior in the same directions (to varying degrees) across the three experiments, and this level of behavior modulation is why we find deficits in some experiments but not others. When compared to control animals, we only found that monkeys with lesions to the VS had deficits in any of the experiments. VS monkeys had deficits in two out of the four token tasks (TkS & TkL). Our within group analysis revealed that these deficits were the result of VS animals not modulating their choice behavior as well as control and amygdala animals in stochastic and deterministic (with a high reward rate) learning environments. We were able to fit RL models to this data, and further characterize what was driving these different levels of performance across experiments. We found that VS animals had lower learning rates for gain cues in TkS and TkL. In the context of this analysis, these cue specific learning rates can be seen as reflections of the value animals hold for said cue in its environmental setting. Thus

the deficits we find in TkS and TkL come from VS animals not valuing the gain cues as much as the other two groups did in these two experiments.

6.1.1 Conclusions

Using our method of testing the monkeys on several different learning tasks we were able to replicate the inconsistent results in the RL literature. This method allows us to make several important conclusions about the role of the VS and amygdala in learning and about RL as a whole. We were able to narrow down the role of the VS and amygdala and make a number of stronger conclusions about their roles in RL.

Neither structure is important for aversive learning

The first conclusion is that neither the VS nor the amygdala seem to play a role in conditioning from losses. This is evident from the results in chapter three and four, where we showed that neither group had learning deficits in conditions with a loss. We also show that adding aversive consequences to learning environments changes how animals behave.

Distinct roles

The second conclusion is that both of these areas appear to only be important when it comes to conditioning from appetitive outcomes in particular conditioning environments. Importantly these contributions from the VS and amygdala appear

to be distinct. To show that the contribution of these two areas are distinct, we have to consider the results from chapter two and the results from (11). This study ran the same VS animals used in this thesis on the What/Where task that we ran monkeys with amygdala lesions on in chapter two. In the (11) study it was found that monkeys with VS lesions only had deficits in the What block type, they behaved similarly to controls in the Where block type. This is in contrasts with the results in chapter two, where we found that monkeys with amygdala lesions had performance deficits in both the What and Where block types (deficits were larger in the What block type).

`This distinction is further supported by the fact that in the token experiments monkeys with amygdala lesions had no deficits in any of the tasks. While, monkeys with VS lesions had deficits in TkS and TkL. These deficits were due to VS monkeys not performing as well as control or amygdala monkeys in the gain/gain conditions, which was evident by the analysis in chapter five.

Distinct motivational roles

The third conclusion is that despite the fact that the VS and amygdala appear to have distinct roles in RL, both structures seem to play a role in motivation, and not in acquiring the stimulus-outcome association. The chapter five analysis revealed that monkeys from all three groups modulated their behavior in the same directions across the token tasks, performing the best in TkS. This level of

behavior modulation is where the VS monkeys differed from the other two groups, they did not increase their performance as much as the other two groups in TkS, and they lowered their performance in TkL more than the other two groups.

It is the combination of these results that leads to the conclusion that these deficits are due to motivation and not learning the stimulus-outcome contingency. The fact that both lesions groups had performance deficits in the What/Where experiment, but minimum (VS group) to no (amygdala group) deficits in the token tasks, proves that both groups have the ability to learn stimulus-outcome relationships. In the token tasks the level to which they perform is determined by environmental (task) variables. The important thing is, this fact is true for control animals as well, as we show in chapter five.

This performance modulation can also been seen in the individual token experiments. The fact that all monkeys performed at different levels across the conditions speaks to this point because the cues are the same across the blocks. Thus, when the +2 cue is learned much faster in some conditions but slower in others speaks to motivation and not ability. Further supporting this fact, we show that the absolute value difference between the two cues is not what drives better performance in some conditions versus others.

Another point to consider about the results regarding the lesion groups is in the What/Where experiment, lesion animals tended to have faster reaction times

when compared to controls. This trend is reversed for both groups in the token experiments (lesion animals tend to respond slower than controls). In fact the VS monkeys have been shown to exhibit a speed accuracy trade off (10, 11).

Reinforcement learning

In this thesis we also present data that is inconsistent with current RL theories. The first of which is how different learning environments have distinct effects on the motivational component of learning. Standard RL theories do not really take this into to account (evidence in this thesis suggest that the metrics they do have for motivation do not accurately capture it). This point is exemplified by the fact that all monkeys performed their best in TkS. This was the only token experiment on a stochastic schedule, current RL theories cannot account for this. The stochastic learning environment should be harder to learn in when compared to a deterministic learning environment. Performing better in a stochastic learning environments has to be due to motivation and not learning ability. From a behavioral perspective this result is not surprising. It is easy to imagine why deprived animals would be more motivated (greedier) in lean and or stochastic reward environments. This is a well-established behavior law, and this law is problematic for standard RL theories, because standard theories assume that animals are always trying to maximize their gains and minimize their losses (in a very specific and micro aspect). The combination of experiments in this thesis

shows that performance is not fixed. Furthermore, one should be careful at assuming they are viewing peak behavioral performance in any single learning environment.

Single value axis?

We also show that monkeys treated learning from loses different than learning from gains. In all but the last token experiment (TkL), learning was always poor in the loss/loss condition. Consistent with this finding, other work has found there to be a difference in how subjects treat gains and losses (194-196). This finding presents problems for the single value axis that is typically assumed in current RL theories.

Finally, we provide insight into a better framework for future experimental design. We show the value of understanding behavior and of testing animals on a number of different learning tasks.

