BASAL GANGLIA INVOLVEMENT IN THE REINFORCEMENT LEARNING OF PHYSICAL AND COGNITIVE ACTIONS

by

Patryk Alix Laurent

B.A. Cognitive Sciences, University of Virginia, 2001

Submitted to the Graduate Faculty of

the Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

CENTER FOR NEUROSCIENCE

This dissertation was presented

by

Patryk Alix Laurent

It was defended on

November 13th, 2009

and approved by

Erik D. Reichle, Department of Psychology, University of Pittsburgh

Julie A. Fiez, Department of Neuroscience, University of Pittsburgh

Peter L. Strick, Department of Neurobiology, School of Medicine, University of Pittsburgh

Marc A. Sommer, Department of Neuroscience, University of Pittsburgh

Mark E. Wheeler, Department of Psychology, University of Pittsburgh

A. David Redish, Department of Neuroscience, University of Minnesota

Dissertation Director: Erik D. Reichle, Department of Psychology, University of Pittsburgh

BASAL GANGLIA INVOLVEMENT IN THE REINFORCEMENT LEARNING OF PHYSICAL AND COGNITIVE ACTIONS

Patryk Alix Laurent, PhD

University of Pittsburgh, 2010

Work is presented aimed at understanding the function of the basal ganglia in reward-related learning. Behavioral, fMRI, and computational techniques are used to examine basal ganglia activation during the reinforcement of physical (i.e., motoric) and cognitive (i.e., non-motoric) actions. In a single experiment design, participants received positive and negative reinforcement for performing actions in one of four possible directions depending on a color cue stimulus. During different phases of the experiment, participants performed either hand movements, eye movements, or covert attention shifts. Behavioral and fMRI data collected during the task were used to test predictions from simulated Reinforcement Learning (RL) agents trained on the same sequences of stimulus, action, and outcome experienced by the human participants. Behavioral data showed that participants were able to learn the three types of action equally well and at similar rates, providing behavioral evidence that a common algorithm might be involved. Further, RL simulations fit the learning of the three types of action equally well, suggesting that RL might be that common algorithm. A deconvolution analysis of striatal fMRI BOLD data suggested that: (1) the striatum computed reward prediction errors for both physical and cognitive actions, and (2) this computation was localized to different regions of the striatum depending on the type of action that was being rewarded. The localization of these computations replicates prior findings implicating those regions in action-specific voluntary control, but extends them to include the fact that these regions compute action-specific reward-prediction errors. Together, these data suggest that distinct circuits linking neocortex with the basal ganglia are involved in RL-related computations for the actions controlled by those circuits. The results point to a theoretical framework in which the basal ganglia

mediate the reinforcement of actions whose control is delegated to local neocortical regions. Implications for theories of learning, theories of cognitive control, and mapping RL to the basal ganglia are discussed. Finally, the possibility that phasic dopamine might mediate the prediction error signals is considered in view of some theoretical challenges such as its non-specificity, its asymmetric representation of negative reinforcement, and its presence during non-rewarding events.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The research reported in this dissertation arose from numerous conversations and interactions with the faculty and graduate students at the Center for Neuroscience at the University of Pittsburgh (CNUP), the Center for the Neural Basis of Cognition (CNBC), and the Learning Research and Development Center (LRDC). I am very thankful for the supportive and intellectually stimulating environment.

I would also like to acknowledge the funding sources that made this research possible. These include National Institutes of Health grant NIH R01 HD053639 to Erik D. Reichle and Tessa Warren, and a National Science Foundation Integrative Graduate Education and Traineeship (IGERT) fellowship provided by the Center for the Neural Basis of Cognition.

As for my research training, I am especially indebted to Dr. Chip Levy for introducing me to academic research and the importance of good computational modeling, to Dr. Julie Fiez for making this whole project possible as my IGERT cross-over mentor, and to Dr. Erik Reichle for his advising and insight.

Finally, I would like to thank my wife Dr. Abby Laurent, for her support and encouragement.

1.0 INTRODUCTION

Learning, the acquisition through experience of behavioral responses, skills, and knowledge that are not prewired into a neural system, is an important area of research in cognitive neuroscience. How does learning occur in the brain? According to current neuroscience theory, learning takes place through the modification of interneuronal connections called *synapses*. Synapses mediate the transmission of information from one neuron (e.g., a sensory neuron) to another neuron (e.g., a motor neuron) through the release and detection of chemical and/or electrical signals. Once sufficient numbers of these synapses have been sufficiently modified, the organism is then said to have learned, because its behavioral output will have changed in response to the same stimulus. For example, before learning, an organism upon hearing an auditory stimulus such as "press the A button" or "press the B button" may not generate the desired behavior. The organism may perform an incorrect action either randomly or arbitrarily. However, after many repeated trials during which the organism is positively reinforced when it performs the task correctly, its behavior may converge to the reinforced behavior.

A couple of questions immediately arise from the previous example. First, if learning is indeed achieved through the modification of synapses, which synapses are the ones that are modified? And relatedly, how did those particular synapses get modified while others did not? Is there any organization to the localization of learning-related changes?

What exactly has been learned? The organism has likely learned many different things during the course of the experiment – presumably with corresponding synaptic changes in several parts of its nervous system. Perhaps the most salient thing learned is the successful mapping of the stimulus to the appropriate response. Learning to make a response given a stimulus and reinforcement is called *operant conditioning* or *instrumental conditioning*, and has a long history of study. How-

ever, it is known that operant conditioning is not the only form of learning that occurs in the brain. Researchers have identified many different kinds of learning and have begun to map these to different brain regions. Doya, for example, identifies three forms of learning (Doya, 1999, 2002). The first form of learning is related to learning the statistics of the environment, leading to improved perception and/or motor control. In the above example of learning to press buttons labeled "A" or "B", this learning might involve improvements in auditory perception, required to discriminate the two possible command stimuli. It may also involve visual learning, required to identify the writing on the two buttons or their spatial configuration. Or this statistical learning may involve fine tuning motor cortex representations to generate hand movements that are sufficiently accurate to press one button and not the other. This general form of "statistical learning" is attributed to the cerebral cortex, the convoluted surface portion of the brain.

The second form of learning identified by Doya is related to compensating for errors due to changes in the dynamics of the environment compared to the organisms predictions for the environment. Such changes might occur, for example, in body movements while carrying a heavy object. In the above example task, this type of learning may be involved in adapting to the differences in which the voice command is given (e.g., a high pitched voice or a lower pitched voice, or accompanied by distortion), as well as in the adjustments made during motor control as various amounts of error in the muscle movements require adjustments in the movements and posture of the participant. This "error-correcting" form of learning is attributed to the cerebellum, the "little brain" that is located below the rest of the brain, posterior to the brainstem.

The third form of learning identified by Doya is learning to predict the correct amount and type of reinforcement feedback as a result of performing actions, which allows selecting actions that are more likely to lead to reward. In the above example task, this form of learning is involved in making the choice to either press "A" or "B" given the history of reinforcement. This "reinforcement" learning is attributed to the basal ganglia, structures that are not readily visible in most depictions of the brain because they are buried beneath the cerebral cortex, and are thus known as "subcortical structures". Despite the fact that these different forms of learning may be variously engaged in the service of any task, either simultaneously, sequentially or in some combination of these, the use of carefully-designed experiments allow researchers to study them independently.

Although this last form of learning – called Reinforcement Learning (RL) – has been attributed to the basal ganglia, there remain many unanswered questions about the role of the basal ganglia in RL. This dissertation seeks to answer the following questions about basal ganglia involvement in learning through reinforcement: Is there neural activity in the basal ganglia that relates to the learning of physical actions? Is there neural activity in the basal ganglia that relates to the learning of cognitive "actions" (i.e., operations that do not require physical movements, such as shifting covert attention)? Is the neural activity found in the same part of the basal ganglia (e.g., the nucleus accumbens) or in action-specific regions? And finally, does the localization of action-specific neural activity coincide with neural activity related to the actions being learned? That is, do both learning and action take place in the same location?

Effectively, these questions aim to elucidate critical information about how RL may be implemented in the brain's neural circuits. This information is important because without it, progress in achieving a deep understanding of how learning occurs in the brain will be difficult.

1.1 REINFORCEMENT LEARNING (RL)

RL has long parallel histories in computer science and behavioral psychology. RL can be formally defined as the general problem of learning to predict reinforcement in order to select actions to maximize the amount of reward that an agent (i.e., an organism or a robot) will obtain from its environment. Thus, RL refers to the problem of, as well as solutions for, learning to select actions to maximize the amount of reward that an agent or organism can obtain from its environment (Sutton and Barto, 1998). In this context, an 'agent' is any autonomous system that is capable of initiating or selecting actions, for example, a robot or other autonomous machine. A slot machine would typically *not* be considered an agent, because although it has internal machinery and appears to select actions (i.e., to choose what to display), it does not have autonomy: that is, an agent (usually a human) pulls the lever to directly cause the resulting actions.

A complete description of an RL problem requires three things to be specified: (1) a list of the possible actions that an agent can take from any given state, (2) a description of how selected actions

move the agent through the environment (i.e., a *state transition function*), and (3) a description the amount of reinforcement that is received by an agent upon entering a state or performing a particular action (i.e., a *reinforcement function*).

Once the problem has been specified, there are three important issues specific to the implementation of RL itself. The first issue in specifying RL algorithms is the type of values that are learned by the agent. Because the RL problems involve maximizing reward, optimal algorithms for solving them store the amount of reward that can be obtained (i.e., reward predictions) in some form or another. Reward predictions in computational simulations of RL agents are stored in a structure generally known as a **value function**. There are two main kinds of value functions. One kind of value function is a state-based value function, where reward predictions are associated with specific states. By convention, state-based value functions are notated as $V(s)$, i.e., the values associated with each state s. The other kind of value instead associates values with each action that could be taken from a particular state. Action-based value functions are frequently notated as $Q(s, a)$, i.e., the value associated with taking action a from state s. In the former case, where values are simply associated with states, algorithms are either equipped with or must learn an internal model so that they can predict how a selected action leads them to the next state, thereby allowing the algorithm to infer the value of an action. In the latter case, the algorithms simply associate values directly with actions without regard to the consequences of those actions.