Bibliography

1. J. C. Houk, J. L. Adamas, A. G. Barto, "A model of how the basal ganglia generates and uses neural signals that predict reinforcement." in Models of information processing in the basal ganglia., J. C. Houk, J. L. Davis, D. G. Beiser, Eds. (MIT Press, Cambridge, MA, 1995), pp. 249-274.
2. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997).
3. R. A. Rescorla, A. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement". (1972), vol. Vol. 2.
4. B. B. Averbeck, V. D. Costa, Motivational neural circuits underlying reinforcement learning. *Nat Neurosci* **20**, 505-512 (2017).
5. H. M. Bayer, P. W. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129-141 (2005).
6. J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, R. J. Dolan, Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron* **38**, 329-337 (2003).
7. J. O'Doherty *et al.*, Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452-454 (2004).
8. W. Schultz, Behavioral dopamine signals. *Trends Neurosci* **30**, 203-210 (2007).
9. M. Matsumoto, O. Hikosaka, Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837-841 (2009).
10. V. D. Costa, O. Dal Monte, D. R. Lucas, E. A. Murray, B. B. Averbeck, Amygdala and Ventral Striatum Make Distinct Contributions to Reinforcement Learning. *Neuron* **92**, 505-517 (2016).

11. K. M. Rothenhoefer *et al.*, Effects of ventral striatum lesions on stimulus versus action based reinforcement learning. *Journal of Neuroscience* (2017).
12. B. Balleine, S. Killcross, Effects of ibotenic acid lesions of the Nucleus Accumbens on instrumental action. *Behavioural Brain Research* **65**, 181-193 (1994).
13. R. De Borchgrave, J. N. P. Rawlins, A. Dickinson, B. W. Balleine, Effects of cytotoxic nucleus accumbens lesions on instrumental conditioning in rats. *Experimental Brain Research* **144**, 50-68 (2002).
14. R. Vicario-Feliciano, E. A. Murray, B. B. Averbeck, Ventral striatum lesions do not affect reinforcement learning with deterministic outcomes on slow time scales. *Behav Neurosci* **131**, 385-391 (2017).
15. S. Ravel, B. J. Richmond, Dopamine neuronal responses in monkeys performing visually cued reward schedules. *European Journal of Neuroscience* **24**, 277-290 (2006).
16. E. C. J. Syed *et al.*, Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nat Neurosci* **19**, 34-36 (2016).
17. H. Seo, D. Lee, Behavioral and neural changes after gains and losses of conditioned reinforcers. *J Neurosci* **29**, 3627-3641 (2009).
18. H. Kim, S. Shimojo, J. P. O'Doherty, Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol* **4**, e233 (2006).
19. M. Pessiglione, B. Seymour, G. Flandin, R. J. Dolan, C. D. Frith, Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042-1045 (2006).
20. J. A. Parkinson *et al.*, The role of the primate amygdala in conditioned reinforcement. *J Neurosci* **21**, 7770-7780 (2001).
21. R. N. Cardinal *et al.*, Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. *Behav Neurosci* **116**, 553-567 (2002).
22. P. H. Rudebeck, E. A. Murray, Amygdala and Orbitofrontal Cortex Lesions Differentially Influence Choices during Object Reversal Learning. *Journal of Neuroscience* **28**, 8338-8343 (2008).
23. P. H. Rudebeck, J. A. Ripple, A. R. Mitz, B. B. Averbeck, E. A. Murray, Amygdala Contributions to Stimulus-Reward Encoding in the Macaque Medial and Orbital Frontal Cortex during Learning. *J Neurosci* **37**, 2186-2202 (2017).

24. B. W. Balleine, A. S. Killcross, A. Dickinson, The Effect of Lesions of the Basolateral Amygdala on Instrumental Conditioning. *The Journal of Neuroscience* **23**, 666-675 (2003).
25. A. Izquierdo, E. A. Murray, Selective bilateral amygdala lesions in rhesus monkeys fail to disrupt object reversal learning. *J Neurosci* **27**, 1054-1062 (2007).
26. Y. K. Takahashi, A. J. Langdon, Y. Niv, G. Schoenbaum, Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. *Neuron* **91**, 182-193 (2016).
27. R. S. Sutton, A. G. Barto, *Introduction to Reinforcement Learning* (MIT Press, 1998).
28. B. F. Skinner, *Science and human behavior*, Science and human behavior. (Macmillan, Oxford, England, 1953), pp. x, 461-x, 461.
29. B. F. Skinner, C. B. Ferster, *Schedules of Reinforcement* (B. F. Skinner Foundation, 2015).
30. A. Neuringer, Operant variability: Evidence, functions, and theory. *Psychonomic Bulletin & Review* **9**, 672-705 (2002).
31. R. J. Herrnstein, ON THE LAW OF EFFECT1. *Journal of the Experimental Analysis of Behavior* **13**, 243-266 (1970).
32. R. Rescorla, Pavlovian conditioning. It's not what you think it is. *The American psychologist* **43 3**, 151-160 (1988).
33. J. A. Harris, B. J. Andrew, E. J. Livesey, The content of compound conditioning. *Journal of Experimental Psychology: Animal Behavior Processes* **38**, 157-166 (2012).
34. D. A. Williams, "Building a Theory of Pavlovian Conditioning From the Inside Out" in The Wiley Blackwell Handbook of Operant and Classical Conditioning. (2014), pp. 27-52.
35. J. Gibbon, M. D. Baldock, C. Locurto, L. Gold, H. S. Terrace, Trial and intertrial durations in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes* **3**, 264-284 (1977).
36. S. J. Weiss, "Instrumental and Classical Conditioning" in The Wiley Blackwell Handbook of Operant and Classical Conditioning. (2014), pp. 417-451.
37. J. J. Pear, The Science of Learning. (2014).
38. M. A. Boyle, A. N. Hoffmann, J. M. Lambert, Behavioral contrast: Research and areas for investigation. *Journal of Applied Behavior Analysis* **51**, 702-718 (2018).