The second issue is how the reward predictions are stored. The simplest storage mechanism is what is termed a *lookup-table*: that is, when a value is associated with a state or state-action pair, that value is kept in a perfect memory system that is guaranteed to return exactly that value when it is retrieved. Such a mechanism is very stable and effective, but can use enormous amounts of computer memory in simulations. Neural networks and other function approximation techniques have been explored because of this problem of computer memory, and the fact that lookup-tables do not generalize values between similar states. Unfortunately, a known problem with function approximation is that the learning is very slow and often unstable, leading to failures to store correct reward predictions within reasonable times.

The third implementation issue is how the reward predictions are learned. It should be noted that although many algorithms exist for learning reward predictions, each with its own advantages

and disadvantages, only a few of these algorithms are biologically plausible and therefore of interest to cognitive neuroscience. For example, at one extreme, RL algorithms exist that iterate and update the values through the entire space of states that could ever be encountered (e.g., the *value-iteration* algorithm), whereas at the other extreme updates are made to a single state at a time according to semi-random exploration of the more likely portions of the state space (e.g., the *trajectory sampling* algorithm.) Although the value-iteration algorithm is guaranteed to lead to optimal behavior, it is rarely reasonable for either computer simulations or real organisms to repeatedly visit every possible state in the simulated environment because there are too many due to a combinatorial explosion. The actions required to perform a task (e.g., making a sandwich) could be performed in any number of timings and orders (e.g., combining the various ingredients, which ones should be toasted, whether to use a knife or spoon to spread toppings, etc.). Thus, the idea of value iteration is not likely to be biologically plausible or a reasonable approximation of animal learning. The serial nature of trajectories can mimic learning through experience with an environment, and/or through experience with a mental model of an environment.

Mathematically, RL algorithms are formally concerned with maximizing the total expected reinforcement an agent will obtain. The simplest learning situation are those in which episodes have a single action after which reinforcement is obtained, but more generally, reinforcement is obtained after one or more actions. RL algorithms generally update reward predictions for a given state using some form of the *Rescorla-Wagner learning rule* (Rescorla and Wagner, 1972) by comparing their current reward predictions, typically initialized to zero or random numbers at the beginning of the simulation, to the reward that is actually obtained during the transition to the subsequent state. (For more detail on the Rescorla-Wagner learning rule, see section 1.2 on page 9.) This difference is called "temporal-difference" and the use of this difference to update values is called *temporal-difference learning*.

Thus, in general the total cumulative value associated with any state (or, in alternative formulations, action from a particular state) is:

$$
\begin{aligned}
V(currentState) &= \text{total reward predicted from currentState} \\
V(currentState) &= r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^3 r_4 + \ldots \\
V(currentState) &= r_0 + \gamma(r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots) \\
V(currentState) &= r_0 + \gamma V(nextState)
\end{aligned}
$$

Here, r refers to reinforcement, and γ ("gamma") refers to a discount factor that can be set to a value less than one so that immediate reward is more valuable than future reward. If the value of a state does not adequately predict its future reward, the left and right sides of the above equation are not equal. The difference is conventionally written as δ ("delta"):

$$
\begin{aligned}
\delta &= r_0 + \gamma V(nextState) - V(currentState) \\
V_{new}(currentState) &= V(currentState) + \delta
\end{aligned}
$$

For a simple example of an RL program and the types of information typically stored by an RL agent, see Figure 1.1 (page 16).

The preceding discussion has focused on defining RL from the perspective of computer science research. However it is important to note that research on RL takes place in a variety of disciplines and is motivated by wide-ranging interests. In neuroscience and psychology in particular, RL is studied because it is a potential explanation for how behaviors may be learned through positive and negative feedback. There are a number of reasons that RL is a particularly interesting form of learning on which to study from a neuroscience perspective. One specific reason is that RL models can be implemented and make specific numerical predictions about behavior. RL simulations output behaviors that can be compared directly to the behaviors performed by a human. Even more interesting for neuroscience research is that the result of an internal computation required by most RL algorithms – the computation of reward-prediction error – appears to be transmitted in the phasic activity of dopamine-releasing neurons. Thus, RL simulations also output values that can be used to test hypotheses about brain mechanisms involved in RL. In general, RL offers a rich theoretical framework for research because of a convergence of results from behavioral neuroscience, neuroanatomy, functional neuroimaging, computer science, and mathematics. This convergence

has begun to improve our understanding about how positive and negative feedback alter the brain at the behavioral, systems, and synaptic levels.

More recently, possible connections between RL and learning in the brain has been investigated more closely. Studies suggest how RL mechanisms could lead to complex behaviors that organisms perform. For example, simulations of simple RL agents have been successfully used to explain how perceptual, cognitive and motor processes may be learned and coordinated to enable efficient, intelligent behavior (Ballard et al., 1997; Botvinick et al., 2008; Laurent, 2008; Reichle and Laurent, 2006).

The link between RL and computation in the brain has a basis in neurophysiology, that is, the study of the electrical activity of cells. Neurophysiology studies have shown that a key quantity in RL algorithms, the difference between successive reward predictions or "reward-prediction error", is conveyed by the phasic activity of dopamine-releasing neurons controlled by the basal ganglia. Functional MRI studies have mirrored these findings in the basal ganglia of humans, specifically in the striatum which is one of the main targets of dopamine release. Further, neuroanatomical studies suggest that the basal ganglia not only maintain connections with motor control regions (Gerardin et al., 2003), but also share the same types of connections with cognitive control regions (Lehéricy et al., 2004a,b; Leichnetz, 2001; Middleton and Strick, 1996, 2000).

Together, these prior results all point to a single as of yet unresolved question which is the main focus of this dissertation: Are the basal ganglia involved in the RL of physical and cognitive actions? More specifically, when humans are reinforced for performing actions, such as saccadic eye movements, are reward-prediction errors computed for each saccade based on previously stored reward predictions? Are the same signatures of learning present if the actions are body movements (i.e., skeletomotor), as well as non motoric "actions", like covert shifts of visuospatial attention?

The experiments presented in Chapter 3 of this dissertation are aimed squarely at answering these questions. Performing these experiments will also allow us to investigate whether distinct basal ganglia circuits are involved in RL for different types of actions, providing a specific neural substrate for the computations posited by RL. In particular, the experiments will yield information as to whether the computation of reward-prediction errors is differentially localized in the striatum depending on the type of action that is being reinforced. This information can then be compared to

what is already known about the functional organization of the basal ganglia through anatomical, electrophysiological, and neuroimaging studies. The mapping of RL computations to a neural network substrate will generate computational and neuroscientific insights into how the brain achieves control of itself, and may also result in the eventual development of fine-tuned diagnosis techniques for basal ganglia disorders.

It is the hope that carrying out a neuroscientific study of the mapping between RL computations and a particular neural substrate will lead to a number of theoretical insights. One such benefit is the ability to make sense of certain patterns of neural activity that might otherwise be difficult to explain. For example, it is known that the phasic activity of dopamine neurons does not only transmit reward-prediction errors, but also conveys non-reward-related signals like positive phasic responses to novel and salient objects. These responses appear to be incompatible with the hypothesis of mapping RL to dopamine. However, given the assumption that phasic dopamine is involved in RL, theoretical simulations can show that these non-reward responses can be reconciled with – and in fact, can emerge from – basic RL principles. This question is examined in more detail in Chapter 4 and in an included article (Laurent, 2008).

1.2 HISTORY OF REWARD-RELATED LEARNING

Historically, behavioral psychology studies of RL have their roots at the beginning of the 20th century in experimental paradigms developed by behavioral scientists such as Ivan Pavlov, Edward Thorndike, and later, B.F. Skinner. Reinforcement learning relates most directly to instrumental, or operant conditioning. Before describing operant conditioning, it is important to describe another form of learning that is thought to be a component of that form of learning: Pavlovian, or classical conditioning.

Classical conditioning experiments make use of the fact that certain stimuli (e.g., food) elicit reflexive responses (e.g., salivation). The experimental procedure involves introducing a conditioning stimulus (e.g., a bell) in the agent's environment prior to the elicitation of the reflexive response. Through repeated presentations of the conditioning stimulus and the elicitation of the

reflexive response, the conditioning stimulus alone is sufficient for the respondent to generate the reflexive response.

Operant, or instrumental conditioning, was investigated by researchers such as Thorndike and Skinner and is more directly related to what is currently termed 'reinforcement learning'. The principal difference between classical and instrumental conditioning is that in instrumental conditioning, a behavioral response is required before the eliciting event (i.e., the 'reward' in reinforcement learning parlance) was provided. The most common example is that of a rat pressing a lever in order to obtain a food pellet.

Many paradigms were developed based on the initial findings of these researchers, with different animal species. Results from these studies have validated the concept of reinforcement learning for a variety of behaviors and stimuli. Through these experiments a number of interesting phenomena have also been discovered. For example, while studying the importance of the temporal relationship between conditioning stimuli and the eliciting stimulus (or behavior), a researcher named Leon Kamin observed that if one stimulus had been conditioned, no learning would occur if a second stimulus was simultaneously presented. This phenomenon, discovered in the late 1960s, was termed *blocking*. This surprising result showed that the temporal order, i.e., the fact that a stimulus preceded reward, was not sufficient to generate learning.