39. D. M. Brethower, G. S. Reynolds, A FACILITATIVE EFFECT OF PUNISHMENT ON UNPUNISHED BEHAVIOR1. *Journal of the Experimental Analysis of Behavior* **5**, 191-199 (1962).
40. J. Crosbie, A. M. Williams, K. A. Lattal, M. M. Anderson, S. M. Brown, SCHEDULE INTERACTIONS INVOLVING PUNISHMENT WITH PIGEONS AND HUMANS. *Journal of the Experimental Analysis of Behavior* **68**, 161-175 (1997).
41. P. Dayan, B. W. Balleine, Reward, Motivation, and Reinforcement Learning. *Neuron* **36**, 285-298 (2002).
42. W. Schultz, Getting Formal with Dopamine and Reward. *Neuron* **36**, 241-263 (2002).
43. I. P. Pavlov, *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*, Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. (Oxford Univ. Press, Oxford, England, 1927), pp. xv, 430-xv, 430.
44. J. Konorski, *Integrative activity of the brain : an interdisciplinary approach* (1967).
45. A. R. Wagner, S. E. Brandon, "Evolution of a structured connectionist model of Pavlovian conditioning (AESOP)" in Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory. (Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1989), pp. 149-189.
46. S. Kuukasjarvi, Attractiveness of women's body odors over the menstrual cycle: the role of oral contraceptives and receiver sex. *Behavioral Ecology* **15**, 579-584 (2004).
47. R. Thornhill, S. Gangestad, The scent of symmetry: A human sex pheromone that signals fitness? *Evolution and Human Behavior* **20**, 175-201 (1999).
48. D. Singh, P. M. Bronstad, Female body odour is a potential cue to ovulation. *Proc Biol Sci* **268**, 797-801 (2001).
49. S. L. Miller, J. K. Maner, Scent of a woman: men's testosterone responses to olfactory ovulation cues. *Psychol Sci* **21**, 276-283 (2010).
50. K. M. Durante, V. Griskevicius, J. A. Simpson, S. M. Cantú, N. P. Li, Ovulation leads women to perceive sexy cads as good dads. *Journal of Personality and Social Psychology* **103**, 292-305 (2012).
51. R. N. Cardinal, J. A. Parkinson, J. Hall, B. J. Everitt, Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci Biobehav Rev* **26**, 321-352 (2002).

52. L. H. Corbit, B. W. Balleine, Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. *J Neurosci* **25**, 962-970 (2005).
53. M. Cador, T. W. Robbins, B. J. Everitt, Involvement of the amygdala in stimulus-reward associations: Interaction with the ventral striatum. *Neuroscience* **30**, 77-86 (1989).
54. W. A. Hershberger, An approach through the looking-glass. *Animal Learning & Behavior* **14**, 443-451 (1986).
55. P. C. Holland, Differential effects of omission contingencies on various components of Pavlovian appetitive conditioned responding in rats. *J Exp Psychol Anim Behav Process* **5**, 178-193 (1979).
56. J. A. Harris, B. J. Andrew, D. W. S. Kwok, Magazine approach during a signal for food depends on Pavlovian, not instrumental, conditioning. *Journal of Experimental Psychology: Animal Behavior Processes* **39**, 107-116 (2013).
57. M. Guitart-Masip *et al.*, Action controls dopaminergic enhancement of reward representations. *Proceedings of the National Academy of Sciences* **109**, 7511-7516 (2012).
58. T. A. Stalnaker *et al.*, Dopamine neuron ensembles signal the content of sensory prediction errors. *Elife* **8** (2019).
59. L. T. Coddington, J. T. Dudman, The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nat Neurosci* **21**, 1563-1573 (2018).
60. T. S. Braver *et al.*, Mechanisms of motivation–cognition interaction: challenges and opportunities. *Cognitive, Affective, & Behavioral Neuroscience* **14**, 443-472 (2014).
61. A. Westbrook, T. S. Braver, Dopamine Does Double Duty in Motivating Cognitive Effort. *Neuron* **89**, 695-710 (2016).
62. A. A. Hamid *et al.*, Mesolimbic dopamine signals the value of work. *Nat Neurosci* **19**, 117-126 (2016).
63. A. Nair *et al.* (2020) Opportunity cost determines action initiation latency and predicts apathy. (Center for Open Science).
64. M. W. Howe, P. L. Tierney, S. G. Sandberg, P. E. M. Phillips, A. M. Graybiel, Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**, 575-579 (2013).
65. E. T. Higgins, Value from hedonic experience and engagement. *Psychol Rev* **113**, 439-460 (2006).

66. J. E. Aberman, J. D. Salamone, Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience* **92**, 545-552 (1999).
67. M. Koch, A. Schmid, H.-U. Schnitzler, Role of nucleus accumbens dopamine D1 and D2 receptors in instrumental and Pavlovian paradigms of conditioned reward. *Psychopharmacology* **152**, 67-73 (2000).
68. K. Ishiwari, S. M. Weber, S. Mingote, M. Correa, J. D. Salamone, Accumbens dopamine and the regulation of effort in food-seeking behavior: modulation of work output by different ratio or force requirements. *Behav Brain Res* **151**, 83-91 (2004).
69. S. Mingote, S. M. Weber, K. Ishiwari, M. Correa, J. D. Salamone, Ratio and time requirements on operant schedules: effort-related effects of nucleus accumbens dopamine depletions. *European Journal of Neuroscience* **21**, 1749-1757 (2005).
70. J. D. Salamone, M. S. Cousins, S. Bucher, Anhedonia or anergia? Effects of haloperidol and nucleus accumbens dopamine depletion on instrumental response selection in a T-maze cost/benefit procedure. *Behav Brain Res* **65**, 221-229 (1994).
71. M. S. Cousins, A. Atherton, L. Turner, J. D. Salamone, Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. *Behav Brain Res* **74**, 189-197 (1996).
72. F. Denk *et al.*, Differential involvement of serotonin and dopamine systems in cost-benefit decisions about delay or effort. *Psychopharmacology* **179**, 587-596 (2005).
73. R. N. Cardinal, Impulsive Choice Induced in Rats by Lesions of the Nucleus Accumbens Core. *Science* **292**, 2499-2501 (2001).
74. R. N. Cardinal, N. J. Howes, Effects of lesions of the nucleus accumbens core on choice between small certain rewards and large uncertain rewards in rats. *BMC Neuroscience* **6**, 37 (2005).
75. J. R. Taylor, T. W. Robbins, Enhanced behavioural control by conditioned reinforcers following microinjections of d-amphetamine into the nucleus accumbens. *Psychopharmacology (Berl)* **84**, 405-412 (1984).
76. D. F. Fiorino, A. G. Phillips, Facilitation of Sexual Behavior and Enhanced Dopamine Efflux in the Nucleus Accumbens of Male Rats afterd-Amphetamine-Induced Behavioral Sensitization. *The Journal of Neuroscience* **19**, 456-463 (1999).
77. Y. Niv, N. D. Daw, D. Joel, P. Dayan, Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507-520 (2007).

78. J. A. Parkinson *et al.*, Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behav Brain Res* **137**, 149-163 (2002).
79. K. Braesicke *et al.*, Autonomic arousal in an appetitive context in primates: a behavioural and neural analysis. *Eur J Neurosci* **21**, 1733-1740 (2005).
80. C. Eisenegger *et al.*, Role of Dopamine D2 Receptors in Human Reinforcement Learning. *Neuropsychopharmacology* **39**, 2366-2375 (2014).
81. V. D. Costa, V. L. Tran, J. Turchi, B. B. Averbeck, Reversal Learning and Dopamine: A Bayesian Perspective. *Journal of Neuroscience* **35**, 2407-2416 (2015).
82. P. J. Cocker, J. G. Hosking, J. Benoit, C. A. Winstanley, Sensitivity to Cognitive Effort Mediates Psychostimulant Effects on a Novel Rodent Cost/Benefit Decision-Making Task. *Neuropsychopharmacology* **37**, 1825-1837 (2012).
83. A. Westbrook *et al.*, Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science* **367**, 1362-1366 (2020).
84. P. E. M. Phillips, G. D. Stuber, M. L. A. V. Heien, R. M. Wightman, R. M. Carelli, Subsecond dopamine release promotes cocaine seeking. *Nature* **422**, 614-618 (2003).
85. A. C. Catania, G. S. Reynolds, A QUANTITATIVE ANALYSIS OF THE RESPONDING MAINTAINED BY INTERVAL SCHEDULES OF REINFORCEMENT1. *Journal of the Experimental Analysis of Behavior* **11**, 327-383 (1968).
86. E. E. Steinberg *et al.*, A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* **16**, 966-973 (2013).
87. B. F. Sadacca, J. L. Jones, G. Schoenbaum, Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife* **5** (2016).
88. M. A. Kaufman, R. C. Bolles, A nonassociative aspect of overshadowing. *Bulletin of the Psychonomic Society* **18**, 318-320 (1981).
89. R. R. Miller, J. E. Witnauer, Retrospective revaluation: The phenomenon and its theoretical implications. *Behavioural Processes* **123**, 15-25 (2016).
90. M. J. Sharpe *et al.*, Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat Neurosci* **20**, 735-742 (2017).
91. M. J. Sharpe *et al.*, Dopamine transients do not act as model-free prediction errors during associative learning. *Nature Communications* **11** (2020).

92. A. R. Wagner, F. A. Logan, K. Haberlandt, Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology* **76**, 171-180 (1968).
93. J. M. Pearce, G. Hall, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev* **87**, 532-552 (1980).
94. J. A. Nevin, R. C. Grace, Behavioral momentum and the Law of Effect. *Behavioral and Brain Sciences* **23**, 73-90 (2000).
95. A. R. Craig, J. A. Nevin, A. L. Odum, "Behavioral Momentum and Resistance to Change" in The Wiley Blackwell Handbook of Operant and Classical Conditioning. (2014), pp. 249-274.
96. C. R. Gallistel, S. Fairhurst, P. Balsam, The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences* **101**, 13124-13131 (2004).
97. J. Lajoie, D. Bindra, Contributions of stimulus-incentive and stimulus-response-incentive contingencies to response acquisition and maintenance. *Animal Learning & Behavior* **6**, 301-307 (1978).
98. L. e. Málková, D. Gaffan, E. A. Murray, Excitotoxic Lesions of the Amygdala Fail to Produce Impairment in Visual Learning for Auditory Secondary Reinforcement But Interfere with Reinforcer Devaluation Effects in Rhesus Monkeys. *The Journal of Neuroscience* **17**, 6011-6020 (1997).
99. B. M. William, MATCHING, UNDERMATCHING, AND OVERMATCHING IN STUDIES OF CHOICE. *Journal of the Experimental Analysis of Behavior* **32**, 269-281 (1979).
100. J. J. McDowell, Matching Theory in Natural Human Environments. *The Behavior Analyst* **11**, 95-109 (1988).
101. O. H. Mowrer, Two-factor learning theory: summary and comment. *Psychol Rev* **58**, 350-354 (1951).
102. D. a. Kahneman, *Thinking, fast and slow* (1st ed. New York : Farrar, Straus and Giroux, [2011] ©2011, 2011).
103. A. M. Andrew, REINFORCEMENT LEARNING: AN INTRODUCTION by Richard S. Sutton and Andrew G. Barto, Adaptive Computation and Machine Learning series, MIT Press (Bradford Book), Cambridge, Mass., 1998, xviii + 322 pp, ISBN 0-262-19398-1, (hardback, £31.95). *Robotica* **17**, 229-235 (1999).
104. B. W. Balleine, A. Dickinson, Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407-419 (1998).