In the following decade, some of these principles of reinforcement were formalized into what became known as the Rescorla-Wagner rule (Rescorla and Wagner, 1972). The Rescorla-Wagner rule is an application of the Widrow & Hoff delta rule to capture the idea that (1) learning is based on the discrepancy between the expected and actual reward, and that (2) once the reward or eliciting stimulus has become fully predicted, learning no longer occurs. This formalization was particularly effective – it adequate captures a number of patterns of behavior reinforcement learning including blocking. However there are a number of behaviors that it does not explain. In particular, it does not explain a particular form of unblocking behavior called downshifting (Holland, 1984). As will be described in detail later (Section 3.6.1 on page 116), downshifting refers to a form of unblocking that is predicted by Rescorla-Wagner to lead to negative (i.e., aversive) values associated with previously blocked conditioning stimuli. One contribution of the present dissertation research will be a proposed extension to the Rescorla-Wagner learning rule that is justified by the experimental

data and overcomes this limitation.

Thus, a number of experimental paradigms have been developed over the past several decades aimed at studying behavioral reinforcement learning in non-human species. These studies have given rise to mathematical descriptions of learning (which have in turn informed RL theory). However, many questions remain about how the brain implements reward-related learning, particularly in humans.

1.3 OVERVIEW

This dissertation is divided into five chapters. The current chapter introduces the theoretical and experimental study of learning, our understanding of the brain, and remaining challenges. Chapter 2 provides background material, including background on accounts of learning, prior neuroscientific findings, and computational models that have been used to study learning. Chapter 3 motivates and describes behavioral, computational and neuroimaging work aimed at understanding how a particular set of brain structures, namely the basal ganglia, mediate learning through reinforcement of physical and cognitive actions that comprise more complex behaviors. The description of the experiment and its results is followed by a discussion of the implications of the findings for learning theories.

Chapter 4 addresses the role of the basal ganglia in novelty and saliency signals, which are qualitatively different from reward-related learning signals described in Chapter 3. The chapter includes an article in which computational simulations of RL agents are used to examine whether novelty/saliency signals can be explained within an RL framework. The simulations reported in that chapter show that novelty and saliency responses can emerge from an RL agent in an environment where its perceptual capabilities are limited. The chapter also discusses how this apparent dual function of the basal ganglia dovetails with the hypothesis of phasic dopamine as a signal of both reward-prediction error and saliency.

Chapter 5 summarizes the work in the dissertation and offers some conclusions.

1.4 RL AS A MODEL OF BASAL GANGLIA FUNCTION

RL originates in machine learning and computer science, but has been more recently applied to psychology and neuroscience. RL is a useful theoretical framework for neuroscience research for a number of reasons. One reason is that it makes quantitative predictions about how histories of reinforcement should influence behavior. To the extent that organisms generally produce behaviors aimed at maximizing their reinforcement, RL should be a good description of the behaviors of those organisms.

The main neuroscientific interest in RL comes from possibility that RL may explain both animal behavior as well as patterns of neural activity in a behaving animal's brain. Much research has been aimed at investigating whether the results of particular computations in RL, namely, the calculation of the difference between expected and received reinforcement, can be mapped to neural substrates. In RL parlance, this quantity is called "reward-prediction error". As will be described below (as well as in Section 2.1), experimental electrophysiological and neuroimaging evidence has begun to emerge which suggests a link between this quantity and the behavior of certain neurotransmitters.

Although reward-prediction error can be expressed mathematically (and will be, below), it is at this point useful to consider its qualitative properties:

- When an unpredicted reward occurs, reward-prediction error is positive
- When a predicted reward occurs, reward-prediction error is zero (non-existent)
- When a predicted reward fails to occur, reward-prediction error is negative
- For temporal-difference learning: When positive or negative reinforcement is predicted, a corresponding reward-prediction error signal should appear at the earliest cue that reliably predicts that reinforcement[1]

Importantly, the magnitude of these signals are hypothesized to be proportional to the quantitative difference between predicted and obtained reinforcement (i.e., they are not "all-or-nothing").

[1] This holds so long as that cue itself is not predictable. Temporal difference learning mechanisms attempt to predict both primary reinforcement as well as secondary reinforcement, which is expressed in terms of predicted cues or states that themselves predict reinforcement. Thus, for example, if a reward cue is entirely predicted by another cue such as an instruction cue, it can be demonstrated that reward-prediction errors occur to the instruction cue rather than to the reward cue in temporal-difference learning models.

Because RL algorithms posit that these numerical predictions are continually (but gradually) updated on each trial of the experiment, RL predicts that there will be continuously-valued signals present during the course of learning. To test these predictions, a mathematical model can be simulated that generates these continuous values, and that analysis entails looking for correlations with the sequence of continuous values as learning proceeds. The most common way to generate regressors for this analysis is to write a computer program to simulate the RL agent that "goes through the motions" that the human participant went through (i.e., observes the same stimuli, performs the same actions, and receives the same reinforcement). This use of RL simulations to make moment-to-moment predictions about latent brain activity is considered a form of "model-based analysis" (see section 2.1.1, page 19.)

The observation of reward-prediction error related signals in the basal ganglia has led to growing interest in the possible connection between the basal ganglia and RL, although current understanding is far from complete. Evidence has started to accumulate suggesting that the basal ganglia are involved in the reinforcement of physical actions. Classically, the majority of findings focus on the nucleus accumbens (the anterior and ventral portion of the striatum where the caudate nucleus and putamen meet) as the locus of reward-prediction error processing, although more recently there has been increased focus on the dorsal striatum (Atallah et al., 2007; Balleine et al., 2007; Delgado et al., 2003; O'Doherty et al., 2004; Volkow et al., 2002).

Despite this interest, it is still not clear how the nucleus accumbens, or the striatum in general, might function to associate rewards with particular states or actions. Answering this question is critical to improving our understanding of the relationship between the basal ganglia and RL. At its core, RL is about allowing reinforcement to guide action selection, and therefore mapping RL to the brain requires a description of how reinforcement is learned and how stored values can be used to determine which actions to execute.

This dissertation proposes that, in contrast to the idea that the nucleus accumbens mediates all reinforcement, that distinct regions of the striatum are involved in reinforcing distinct kinds of actions. This idea, if correct, would provide a potential answer to this quandary, because it would suggest that actions are reinforced by making changes to particular parts of the striatum (e.g., parts of the striatum that are anatomically connected to the neocortical regions controlling those actions).

1.5 CHALLENGES TO RL AS A MODEL OF BASAL GANGLIA FUNCTION

Although RL is a computationally appealing framework for understanding basal ganglia function, there are some important challenges to its acceptance by the broader neuroscience community. A general challenge arises due to the sheer number of RL algorithms that exist – cognitive neuroscientists are especially interested in studying RL algorithms that are biologically-plausible (i.e., that can be implemented using neuronal mechanisms that are known today). However, even within this constraint, the following challenges highlight the need for continuing research into brain mechanisms of RL. Experimental and computational results presented in this dissertation will be discussed with respect to these challenges.

Exploration and exploitation. *Exploration versus exploitation* describes a tradeoff that is relevant in many forms of RL algorithms, including the trajectory sampling algorithm mentioned previously. *Exploitation* refers to the notion that, to maximize reward, a RL agent should select actions that have led to good outcomes in the past. For example, a simple implementation of this idea is the policy that states: select the action leading to the state with the maximum reward. *Exploration*, on the other hand, refers to the notion that, to maximize reward, an RL agent should also sometimes select actions that have not been tried before because they may lead to better outcomes than previously obtained. Thus, exploration and exploitation manifest themselves through effects on the agent's observable behavior and on the internal computations performed.

Because exploration and exploitation have effects on behavior exhibited by RL agents (i.e., the specific sequence of selected actions), as well as on the internal computations performed by these agents (i.e., the specific sequence of reward-prediction updates due to the received outcomes), understanding these parameters in humans is an important challenge to RL as a model of basal ganglia function. There are a number of possible mechanisms that could be singly or jointly involved in implementing these parameters in humans. One possibility is that the hippocampus, given its involvement in exploratory locomotion, may drive exploratory behavior. It is known that the hippocampus replays sequences of sensory and/or motor experiences from the day during sleep, and that these replays trigger activity in the striatum (Pennartz et al., 2004), and that the hippocampus is involved in exploratory behavior (Good and Honey, 1997; Thinus-Blanc et al., 1991). Another

possibility is that activation of the cholinergic system may drive a switch between exploratory and exploitative behaviors, given that acetylcholine appears to be involved in the storage versus retrieval of memories (Doya, 2002). A third possibility is that the exploration and exploitation arise due to activation of the dopamine system itself: such activation by novel stimuli could give rise to exploratory behaviors (Wittmann et al., 2007). Finally, the prefrontal cortex may control exploration in some circumstances, presumably using working memory to keep track of which actions or locations had been previously visited.

Off-line RL. The previous point relates specifically to on-line reinforcement while an agent is behaving during periods of wakefulness. It is also possible that agents can learn from reinforcement in an off-line manner (e.g., during periods of rest). RL algorithms could benefit from exploration within a previously learned model of the environment rather than through direct experience with the environment. Indeed, brain mechanisms such as hippocampal replay of previously experienced event sequences could contribute to between-session learning through hipocampal-striatal projections (Johnson and Redish, 2005; Lansink et al., 2009; Pennartz et al., 2004). Thus, future studies could be performed to quantify the extent to which off-line learning occurs and the brain mechanisms and algorithms that underly such off-line learning.

Neural Network Implementations of RL. An important part of mapping a computational framework to the brain involves understanding how a neural network might be able to implement those computations. One approach is to examine implementations of RL in artificial neural networks. A number of such network implementations of RL have been proposed and a few of these have been successful in a particular domain (e.g., Tesauro, 1994, in backgammon). However, in common practice it has been found that RL in neural networks is either too slow (i.e., requiring a very large number of training trials at small learning rates), or too unstable (i.e., the weights can saturate or oscillate wildly at high learning rates). Thus, an open challenge is to determine what aspects of basal ganglia circuitry might be involved in accelerating and stabilizing RL. One possibility is that the integration of error-correcting gradient descent with RL gradient descent may accelerate RL. Such an approach was used in the *residual algorithm* for RL Baird (1999). The combination of error-correction and reward-prediction gradients is novel, and given the theories suggesting that the cerebellum is involved in error-correcting learning, this algorithm points to potential future re-

search into the contributions of the cerebellum to RL in the basal ganglia. This avenue of research may be fruitful because it has already been demonstrated that there is a direct connection between the cerebellum and basal ganglia (e.g., Hoshi et al., 2005).