105. D. A. Gottlieb, E. L. Begej, "Principles of Pavlovian Conditioning" in The Wiley Blackwell Handbook of Operant and Classical Conditioning. (2014), pp. 1-25.
106. T. L. Davidson, A. M. Altizer, S. C. Benoit, E. K. Walls, T. L. Powley, Encoding and selective activation of "metabolic memories" in the rat. *Behav Neurosci* **111**, 1014-1030 (1997).
107. P. C. Holland, J. J. Straub, Differential effects of two ways of devaluing the unconditioned stimulus after Pavlovian appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes* **5**, 65-78 (1979).
108. A. Dickinson, B. Balleine, Motivational control of goal-directed action. *Animal Learning & Behavior* **22**, 1-18 (1994).
109. S. Siegel, Pavlovian conditioning analysis of morphine tolerance. *NIDA Res Monogr*, 27-53 (1978).
110. J. R. Macrae, M. T. Scoles, S. Siegel, The Contribution of Pavlovian Conditioning to Drug Tolerance and Dependence. *Addiction* **82**, 371-380 (1987).
111. R. Sinha, C. S. Li, Imaging stress- and cue-induced drug and alcohol craving: association with relapse and clinical implications. *Drug Alcohol Rev* **26**, 25-31 (2007).
112. K. Van Den Akker, G. Schyns, A. Jansen, Learned Overeating: Applying Principles of Pavlovian Conditioning to Explain and Treat Overeating. *Current Addiction Reports* **5**, 223-231 (2018).
113. Y. K. Takahashi *et al.*, Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron* **95**, 1395-1405.e1393 (2017).
114. N. J. Mackintosh, Ed., *Animal learning and cognition* (Academic Press, New York, 1994), 1 Ed, p 379.
115. B. W. Balleine, M. Liljeholm, S. B. Ostlund, The integrative function of the basal ganglia in instrumental conditioning. *Behav Brain Res* **199**, 43-52 (2009).
116. L. H. Corbit, B. W. Balleine, Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**, 962-970 (2005).
117. L. H. Burns, T. W. Robbins, B. J. Everitt, 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* **55**, 167-183 (1993).

118. G. B. Bissonette, R. N. Gentry, S. Padmala, L. Pessoa, M. R. Roesch, Impact of appetitive and aversive outcomes on brain responses: linking the animal and human literatures. *Frontiers in Systems Neuroscience* **8** (2014).
119. L. G. Ungerleider, M. Mishkin, "Two cortical visual systems" in Analysis of Visual Behavior, D. J. Ingle, M. A. Goodale, R. J. W. Mansfield, Eds. (MIT Press, Cambridge, 1982), pp. 549-587.
120. M. A. Goodale, A. D. Milner, Separate visual pathways for perception and action. *Trends Neurosci* **15**, 20-25 (1992).
121. H. Barbas, Anatomic organization of basoventral and mediodorsal visual recipient prefrontal regions in the rhesus monkey. *J Comp Neurol* **276**, 313-342. (1988).
122. M. J. Webster, J. Bachevalier, L. G. Ungerleider, Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cereb Cortex* **4**, 470-483 (1994).
123. B. B. Averbeck, J. Lehman, M. Jacobson, S. N. Haber, Estimates of projection overlap and zones of convergence within frontal-striatal circuits. *J Neurosci* **34**, 9497-9505 (2014).
124. S. N. Haber, K. S. Kim, P. Mailly, R. Calzavara, Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J Neurosci* **26**, 8368-8376 (2006).
125. A. B. Sereno, J. H. Maunsell, Shape selectivity in primate lateral intraparietal cortex. *Nature* **395**, 500-503 (1998).
126. T. J. Bussey, S. P. Wise, E. A. Murray, Interaction of ventral and orbital prefrontal cortex with inferotemporal cortex in conditional visuomotor learning. *Behav Neurosci* **116**, 703-715 (2002).
127. E. O. Neftci, B. B. Averbeck, Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence* **1**, 133-143 (2019).
128. C. A. Taswell, V. D. Costa, E. A. Murray, B. B. Averbeck, Ventral striatum's role in learning from gains and losses. *Proc Natl Acad Sci U S A* **115**, E12398-E12406 (2018).
129. N. F. Parker *et al.*, Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience* **19**, 845-854 (2016).
130. B. Lau, P. W. Glimcher, Value representations in the primate striatum during matching behavior. *Neuron* **58**, 451-463 (2008).
131. K. Samejima, Y. Ueda, K. Doya, M. Kimura, Representation of action-specific reward values in the striatum. *Science* **310**, 1337-1340 (2005).