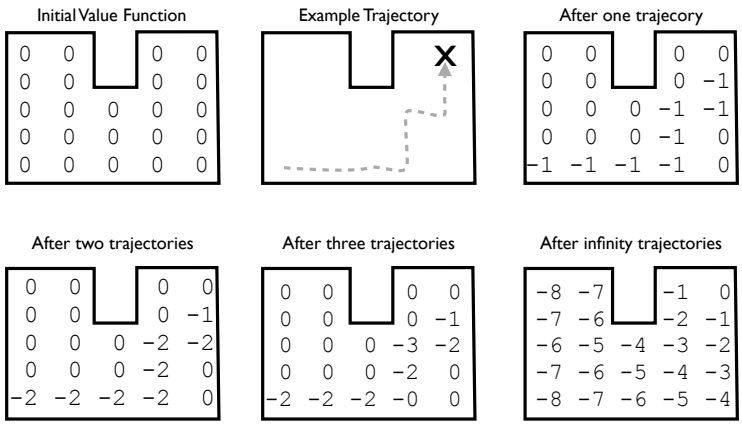

Figure 1.1: **What is Learned in Simple Reinforcement Learning (RL) Problem**. In this problem, there is a goal state in the upper-left of the maze. A "reinforcement function" is defined so that the agent received a -1 (i.e., negative reinforcement) for every movement in the maze. The agent is configured to select actions leading to high rewards. RL equations on the left define the value function as the sum of cumulative expected rewards. Differences between successive reward predictions are used to learn from experience. In this case, experience is obtained through "trajectory sampling" in which a simulated agent begins at some state in the state space, and selects either the best action or, with some small probability, a random exploratory action. An example trajectory is shown in the center-top panel. For simplicity in this example, this exact trajectory will be selected on three subsequent trials. After a sufficient number of trajectories through the state space, a value function like the one displayed at the bottom right has been learned and the RL problem is "solved". The agent can follow an optimal path to the goal state from any point in the state space by simply selecting actions that lead to the next best state (i.e., by selecting the state that results in the last amount of predicted punishment).

2.0 BACKGROUND

2.1 FUNCTIONAL MRI STUDIES

More recently, cognitive neuroscience researchers have attempted to use fMRI to understand the neural mechanisms of reward-related learning in humans. The general approach in this research has been to track trial-to-trial changes in behavior and to use this behavior along with computational models of RL to generate predictions about plasticity processes in the brain that led to the behavioral changes. The trial-to-trial nature of the predictions arises because the variability in behavior depends on the history of stimuli, selected actions and rewards up to that trial. As a result, block design approaches in which there are many repeated trials of a particular type are not typically used in these studies. Instead, paradigms typically use event-related designs. *Event-related designs* are based on the idea that a particular stimulus explains or "accounts for" the trial-to-trial variability in the fMRI signal. Thus, event-related designs are particularly well-suited to investigating the relationship between RL models and brain activity.

There are two kinds of event-related design: slow event-related designs, and fast event-related designs. In slow event-related designs a lengthy time interval (exceeding 6-8 seconds) must be inserted between the events of interest in a study. This is because the hemodynamic response that is measured with fMRI takes that amount of time to return to near baseline levels; without this delay, the hemodynamic responses from sequential trials would overlap and this become problematic for measurement of the responses. Slow event-related design approaches were used in previous model-based studies of reward-related learning, like McClure et al. (2003), O'Doherty et al. (2003) and Tanaka et al. (2004). The advantage of a slow event-related design, namely that the shape of the hemodynamic response can be clearly observed and separated from other fMRI effects, is countered

by its main disadvantage which is a reduction in the number of trials that can be performed during an experiment compared to fast-event related designs, discussed next. Consequently, slow event-related designs may have less statistical power compared to other designs because of fewer trials.

A more recent alternative to slow event-related designs are fast event-related designs, which is what is used in the present dissertation experiment. Fast event-related designs do not require long time intervals between subsequent trials, but they do require that random distribution of time intervals between the events of interest is used. The insertion of random amounts of time between the events of interest is termed *jitter*. With sufficient randomness between events, the signals of interest can be statistically separated from each other using techniques such as deconvolution analysis. This approach was used in more recent studies of RL in the basal ganglia (e.g., Haruno and Kawato, 2006).

Another aspect of fMRI technique is the taking into account of anatomical variability from one participant to the next. Group-level analyses in fMRI requires experimenters to apply a transformation that maps each individuals brain into a common space so that statistics can be done on the fMRI data. The majority of prior experimental approaches have relied on purely affine transformations (e.g., AFNI provides a 12-parameter function involving shearing, scalings and rotating) based on coarse anatomical landmarks such as the anterior and posterior commissures. However, the basal ganglia, especially the dorsal striatum which are of particular interest in RL studies, are highly variable in shape from participant to participant. Testing highly specific hypotheses about localization of function within the basal ganglia is likely to benefit from a finer mapping of anatomical structures between participants. The present study used a relatively new high-dimensional nonlinear warping algorithm named ANTS that has been evaluated as the best performing method among 14 tools in a recent review in Neuroimage (Klein et al., 2009). In Appendix A, a visual comparison of the automated output of ANTS compared to the automated low-dimensional co-registration procedure used by AFNI.

Prior studies have only made very general hypotheses about the involvement of different regions in the striatum for the computation of RL quantities such as reward-prediction error, e.g., "dorsal" versus "ventral" striatum. These hypotheses are generally based on findings in rats (Atallah et al., 2007) or in drug abusers (Volkow et al., 2002) suggesting that the dorsal striatum is

involved in less flexible behaviors compared to the ventral striatum. A recent review implicates two regions of the dorsal striatum in decision-making: the caudate and dorsomedial striatum are implicated in flexible, goal-directed decision-making, whereas the dorsolateral striatum is implicated in less flexible, automatic decision-making (Balleine et al., 2007).

Prior investigators have focused on single types of actions. Generally, all studies of instrumental conditioning have involved button presses in human studies, or saccadic eye movements in non-human primates. The use of this limited set of actions is likely to limit the range of data observed in those studies. In contrast, in the present study, participants performed different kinds of actions to maximize reward. The use of different actions allowed testing very specific anatomical predictions about the regions of the striatum that might be involved in the reinforcement learning of those actions.

Finally, most prior studies of RL in the basal ganglia have assumed that the shape of the hemodynamic response is the same for both positive and negative reward-prediction error, and that this shape is simply modulated by the sign and degree of reward-prediction error (however, the more recent study by Haruno and Kawato (2006) does investigate the question of shape more closely). Studies focused on studying striatal responses to positive and negative outcome (which are closely related to positive and negative reward-prediction error) have shown that these signals do in fact have different shapes, and the modulation of these shapes is not linear but is rather dependent on whether outcome is positive or negative (Delgado et al., 2003, 2000). Thus in the present study, positive and negative reward-prediction errors were allowed to have separate shapes, dictated by the data using deconvolution analysis.

2.1.1 MODEL-BASED ANALYSIS

Model-based analysis is the idea that one can use computational models to make predictions about latent variables (for example, brain activity) based on internal variables in a computational model (e.g., Corrado and Doya, 2007; Tanaka et al., 2004). Model-based analysis was important in testing the main hypotheses of this dissertation about reward-prediction error computations. The way in which model-based analysis was used in the dissertation is illustrated in Figure 2.1 on page 26. As shown in that figure, the observed sequence of human behavior was used to fit the performance

of an RL agent simulation. Then, the modeled sequence of reward-prediction errors generated internally by the RL agent were used as regressors hypothesized to correlate with the brain activity of participants performing the task.

This dissertation makes use of event-related model-based analysis to test the hypothesis that there are signals that reflect trial-to-trial changes in reward-prediction error that can be localized (1) to the striatum, and, (2) to particular regions of the striatum depending on the type of action that is being reinforced.

2.2 DOPAMINE

Although the focus of this dissertation is on the involvement of the basal ganglia in RL, a potentially important aspect of the basal ganglia are two of its structures known as the ventral tegmental area and the substantia nigra pars compacta. These structures are of particular interest because they release the neurotransmitter *dopamine*, which is a catecholamine neurotransmitter. These cell bodies release dopamine both within the basal ganglia (e.g., the striatum and other basal ganglia nuclei) as well as through the rest of the brain, including extensive projections to the prefrontal cortex.

Dopamine has long been thought to be involved in reward-related learning, and more recently has been a candidate for the reward-prediction error signal in RL. Indeed, probably one of the earliest neuroscientific studies of reinforcement learning was a study performed by James Olds and Peter Milner in which they implanted electrodes into the medial forebrain bundle (which targets the ventral tegmental area) and ventral tegmental area itself. Olds and Milner found that stimulating these areas was a highly effective stimulus for reinforcing behavior. Indeed, rats who are allowed to press a lever to stimulate these regions continue to do so until exhausted. Research thus suggests that natural reinforcers (as well as acquired reinforcers) act at least in part through activation of the dopamine system.

Although dopamine may be involved in reinforcement learning, the fMRI methodology used in this dissertation does not allow direct inferences about dopamine activation or presence. There are, however, some reasons to suspect that dopamine may be related to the fMRI BOLD signal.

First, the metabolic activity that is the neurophysiological basis of the fMRI signal is known to more closely reflect the action of afferent synaptic input into a given brain region rather than the efferent spiking output of that region (Logothetis et al., 2001). Because the striatum is a target of dopamine-releasing neurons of the subtantia nigra, dopamine acts on striatal synapses, the striatum is expected to show BOLD responses that are hypothesized to be due to dopamine. Second, a recent study concluded that, at least in the ventral parts of the striatum (e.g., the nucleus accumbens), BOLD activation does indeed reflect the action of dopamine on postsynaptic D1 receptors there Knutson and Gibbs (2007).

However, it is less clear whether focalized BOLD responses in the striatum can be attributed to dopamine (although such activations been observed in prior studies, (e.g., Hare et al., 2008; McClure et al., 2003; O'Doherty et al., 2003; Tricomi et al., 2009)). Indeed, dopamine neurons project to the entire striatum and are thought to fire homogeneously as a group. This suggests that dopamine is released to the entire striatum as a **global** reinforcement signal. Thus, one might predict that striatal BOLD activation would not reliably show regional differences due to dopamine. One possibility however is that BOLD activity in the striatum is not solely due to dopamine, but rather the combination of dopamine and some other source of neural activity. For example, cortico-striatal projections are expected to be active in an action-dependent manner; this activity might then be permissive for dopamine to be preferentially released by synapses entering those portions of the striatum. Similarly, dopamine release might be inhibited in other striatal regions whose cortico-striatal afferents are relatively inactive. This could further depend on the 'up' or 'down' excitability states of the striatal medium spiny neurons, which may be in part dependent on the activation of cortico-striatal projections and on the task being performed.

Another possible explanation for attributing differentially-localized BOLD in the striatum to dopamine is that this BOLD response may reflect synaptic plasticity processes (i.e., LTP and/or LTD) due to the local convergence of neocortical and dopamine activity. That is, the BOLD response might be more significant in places where dopamine interacts with afferent neocortical activity because of metabolic increases involved in plasticity (e.g., Schultz, 1998; Suri and Schultz, 1998).

2.3 COMPUTATIONAL MODELS OF BASAL GANGLIA FUNCTION

Developing computational models of the basal ganglia has been challenging. One reason for this is their complex patterns of connectivity: the basal ganglia contain multiple pathways and multiple levels of inhibitory connections. Another source of difficulty is that there are many aspects of basal ganglia organization which have as-of-yet undetermined, or still widely-debated, significance.

Most importantly, however, there is a lack of known constraints on the possible localization of function in the basal ganglia. For example, it is not clear whether all reinforcement-related activations take place in the ventral striatum (i.e., the nucleus accumbens) or in the dorsal striatum. Knowing about the localization of reinforcement-related activation and computations will help build models that are more likely to be correct.

A number of conceptual and computational models of basal ganglia function have been proposed and are discussed briefly. Although this list is not exhaustive, the models described below are widely discussed and capture the range of models that are frequently described in the literature.

Direct-and-Indirect. The earliest models of basal ganglia function were based on the notion of direct and indirect pathways through the basal ganglia. These models were based on observations of hypo- and hyperkinetic dysfunction observed in basal ganglia disorders (Albin et al., 1989). However in a recent neuroanatomical study, basal ganglia researchers have found evidence against the direct and indirect pathways (Lévesque and Parent, 2005), suggesting that the use of this division in the development of further models should be re-evaluated.

Striatum as Lateral Inhibitory Network for Selection. One of the earliest proposals for the computations performed in the basal ganglia revolved around the striatum as a lateral inhibitory network (Wickens, 1993). Computationally, the idea is that the striatum performs a "winners-takes-all" operation on afferent input, thereby only allowing the strongest cortico-striatal input to pass through the circuit. This proposal arose from early observations that the majority of the synapses in the striatum, in particular synapses formed between the main projection neurons of the striatum (i.e., the *medium spiny neurons*) were GABA-releasing (Bolam et al., 2000; Tepper and Bolam, 2004). This function would be particularly important for cases in which striatal input converged from multiple neocortical sources because presumably the striatum could be involved in selecting

one of those sources over another. However, it has recently been suggested that the effect of these inhibitory synapses is too weak to effectively implement a winners-take-all operation (Tepper et al., 2004).

Dimensionality Reduction. Despite the anatomical evidence for lateral inhibitory connections in the striatum, there is little electrophysiological evidence for connectivity that is functional. This idea, combined with the reduction in the numbers of neurons as one moves from neocortex, to the striatum, to the globus pallidus, has been used to develop models of basal ganglia function where their function are assumed to be to reduce the dimensionality of neocortical information (Bar-Gad et al., 2000, 2003). The reduction of dimensionality is hypothesized to be critical to the learning and control in brain networks because without such a reduction in dimensionality, learning based on the patterns would suffer from the "curse of dimensionality". This means, for example, in the case of learning to map a state to an action, there would need to be an "unrealistically large" number of training examples because each example would be very different in a high dimensional space. The reduction-of-dimensionality model was also proposed by a separate group of researchers (Joel et al., 2002). Although the idea of dimensionality-reduction is a computationally powerful, its localization to the basal ganglia has not been borne out by significant experimental confirmation to-date.

Action Selection in Recurrent Loops. In a model developed by Leblois et al. (2006), recurrent activation through cortico-basal ganglia loops is proposed as a mechanism through which weak neocortical activation patterns are amplified and thereby selected. This model demonstrated how one of two weakly activated action representations in the neocortex could compete within the basal ganglia circuitry until one emerged as a clear winner. Although this model did not implement a RL mechanism (e.g., implemented by a phasic dopamine signal), future variants of this model will do so (T. Boraud, personal communication). Given that such a mechanism has been shown to work in an unstructured recurrent neural network (Izhikevich, 2007), it should be particularly successful in this more structured model.

Actor-Critic Models. Actor-critic models are based on the distinction between learning reward predictions and selecting actions based on them (Joel et al., 2002). Some of the earliest efforts to map RL to the basal ganglia were actor-critic models. This division between actor and critic has

been proposed to map to the computations of the ventral and dorsal striatum, respectively (e.g., O'Doherty et al., 2004). However, it is not clear why the separation between actor and critic is necessary. Many RL models do not posit a separation between actor and critic: The same circuit involved in storing reward predictions may indeed be the same circuit that allows those reward predictions to select previously reinforced actions with increased probability (i.e., reinforcement and action selection could be mediated by the same cortico-striatal-thalamocortical loops, by introducting reinforcement into models like the Leblois et al. (2006) model).

ACT-R. ACT-R Anderson et al. (2004) is discussed more thoroughly in Chapter 3), but is mentioned here for completeness. ACT-R is a model of cognitive control that makes a set of very specific assumptions about basal ganglia function. Distinct roles attributed to each nucleus in the basal ganglia: striatum as pattern recognizer, pallidum as action selection, and thalamus as execution. This model overlaps to some extent with the Leblois et al. (2006) model because it presumes that the pallidum is involved in the selection of action. Despite the lack of fMRI evidence consistent with ACT-R, the model does a very good job of fitting behavioral data.

'Gating' Sensory Input into Prefrontal Cortex. Although not a model of basal ganglia function per se, O'Reilly and colleagues (Hazy et al., 2006) have proposed a model in which the basal ganglia is a key structure. According to the model, the prefrontal cortex originates cognitive control and acts as a "central executive", having properties like working memory and high levels of connectivity that make it especially well-suited to maintaining goals and controlling arbitrary actions. However, these same properties are what make the prefrontal cortex highly likely to be distracted by conflicting or task-irrelevant information. They therefore propose that the basal ganglia act as a "gate" that prevents sensory information from entering into the prefrontal cortex until such information is necessary. This gate is trained through an internal algorithm, namely, temporal-difference learning, to learn to predict when cues are likely to lead to reward. Because these cues predict reward, their presence reflects opportune moments to open the "gate" allowing new sensory information and/or task goals to influence the prefrontal cortical representations. There are two potential mechanisms by which the basal ganglia could have this effect: in one mechanism, the basal ganglia themselves are able to remove tonic inhibition from the prefrontal cortex, thereby allowing additional activation to result in rapid updated. In the other mechanism, phasic dopamine

– which is released from the substantia nigra pars compacta and ventral tegmental areas, structures under the control of the basal ganglia and prefrontal cortex – stimulates D2 dopamine receptors in the prefrontal cortex and allows an update to occur. (O'Reilly, 2006)

2.4 CONCLUSION

This chapter has provided a overview of the history of reward-related learning, described experimental methodology and its limitations, and has listed a number of theoretical accounts that have been previously advanced for understanding the function of the basal ganglia. The current state of affairs suggests that the data are not sufficiently constraining to strongly prefer particular theoretical accounts over others.

The main goal of this dissertation is to seek additional evidence that might provide additional constraints to show which of these models are more likely to be correct. It is the hope that the experimental results will elucidate the nature of basal ganglia computations and build on these existing models to generate a conceptual framework that has a significant amount of explanatory power.