132. M. Seo, E. Lee, B. B. Averbeck, Action selection and action value in frontal-striatal circuits. *Neuron* **74**, 947-960 (2012).
133. E. Lee, M. Seo, O. Dal Monte, B. B. Averbeck, 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).
134. J. J. Paton, M. A. Belova, S. E. Morrison, C. D. Salzman, The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439**, 865-870 (2006).
135. M. A. Belova, J. J. Paton, C. D. Salzman, Moment-to-moment tracking of state value in the amygdala. *J Neurosci* **28**, 10023-10030 (2008).
136. A. N. Hampton, R. Adolphs, M. J. Tyszka, J. P. O'Doherty, Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex. *Neuron* **55**, 545-555 (2007).
137. M. G. Baxter, E. A. Murray, The amygdala and reward. *Nat Rev Neurosci* **3**, 563-573 (2002).
138. C. D. Salzman, S. Fusi, Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. *Annu Rev Neurosci* **33**, 173-202 (2010).
139. P. H. Rudebeck, A. R. Mitz, R. V. Chacko, E. A. Murray, Effects of amygdala lesions on reward-value coding in orbital and medial prefrontal cortex. *Neuron* **80**, 1519-1531 (2013).
140. D. G. Amaral, J. L. Price, A. Pitkanen, S. T. Carmichael, "Anatomical organization of the primate amygdaloid complex" in The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunctdion, J. P. Aggleton, Ed. (Wiley-Liss, New York, 1992), pp. 1-66.
141. D. G. Amaral, J. L. Price, Amygdalo-cortical projections in the monkey (Macaca fascicularis). *Journal of Comparative Neurology* **230**, 465-496 (1984).
142. V. D. Costa, A. R. Mitz, B. B. Averbeck, Subcortical Substrates of Explore-Exploit Decisions in Primates. *Neuron* **103**, 533-545 e535 (2019).
143. E. L. Peck, C. J. Peck, C. D. Salzman, Task-dependent spatial selectivity in the primate amygdala. *J Neurosci* **34**, 16220-16233 (2014).
144. R. J. Morecraft *et al.*, Amygdala interconnections with the cingulate motor cortex in the rhesus monkey. *J Comp Neurol* **500**, 134-165 (2007).
145. J. Seidlitz *et al.*, A population MRI brain template and analysis tools for the macaque. *Neuroimage* **170**, 121-131 (2018).
146. R. W. Cox, AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* **29**, 162-173 (1996).

147. B. M. Basile, C. L. Karaskiewicz, E. C. Fiuzat, L. Malkova, E. A. Murray, MRI Overestimates Excitotoxic Amygdala Lesion Damage in Rhesus Monkeys. *Frontiers in Integrative Neuroscience* **11**, 12 (2017).
148. V. Willenbockel *et al.*, Controlling low-level image properties: the SHINE toolbox. *Behavior research methods* **42**, 671-684 (2010).
149. W. F. Asaad, E. N. Eskandar, A flexible software tool for temporally-precise behavioral control in Matlab. *Journal of neuroscience methods* **174**, 245-258 (2008).
150. V. D. Costa, V. L. Tran, J. Turchi, B. B. Averbeck, Reversal learning and dopamine: a bayesian perspective. *J Neurosci* **35**, 2407-2416 (2015).
151. A. I. Jang *et al.*, The Role of Frontal Cortical and Medial-Temporal Lobe Brain Areas in Learning a Bayesian Prior Belief on Reversals. *J Neurosci* **35**, 11751-11760 (2015).
152. J. H. Zar, *Biostatistical Analysis* (Prentice Hall, Upper Saddle River, 1999).
153. S. Olejnik, J. Algina, Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. *Contemporary Educational Psychology* **25**, 241-286 (2000).
154. S. J. Gershman, B. Pesaran, N. D. Daw, Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values. *Journal of Neuroscience* **29**, 13524-13531 (2009).
155. B. Lau, P. W. Glimcher, Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys. *Journal of the Experimental Analysis of Behavior* **84**, 555-579 (2005).
156. R. Bartolo, B. B. Averbeck, Prefrontal Cortex Predicts State Switches during Reversal Learning. *Neuron* (2020).
157. B. K. Chau *et al.*, Contrasting Roles for Orbitofrontal Cortex and Amygdala in Credit Assignment and Learning in Macaques. *Neuron* **87**, 1106-1118 (2015).
158. P. H. Rudebeck, R. C. Saunders, D. A. Lundgren, E. A. Murray, Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes. *Neuron* **95**, 1208-1220 e1205 (2017).
159. N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, R. J. Dolan, Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876-879 (2006).
160. G. E. Alexander, M. R. DeLong, P. L. Strick, Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci* **9**, 357-381. (1986).