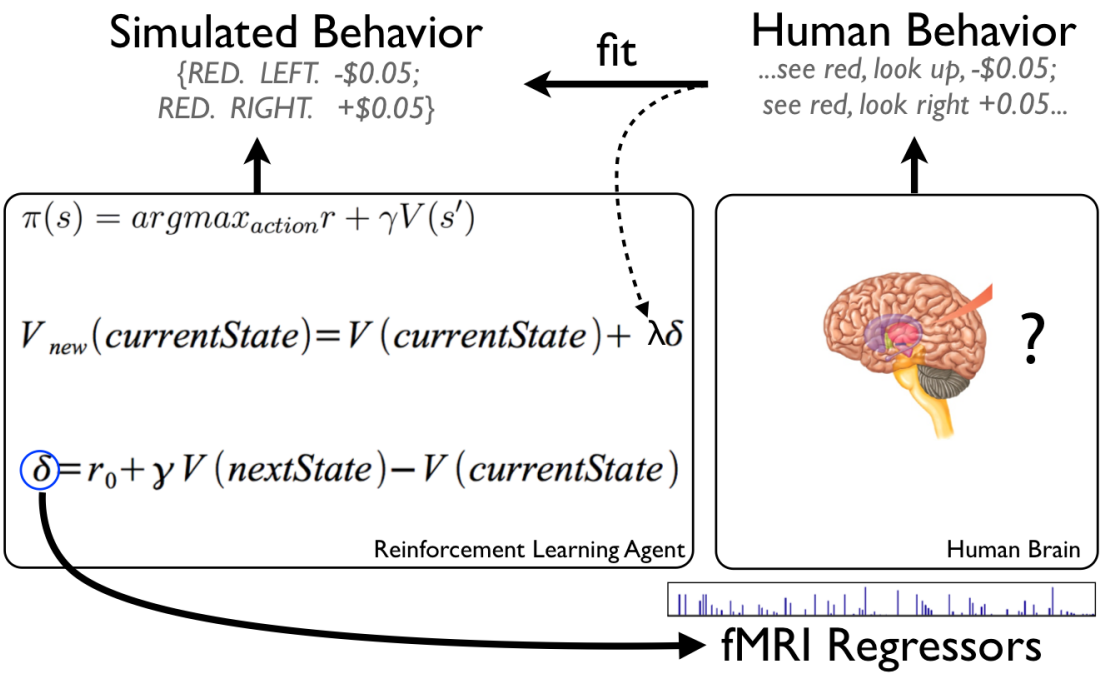

Figure 2.1: **An illustration of the model-based analysis approach used to analyze fMRI data in this dissertation.** This method allows numerical predictions from the internal variables in computational models to be used to make predictions about neuronal activity in human participants. The starting point is human behavior in the upper-right corner. We implement a RL agent that generates simulated behavior. We are able to maximize the fit of the RL agents' behavior to the human participants' behavior by changing the learning rate parameter (λ, in the second equation). Once this fit is maximized a regressor for action-specific reward prediction error can be generated from the interval variable in the model for reward prediction error (δ, in the third equation). See the text on page 19 for more discussion.

3.0 REINFORCEMENT LEARNING OF PHYSICAL AND COGNITIVE ACTIONS

3.1 INTRODUCTION

Are the basal ganglia involved in the reinforcement learning of physical and cognitive actions? Prior evidence from primate physiology and human fMRI studies suggest that there is some involvement of the basal ganglia, however these studies have left many unanswered questions. No clear picture has yet emerged about the nature of computations performed in the basal ganglia, nor about the localization of these computations. Do the basal ganglia compute reward-prediction error for both physical and cognitive actions? Are reward-prediction errors all computed in the same location in the nucleus accumbens, or are they in different locations in the dorsal striatum? If reward-prediction error computations for different actions take place in the basal ganglia, do these computations overlap spatially with regions involved in the control of those actions?

Data from neuroanatomical, electrophysiological and neuroimaging experiments point to possible circuitry that may support computations required by RL. Neuroanatomical studies have shown that there are parallel pathways projecting to the basal ganglia from different regions of neocortex, including skeletomotor, oculomotor and cognitive regions. These parallel pathways could provide a mechanism for the basal ganglia to act on action-specific representations in neocortex. In these circuits the striatum is particularly important because it is considered the main input nucleus of the basal ganglia, receiving substantial innervation from nearly every region of the cerebral cortex. The striatum is also heavily innervated by dopamine-releasing neurons of the subtantia nigra pars compacta, and the ventral tegmental area. Although the inputs from neocortex carry perceptual, motor and goal-related information, the dopamine neurons appear to transmit information related to reward-prediction error.

However, it remains to be shown whether this convergence of this neocortical and dopaminergic input could lead to action-specific RL computations. The hypothesis about the localization of action-specific reward-prediction error computations to different loops is thus based on the idea that this circuitry is involved in allowing the basal ganglia to apply the same type of computation to all afferent activity, i.e., the computation of reward-prediction errors for neocortical action representations.

In this chapter, a behavioral and neuroimaging experimental paradigm is developed that allows direct comparisons of learning rates, asymptotic performance levels, and neural activity while three different classes of actions are positively and negatively reinforced. The paradigm allows both physical and cognitive actions to be studied. For this particular experiment, the effects of reinforcement on three types of actions will be studied: skeletomotor actions (hand movements), oculomotor actions (eye movements), and attention-shift actions (covert shifts of visual-spatial attention). Model-based analysis of fMRI data was used to test the hypothesis that reward-prediction errors are computed in spatially distinct locations that correspond to the regions involved in the physical and cognitive actions that are being reinforced.

Prior to the fMRI study, a behavior pilot study with a larger number of participants was carried out to obtain behavioral measurements of RL, and to determine the parameters of the study. The behavioral data allowed a close examination of the behavioral effects of reinforcement. In particular, the data provided answers to a number of additional questions: Can different actions be learned to the same asymptotic levels and at the same rates? If so, this suggests that a similar learning algorithm is involved in learning the actions. Can the different actions be fit equally well with RL agents? If so, this provides evidence in support of the idea that RL – i.e., selecting actions based on learned reward predictions – may be that common algorithm.

3.2 ARTICLE CONTENT

The next sections contain the text of an article (to be submitted) describing the main experimental results of this dissertation project.

3.3 EMPIRICAL AND THEORETICAL EVIDENCE FOR GLOBAL REINFORCEMENT OF LOCAL CONTROL

An important challenge for the nervous system is to solve the problem of selecting, controlling, and coordinating activity patterns in the brain to generate intelligent, adaptive behaviors leading to positive outcomes for the organism. We propose that a key part of the brain's solution to this problem involves an implementation of Reinforcement Learning (RL) that uses the well-documented, distributed pattern of loop connectivity between neocortex and the basal ganglia. In particular, we propose that the brain (1) delegates control of actions to local areas of cerebral cortex, (2) selects particular actions by using recurrent cortico-basal ganglia loop activation dynamics, and (3) uses global phasic dopamine as a reinforcement signal to modulate these recurrent dynamics to maximize future reward for the organism as a whole. In this framework, the basal ganglia provide a general-purpose mechanism that allows global dopamine to dynamically and selectively reinforce patterns of neocortical activity for physical, as well as cognitive, actions through the updating of reward predictions within cortico-basal ganglia loops. In support of this idea, we provide behavioral, fMRI, and computational evidence. In a behavioral study, we examined participants' performance as they were rewarded for performing hand movements, eye movements, and covert shifts of visual attention in a single task. We found that participants could learn the behaviors equally well, suggesting that a similar algorithm may be used when learning different actions. Furthermore, we found that RL simulations were able to fit the participants' data equally well, suggesting a common learning algorithm based on reward-prediction error for these motor and cognitive (i.e., non-motoric) actions. In an fMRI study, we examined BOLD fMRI signals in the basal ganglia as participants were reinforced for these same three types of action. We found that, depending on the type of action

that is being reinforced, reward-prediction error regressors derived from RL agents correlated with BOLD activity in distinct regions of the striatum, the same general regions to which neocortical representations for those actions are known to project. Specifically, we found that reward-prediction errors for right-hand movements were computed in the left putamen, reward-prediction errors for eye movements were computed in the dorsal striatum and anterior caudate, and reward-prediction errors for covert attention shifts were computed in the anterior caudate. These results suggest that the computation of reward-prediction errors for specific actions is one of the functions implemented by cortico-basal ganglia loops (Alexander, DeLong, & Strick, 1986), and that individual actions that may be composed into a more complex behavior can be reinforced independently. Importantly, these results establish that reward-prediction errors are computed not just for physical actions (i.e., hand movements and eye movements), but for cognitive "actions" as well (i.e., covert attention shifts). In contrast to the traditional view of their primary involvement in real-time motor control, these results highlight a role for the basal ganglia in rapid learning and control of non-motor actions, providing a functional explanation for prior anatomical findings that the basal ganglia are connected to cognitive areas as well as motor areas. These results also demonstrate how a global reinforcement signal, such as that hypothesized to be mediated by dopamine, can be used to learn actions controlled by specific neocortical regions. The separate computation of reward-prediction errors for different neocortical representations helps explain previously problematic results with blocking and downshifting of values (Holland, 1984). Finally, we review some RL simulations that demonstrate (1) how the reinforcement of independent actions can lead to the apparent coordination of those actions into "intelligent" behaviors, and (2) how the mapping of reward-prediction errors to phasic dopamine helps explain its involvement in apparently non-reward-related activity like novelty and saliency responses. We conclude that the global reinforcement of local delegated control is a viable mechanism available to the brain to the solve problem of learning to control and coordinate physical and cognitive actions into more complex behaviors, much like bonuses to all employees of a corporation can result in improved, coordinated performance.

3.3.1 INTRODUCTION

Humans are able to control a wide variety of actions and organize them into intelligent, adaptive behaviors. Our ability to effectively control such behaviors is quite remarkable when one considers how the size of the human brain, estimated at 100 billion neurons, imposes severe constraints on connectivity between its parts: only a small subset of the possible connections, through which control can be exerted, exist. For example, it is estimated that a given neuron only receives connections from an average of 10,000 others (i.e., .000000001% of the neurons in the brain). This limitation on connectivity follows from a wiring-space constraint and is likely to be, along with energy efficiency, a key factor in brain organization (e.g., Chklovskii, 2004). Further, neurons must be sufficiently depolarized by afferent activity before they fire an action potential and therefore generate output at their synaptic terminals. The controlled activation of particular sets of neurons therefore requires either a significant convergence of connectivity, or a temporal pattern that is able to trigger and/or maintain the desired activation patterns. This combination of sparse connectivity and thresholded firing imposes severe limits on how the brain can control itself. Nevertheless, we do exhibit the ability to efficiently control a large variety of non-instinctual actions and organize them into adaptive and intelligent behaviors – but how?

The answer that we consider in this thesis is that the brain's solution for achieving efficient control of itself is based upon learning through reinforcement. We present a theoretical framework based on Reinforcement Learning (RL) that explains how the brain is able to control and coordinate neocortical activity patterns that lead to physical and/or cognitive actions. At a gross level, the framework states that the brain does this by delegating control of actions to local, specialized regions of the neocortex, and by using a global, non-specific reinforcement signal to adjust the likelihood of performing those actions by modulating the strengths of (mostly) closed loops feeding back to those neocortical action representations. This "globally-reinforced locally-controlled" mechanism allows the appropriate actions to be selected in accordance with their likelihood of leading to positive outcomes, and helps side-step the need for strong and highly specific cortico-cortical connections for exerting control. Also, because the control of the actions is delegated rather than controlled by a central executive, we propose in the General Discussion that it mitigates the homunculus problem that frequently arises in cognitive control theories. At the core of this framework

is the idea that selection of behaviors, and the individual actions that comprise them, are guided by a drive to maximize the overall reward received by the organism. This idea is captured elegantly by RL, which refers generally to computational algorithms wherein actions selected by an organism (the 'agent') are guided by learned anticipated reward, i.e., reward predictions. In the most widely-used implementations of RL, as well as in the proposed framework, these reward predictions are learned through reward-prediction errors (i.e., differences between expected and received reward). Reward-prediction errors have been demonstrated to be an important component of the signal broadcast throughout the brain by phasic dopamine Schultz et al. (1997). In agreement with prior proposals, the global reinforcement signal in this framework is hypothesized to be conveyed by phasic changes in dopamine. It can be demonstrated through simulations that RL-based frameworks are computationally powerful, because the global reinforcement of independent actions can give rise to the coordination of actions that might be represented in distant parts of the brain (e.g., the coordination of various perceptual, cognitive, and motor processes in reading Laurent and Reichle, submitted; Reichle and Laurent, 2006).

Although prior research shows that dopamine is likely to be involved in RL because it conveys reward-prediction error signals, it remains to be shown whether such a globally-broadcast reward-prediction error dopamine signal can in fact reinforce specific actions on a neural systems level. The global nature of the dopamine signal raises questions about whether and where action-specific RL computations are computed and stored. There is evidence of dopamine-dependent plasticity for corticostriatal synapses (Reynolds and Wickens, 2002), and that plasticity may be selective for active synapses (e.g., Bamford et al., 2004). Another finding partially addresses these questions at a coarser systems level: dopamine synthesis capacity in specific regions of the striatum has been shown to differentially correlate to performance of actions in motor versus cognitive domains (Landau et al., 2009). However, the relation of RL signals to the specific actions that are being reinforced, as well as the circuits that compute reward-prediction errors store the resulting reward predictions for different actions, has so far remained elusive. These questions are a motivating factor for the design of the experiment described below. Although recent research has suggested the possibility that reward-prediction error computations used by RL are performed in the basal ganglia, many questions remain. First, it is not certain how these computations map onto the basal ganglia circuitry. Although the cellular content of the striatum is relatively uniform and the overall anatomy

of the basal ganglia is organized into parallel pathways, it is unclear whether the basal ganglia are computationally homogenous, i.e., whether they apply the same computation to all inputs. This is an important point because if, for example, the signatures of RL computations are shown to occur in all parts of the basal ganglia, it is likely that the parallel structure of the basal ganglia is important in performing those computations. On the other hand, if RL computations are only performed in a restricted region of the basal ganglia, then it is more likely that some local property, such as afferent information from a particular cortical region, is important for those computations. In particular, some researchers have proposed that the basal ganglia perform different computations depending on whether inputs arrive in the ventral or dorsal portions of the striatum (for a review, see Wickens et al., 2007).

Testing the idea that the basal ganglia are involved in particular computations requires knowing what information those computations are operating on, that is, what information the basal ganglia receive. The basal ganglia have been traditionally viewed as a motor control structure, not only because basal ganglia disorders like Parkinson's and Huntington's diseases disrupt movements, but also because of the extensive anatomical and functional connections between the basal ganglia and frontal cortex (Gerardin et al., 2003; Lehéricy et al., 2004b). For example, the putamen has been established as a striatal zone for the control of skeletal muscles, such as those used in hand movements (Crutcher and DeLong, 1984). Other research has focused on dorsal-most parts of the caudate and putamen, as involved in the control of saccadic eye movements (Brown et al., 2006, 2007; Calzavara et al., 2007; Gerardin et al., 2003; Sweeney et al., 1996). However, neuroscientists have increasingly acknowledged the fact that the basal ganglia have connections with non-motor neocortical areas, suggesting that they are involved in cognitive control. In particular, fMRI studies have shown that the head of the caudate is more active for covert attention shifts compared to saccades (Perry and Zeki, 2000), and the anterior parts of the striatum are involved in the control and modulation of attention (Corbetta et al., 1991). There also exists anatomical evidence in support of projections to the anterior striatum from parts of the parietal lobe that are hypothesized to be involved in attention(Cavada and Goldman-Rakic, 1991; Leichnetz, 2001). Indeed, methods such as autoradiographic tracing, trans-neuronal virus transport, and diffusion tensor imaging have all been used by neuroatomists to demonstrate connections between the striatum and a variety of frontal, parietal and even temporal regions involved in non-motoric cognition and perception

tasks (Alexander et al., 1986; Lehéricy et al., 2004b; Middleton and Strick, 1996; Webster et al., 1993; Yeterian and Pandya, 1995). Interestingly, the outputs of the basal ganglia project back to regions of neocortex in a very orderly fashion, suggesting a local, closed-loop architecture repeated throughout the involved regions (Middleton and Strick, 2000). Together, these findings suggest the hypothesis that if the basal ganglia do perform a particular kind of computation, that computation may be one that is relatively local and likely identical with respect to each part of neocortex. Thus, one experimental approach would be to investigate whether the involvement of the basal ganglia is similar for different physical actions, such as hand movements and eye movements. And because of their connection to cognitive areas, the same question can be asked about the involvement of the basal ganglia in cognitive "actions" as well (e.g., shifts of covert attention).

From a physiological perspective, the relatively rapid time-course of changes in dopamine (and striatal activity) run counter to the traditional idea that the basal ganglia are involved primarily in the slow acquisition of procedural ("skill") memory. It is now apparent that the basal ganglia are involved in learning during relatively short time-courses (for a review, see Graybiel, 2005). Electrophysiological recordings from dopamine-releasing neurons indicate that their phasic activity during simple reward-related tasks is not simply correlated with reward but rather reflects a "reward-prediction error" signal that updates from trial to trial within a single experimental session, representing a difference between expected and received reward (Schultz et al., 1997). Functional MRI studies have mirrored these findings in the human striatum, which is heavily targeted by the dopamine neurons and have established connections to reinforcement learning (e.g., Haruno and Kawato, 2006; McClure et al., 2003; O'Doherty et al., 2003; Tanaka et al., 2004). In these studies which examined both Pavlovian and instrumental conditioning, running values for reward expectations for each stimulus viewed by and/or action performed by the human participants are updated in a computational RL model. On each trial, the difference between simulated reward expectation and the actual reward on the trial are found to correlate to regions in the ventral and/or dorsal striatum. However, the different paradigms used in these studies found these correlated activations in different regions of the striatum. From the data it is therefore unclear whether the striatum is involved in reward-prediction error in a general sense with arbitrary regions involved in different tasks, or whether specific parts of the striatum and/or neocortex are uniquely involved in reward-prediction error in a task-dependent manner.

From the computational perspective, recent RL simulations of real-time behaviors have demonstrated that a relatively simple learning mechanism based on global reward-prediction errors – importantly, without any centralized control– is able to develop adaptive, complex patterns in which distinct actions appear to be coordinated. In these simulations, reward-prediction error has been used to control and coordinate simulated motor and cognitive actions in real-time behaviors. Cognitive science research has successfully used computer simulations of RL agents to understand how motor processes, as well as perceptual and cognitive processes, may be learned and coordinated to enable efficient, "intelligent" behavior (Ballard et al., 1997; Botvinick et al., 2008; Laurent, 2008; Laurent and Reichle, submitted; Reichle and Laurent, 2006). Interestingly, the majority of RL algorithms work by using reward-prediction error, i.e., the same quantity that is hypothesized to be conveyed by dopamine, to modify the likelihood of selecting an action or the values assigned to states reachable via different actions. Given that RL is a particularly powerful algorithm for explaining the emergence of coordinated "intelligent" behavior, the apparent convergence between RL computations and moment-to-moment activity in the basal ganglia warrants further investigation. In particular, an understanding of the computation and use of RL quantities in the brain could potentially generate important neuroscientific knowledge about how reinforcement leads to intelligent behavior in humans.

Taken together, the above findings challenge the traditional view that the predominant role of the basal ganglia is the slow, procedural learning of motor skills and the real-time control of motor skills. Rather, they raise the possibility that the basal ganglia are involved in the rapid, reward-related learning of both motor and non-motor actions. Although this proposed view of basal ganglia function is appealing, a significant piece of the puzzle is missing before this view can be taken seriously: Are the basal ganglia in fact involved in on-line RL computations for arbitrary physical and cognitive actions? Can it be shown that these computations occur in relation to the known anatomical projections from the neocortex into the basal ganglia? Specifically, are these computations localized to distinct parts of the striatum that receive anatomical projections from the regions whose actions they are reinforcing?

The present study was run to answer these questions. The results suggest that a function of the basal ganglia is to perform RL computations for any physical and cognitive "actions" that are

represented as patterns of activity in the cerebral cortex. The spatial localization of the results suggest that neocortical activity patterns are controlled and selected according to learned reward predictions by recurrent loops through the basal ganglia. The results presented in this thesis suggest that this reinforcement learning is one solution to the problem of learning to perform and coordinate very diverse actions in order to maximize the total reinforcement received by the organism.