161. S. E. Rhodes, E. A. Murray, Differential effects of amygdala, orbital prefrontal cortex, and prelimbic cortex lesions on goal-directed behavior in rhesus macaques. *J Neurosci* **33**, 3380-3389 (2013).
162. P. H. Rudebeck, E. A. Murray, Amygdala and orbitofrontal cortex lesions differentially influence choices during object reversal learning. *J Neurosci* **28**, 8338-8343 (2008).
163. L. M. Romanski *et al.*, Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat Neurosci* **2**, 1131-1136. (1999).
164. B. B. Averbeck, E. A. Murray, Hypothalamic Interactions with Large-Scale Neural Circuits Underlying Reinforcement Learning and Motivated Behavior. *Trends in Neurosciences* (2020).
165. R. Caminiti *et al.*, Computational Architecture of the Parieto-Frontal Network Underlying Cognitive-Motor Control in Monkeys. *eNeuro* **4** (2017).
166. M. A. Steinmetz, B. C. Motter, C. J. Duffy, V. B. Mountcastle, Functional properties of parietal visual neurons: radial organization of directionalities within the visual field. *J Neurosci* **7**, 177-191. (1987).
167. M. Mascaro, A. Battaglia-Mayer, L. Nasi, D. J. Amit, R. Caminiti, The eye and the hand: neural mechanisms and network models for oculomanual coordination in parietal cortex. *Cereb Cortex* **13**, 1276-1286 (2003).
168. R. Desimone, T. D. Albright, C. G. Gross, C. Bruce, Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci* **4**, 2051-2062. (1984).
169. D. L. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* **111**, 8619-8624 (2014).
170. B. B. Averbeck, M. Seo, The statistical neuroanatomy of frontal networks in the macaque. *PLoS Comput Biol* **4**, e1000050 (2008).
171. S. C. Rao, G. Rainer, E. K. Miller, Integration of what and where in the primate prefrontal cortex. *Science* **276**, 821-824. (1997).
172. E. A. Murray, S. P. Wise, Role of the hippocampus plus subjacent cortex but not amygdala in visuomotor conditional learning in Rhesus monkeys. *Behav Neurosci* **110**, 1261-1270 (1996).
173. R. Bartolo, R. C. Saunders, A. R. Mitz, B. B. Averbeck, Information-Limiting Correlations in Large Neural Populations. *The Journal of Neuroscience* **40**, 1668 (2020).

174. B. H. Turner, M. Mishkin, M. Knapp, Organization of the amygdalopetal projections from modality-specific cortical association areas in the monkey. *J Comp Neurol* **191**, 515-543 (1980).
175. D. P. Friedman, J. P. Aggleton, R. C. Saunders, Comparison of hippocampal, amygdala, and perirhinal projections to the nucleus accumbens: combined anterograde and retrograde tracing study in the Macaque brain. *The Journal of Comparative Neurology* **450**, 345-365 (2002).
176. H. T. Ghashghaei, H. Barbas, Pathways for emotion: interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience* **115**, 1261-1279. (2002).
177. H. T. Ghashghaei, C. C. Hilgetag, H. Barbas, Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala. *Neuroimage* **34**, 905-923 (2007).
178. F. T. Russchen, D. G. Amaral, J. L. Price, The afferent input to the magnocellular division of the mediodorsal thalamic nucleus in the monkey, Macaca fascicularis. *J Comp Neurol* **256**, 175-210 (1987).
179. P. S. Goldman-Rakic, L. J. Porrino, The primate mediodorsal (MD) nucleus and its projection to the frontal lobe. *J Comp Neurol* **242**, 535-560. (1985).
180. C. J. Peck, B. Lau, C. D. Salzman, The primate amygdala combines information about space and value. *Nat Neurosci* **16**, 340-348 (2013).
181. C. J. Peck, C. D. Salzman, Amygdala neural activity reflects spatial attention towards stimuli promising reward or threatening punishment. *Elife* **3** (2014).
182. C. J. Peck, C. D. Salzman, The amygdala and basal forebrain as a pathway for motivationally guided attention. *J Neurosci* **34**, 13757-13767 (2014).
183. E. A. Murray, P. H. Rudebeck, Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. *Nature Reviews Neuroscience* (2018).
184. S. M. Groman *et al.*, Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. *Neuron* **103**, 734-746 e733 (2019).
185. T. A. Stalnaker, T. M. Franz, T. Singh, G. Schoenbaum, Basolateral amygdala lesions abolish orbitofrontal-dependent reversal impairments. *Neuron* **54**, 51-58 (2007).
186. R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267-279 (2014).
187. N. W. Schuck, M. B. Cai, R. C. Wilson, Y. Niv, Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron* **91**, 1402-1412 (2016).

188. C. K. Starkweather, S. J. Gershman, N. Uchida, The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron* **98**, 616-629 e616 (2018).
189. D. Durstewitz, N. M. Vittoz, S. B. Floresco, J. K. Seamans, Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron* **66**, 438-448 (2010).
190. M. Sarafyazd, M. Jazayeri, Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364** (2019).
191. R. B. Ebitz, E. Albarran, T. Moore, Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron* **97**, 450-461 e459 (2018).
192. M. P. Karlsson, D. G. Tervo, A. Y. Karpova, Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science* **338**, 135-139 (2012).
193. P. Jean-Richard-Dit-Bressel, S. Killcross, G. P. McNally, Behavioral and neurobiological mechanisms of punishment: implications for psychiatric disorders. *Neuropsychopharmacology* (2018).
194. J. Kubanek, L. H. Snyder, R. A. Abrams, Reward and punishment act as distinct factors in guiding behavior. *Cognition* **139**, 154-167 (2015).
195. E. B. Rasmussen, M. C. Newland, Asymmetry of reinforcement and punishment in human choice. *J Exp Anal Behav* **89**, 157-167 (2008).
196. S. Farashahi, H. Azab, B. Hayden, A. Soltani, On the Flexibility of Basic Risk Attitudes in Monkeys. *J Neurosci* **38**, 4383-4398 (2018).
197. P. Namburi *et al.*, A circuit mechanism for differentiating positive and negative associations. *Nature* **520**, 675-678 (2015).
198. F. Ambroggi, A. Ishikawa, H. L. Fields, S. M. Nicola, Basolateral amygdala neurons facilitate reward-seeking behavior by exciting nucleus accumbens neurons. *Neuron* **59**, 648-661 (2008).
199. B. J. Everitt, K. A. Morris, A. O'Brien, T. W. Robbins, The basolateral amygdala-ventral striatal system and conditioned place preference: further evidence of limbic-striatal interactions underlying reward-related processes. *Neuroscience* **42**, 1-18 (1991).
200. S. Lammel *et al.*, Input-specific control of reward and aversion in the ventral tegmental area. *Nature* **491**, 212-217 (2012).
201. R. Vicario-Feliciano, E. A. Murray, B. B. Averbeck, Ventral striatum lesions do not affect reinforcement learning with deterministic outcomes on slow time scales. *Behav Neurosci* **131**, 385-391 (2017).
202. S. J. Gershman, B. Pesaran, N. D. Daw, Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The*

Journal of neuroscience : the official journal of the Society for Neuroscience **29**, 13524-13531 (2009).
203. B. B. Averbeck, Amygdala and Ventral Striatum Population Codes Implement Multiple Learning Rates for Reinforcement Learning. *IEEE Symposium Series on Computational Intelligence* (2017).
204. K. T. Beier *et al.*, Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. *Cell* **162**, 622-634 (2015).
205. W. Menegas *et al.*, Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *Elife* **4** (2015).
206. A. Beyeler *et al.*, Divergent Routing of Positive and Negative Information from the Amygdala during Memory Retrieval. *Neuron* **90**, 348-361 (2016).
207. A. H. Taub, R. Perets, E. Kahana, R. Paz, Oscillations Synchronize Amygdala-to-Prefrontal Primate Circuits during Aversive Learning. *Neuron* **97**, 291-298 e293 (2018).
208. H. F. Clarke, N. K. Horst, A. C. Roberts, Regional inactivations of primate ventral prefrontal cortex reveal two distinct mechanisms underlying negative bias in decision making. *Proc Natl Acad Sci U S A* **112**, 4176-4181 (2015).
209. J. E. LeDoux, Emotion circuits in the brain. *Annual Review of Neuroscience* **23**, 155-184 (2000).
210. S. Palminteri *et al.*, Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* **76**, 998-1009 (2012).
211. C. Eisenegger *et al.*, Role of dopamine D2 receptors in human reinforcement learning. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **39**, 2366-2375 (2014).
212. N. Lally *et al.*, The Neural Basis of Aversive Pavlovian Guidance during Planning. *J Neurosci* **37**, 10215-10229 (2017).
213. K. Amemori, A. M. Graybiel, Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. *Nature Neuroscience* **15**, 776-785 (2012).
214. C. T. Gross, N. S. Canteras, The many paths to fear. *Nat Rev Neurosci* **13**, 651-658 (2012).
215. M. F. Roitman, R. A. Wheeler, R. M. Carelli, 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 (2005).
216. M. F. Roitman, R. A. Wheeler, P. H. Tiesinga, J. D. Roitman, R. M. Carelli, Hedonic and nucleus accumbens neural responses to a natural reward are regulated by aversive conditioning. *Learning & memory* **17**, 539-546 (2010).
217. A. Badrinarayan *et al.*, Aversive stimuli differentially modulate real-time dopamine transmission dynamics within the nucleus accumbens core and

shell. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**, 15779-15790 (2012).
218. E. A. Budygin *et al.*, Aversive stimulus differentially triggers subsecond dopamine release in reward regions. *Neuroscience* **201**, 331-337 (2012).
219. J. Park, E. S. Bucher, E. A. Budygin, R. M. Wightman, Norepinephrine and dopamine transmission in 2 limbic regions differentially respond to acute noxious stimulation. *Pain* **156**, 318-327 (2015).
220. E. Navratilova, C. W. Atcherley, F. Porreca, Brain Circuits Encoding Reward from Pain Relief. *Trends in Neurosciences* (2015).
221. M. F. Roitman, R. A. Wheeler, R. M. Wightman, R. M. Carelli, Real-time chemical responses in the nucleus accumbens differentiate rewarding and aversive stimuli. *Nat Neurosci* **11**, 1376-1377 (2008).
222. S. Chakraborty, N. Kolling, M. E. Walton, A. S. Mitchell, Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *Elife* **5** (2016).
223. E. A. Murray, P. H. Rudebeck, Specializaations for reward-guided decision-making in the primate ventral prefrontal cortex. *Nature Reviews Neuroscience* (2018).
224. C. A. Taswell *et al.*, Effects of Amygdala Lesions on Object-Based Versus Action-Based Learning in Macaques. *Cerebral Cortex* **31**, 529-546 (2021).
225. M. R. Roesch, Neuronal Activity Related to Reward Value and Motivation in Primate Frontal Cortex. *Science* **304**, 307-310 (2004).
226. A. Litt, H. Plassmann, B. Shiv, A. Rangel, Dissociating Valuation and Saliency Signals during Decision-Making. *Cerebral Cortex* **21**, 95-102 (2011).
227. G. B. Bissonette *et al.*, Separate Populations of Neurons in Ventral Striatum Encode Value and Motivation. *PLoS ONE* **8**, e64673 (2013).
228. K. M. Rothenhoefer, T. Hong, A. Alikaya, W. R. Stauffer, Rare rewards amplify dopamine responses. *Nat Neurosci* **24**, 465-469 (2021).
229. L. H. Burns, T. W. Robbins, B. J. Everitt, 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. *Behav Brain Res* **55**, 167-183 (1993).