3.3.2 HYPOTHESES AND PREDICTIONS

One of the main contributions of this thesis is to demonstrate that there is a mapping between neuroanatomical circuits and the computations proposed by RL, more specifically, a mapping between the cortico-basal ganglia loops and the computation and storage of reward-prediction errors associated with specific actions.

The proposal is that most complex behaviors consist of one or more actions that are selected and combined into sequences based on the changes to recurrent loops through the basal ganglia with cortical regions where the individual actions are represented and controlled. Learning through reinforcement involves the strengthening and weakening of cortico-striatal inputs to these loops in proportion to a global reinforcement signal – that is, reward-prediction error – broadcast by phasic dopamine. The result of this modification in cortico-basal ganglia recurrent loop strength is that neocortical action patterns that are active prior to reward will be more easily activated in the future, and those active prior to punishment will be less easily activated in the future. Because this type of loop circuit exists with connections to most of neocortex, this computation enables the independent reinforcement of any action that is represented in neocortex and the capacity to combine them into arbitrarily complex sequences of actions. A computational simulation based on the anatomy of the basal ganglia has provided a clear demonstration of the use of recurrent activation dynamics to select specific actions in the (non-learning) model proposed by (Leblois et al., 2006). These modelers demonstrated how a mostly-closed recurrent loop architecture from the cortex through the basal ganglia could effectively control and amplify neocortical action representations, even if those action representations overlapped. The addition of learning to that model, using a reward-prediction error signal, could serve as a preliminary implementation of the framework being proposed here.

The other part of the implementation is the modulation of the global reward-prediction er-

ror signal itself. That is, although the main hypothesized role of the cortico-striatal plasticity is proposed to be the recurrent reactivation of neocortical patterns depending on which one leads to reward, a further role of the cortico-striatal plasticity could be to participate in the computation of reward-prediction error. Specifically, once a reward prediction has been successfully learned, the dopamine signal should be inhibited. As hypothesized by several others, this could be achieved, at least in part, through the inhibitory connections that exist from the striatum back to the dopamine neurons (Doya, 1999; Schultz, 1998). However, it remains to be shown whether populations of striatal neurons are activated by the neocortical action representation in proportion to actual reward predicted for those actions, which would thereby allow them to cancel out the activity of the dopamine neurons.

This framework leads to a number of specific predictions. Those that are explicitly tested by the present experiment are listed here, and additional untested predictions of the framework are discussed in the General Discussion. The predictions are: (1) Because dopamine is a globally broadcast signal, different physical actions, such as hand movements and saccadic eye movements, should be influenced by reward in the same way and to the same extent, i.e., similar asymptotic performance levels and learning rates. This will be particularly true if the actions themselves are sufficiently familiar or well established in their neocortical representations, and do not require finer neocortical reorganization. In addition, RL agents trained on the same sequence of stimulus, action, and outcome should equally well fit participant behavior regardless of the type of action being reinforced. (2) Because the basal ganglia are connected with motor as well as cognitive areas, cognitive "actions" with no motoric component, such as covert shifts of attention, should also be rewarded in the same way as overt physical actions. (3) Correlates of the computations performed in the cortico-basal ganglia loops should be the same for the different actions making up a behavior (i.e., whatever correlates are found should be found for all actions because of similar histories of action and reward.) (4) Correlates of the computations in the cortico-basal ganglia loops should be based on the difference between learned reward and the reward received on each trial – i.e., the reward-prediction error that is utilized by most RL algorithms. (5) The reward-prediction error computations should be localized to the intersection of dopamine terminals and the terminals of cortical afferents representing the action(s) involved in the behavior, i.e., regions in the striatum targeted by cortical regions controlling those actions.

The first two predictions are evaluated by the behavioral experiments presented below by testing the hypothesis that learning rates and asymptotic performance are identical when different types of physical and cognitive action are reinforced (i.e., hand movements, eye movements, and convert shifts of attention). The results suggest two main conclusions: First, although these behaviors are controlled by different neocortical regions, they can be learned equally well within in single experimental session. Second, both physical actions and cognitive "actions" appear to be learned through a common underlying learning mechanism. In order to determine this learning mechanism, a finer test of these predictions is then carried out by examining whether the behavior of simulated RL agents fits the human behavior equally well when different actions are reinforced. The results of this analysis suggest that RL may be the common learning mechanism that underlies the reinforcement of distinct types of action.

The final predictions are evaluated by the fMRI experiment by testing the hypothesis that metabolic activity correlating to reward-prediction error will be localized to distinct cortico-striatal locations where the afferent information represents the actions that are to be reinforced. Note that this hypothesis differs in three important ways from prior research on reward and the basal ganglia. First, prior reward research has largely focused on the involvement of the ventral striatum (i.e., nucleus accumbens) and the orbital-frontal cortex as the loci of reward-related learning (Robbins and Everitt, 1992; Rolls, 2000; Stern and Passingham, 1996). A view based on this prior research predicts that the ventral striatum would be active during the reinforcement of all three types of action. Although interesting, carrying out a study to investigate this hypothesis would not help explain how different actions are reinforced. A more theoretically valuable hypothesis to investigate is whether striatal activity differentially reflects the reinforcement of specific types of actions. Thus, we predicted that particular regions of the striatum would be involved in the reinforcement of actions, and that those particular regions should be those receiving information from neocortical regions controlling those types of actions. Second, a substantial amount of prior research has examined the involvement of the striatum in outcome rather than in learned reward predictions (Delgado et al., 2003, 2000). However, in those tasks, the paradigms were set up so that outcome was randomized rather than according to a fixed reward contingency. Because of this, it could be argued that, in those tasks, reward-prediction error was always equal to some proportion of the outcome because no accurate predictions about the reward could be learned. Thus, in such

paradigms, reward-prediction errors would likely only correlate weakly with striatal activity. In contrast, in the current task, the mapping between stimulus action and outcome remained constant for substantial portions of the experiment, allowing participants to effectively and correctly learn to predict rewards. Thus, we predicted that the signals in the striatum would reflect the difference between the learned reward and the actual reward received by the participant on each trial. Third, prior research has emphasized that the dopamine neurons project widely to all parts of the striatum, and that these neurons all fire in the same way. This fact, along with the findings that the hemodynamic response is primarily driven by postsynaptic neurotransmission (Attwell and Iadecola, 2002) suggests that the entire striatum or parts of the striatum in some non-selective pattern, rather than action-specific regions, are likely to appear in studies of striatal BOLD responses. In addition, the fact that dopamine acts directly on the vasculature (Krimer et al., 1998), suggests that action-specific processing due to the joint activity of cortico-striatal terminals and dopaminergic innervation may not be sufficiently distinct to be detected by a study reliant on the BOLD response. Despite these issues, a recent review concludes that the BOLD response in the ventral striatum is attributable to dopamine, and in particular to the activation of D1 receptors there (Knutson and Gibbs, 2007). This suggests that a study that aims to test for BOLD evidence for the computation of reward-prediction error in distinct subregions of the striatum could, in principle, be carried out.

The prediction that the reinforcement of specific actions will give rise to differentially-localized striatal fMRI signals based on local computation has remained largely unexplored. However, two recent neuroimaging studies have examined differential activation of striatal subregions related to dopamine release. The first study was based on individual differences attributed to dopamine receptor distributions in the striatum (e.g., Clatworthy et al., 2009). The second study was based on subregion-specific loss of dopaminergic terminals in elderly populations (Landau et al., 2009). In contrast to these studies, the main prediction in this study is that action specific reward-prediction errors will be computed in distinct striatal subregions when different actions, either physical or cognitive, are being reinforced – for normal participants.

Finally, prior computational simulations are reviewed. These simulations indicate how coordinated, intelligent behavior can emerge in a RL agent when separate, locally-controlled actions are reinforced. This occurs even when the agent is so simple that direct coordination between actions

(as might be mediated by cortico-cortical connections) is not allowed. Together, these results all point to the idea that arbitrary local patterns of cortical activity can be reinforced by global reward, and therefore recurrently excited via loops with the basal ganglia.

3.3.3 METHODS

General procedure. The experimental procedures were approved by the University of Pittsburgh's Internal Review Board. All of the behavioral and fMRI experiments followed the same general design (similar to experiments carried out by Posner, 1980, but requiring three different types of responses: e.g., hand movements, eye movements, or covert shifts of visual attention). At the beginning of each run of trials requiring a particular response type, participants were instructed as to the type of action to perform in making their responses. The type of action could be a hand movement, an eye movement, or a covert shift of visual attention. The sequences of events on each of the three types of trials are visually depicted in Figure 3.1 (page 41). To make the procedure concrete, a trial involving eye movements will be described. During each trial, participants began by fixating the center of the screen. A color cue stimulus then appeared that consisted of five identically colored boxes (red, green, blue, or purple), with one (fixation) box in the center of the screen and the four remaining (target) boxes arranged in the cardinal positions on the screen. Participants were instructed to quickly move their eyes from the fixation box to one of the four peripheral target boxes as soon as the stimulus appeared. The stimulus was displayed for 1 second; four different random probe letters (e.g., A, I, O, or U) then flashed for 100 ms in each of the four target boxes (see Fig. 1). Contingent upon the eye movement that was performed, participants saw one of the probe letters. Participants were instructed to report the identity of the letter that they saw in the target location. Participants were then provided with reinforcement feedback depending on whether the probe letter they reported corresponded to the probe letter displayed in the box in the rewarded location. If they had moved their eyes in the correct (rewarded) location, then "CORRECT +$0.05" was displayed and participants received $0.05; however, if they had moved their eyes in one of the three incorrect (unrewarded) locations, then "INCORRECT -$0.05" was displayed and participants lost $0.05. The selection of rewarded and unrewarded directions is described in the next section.

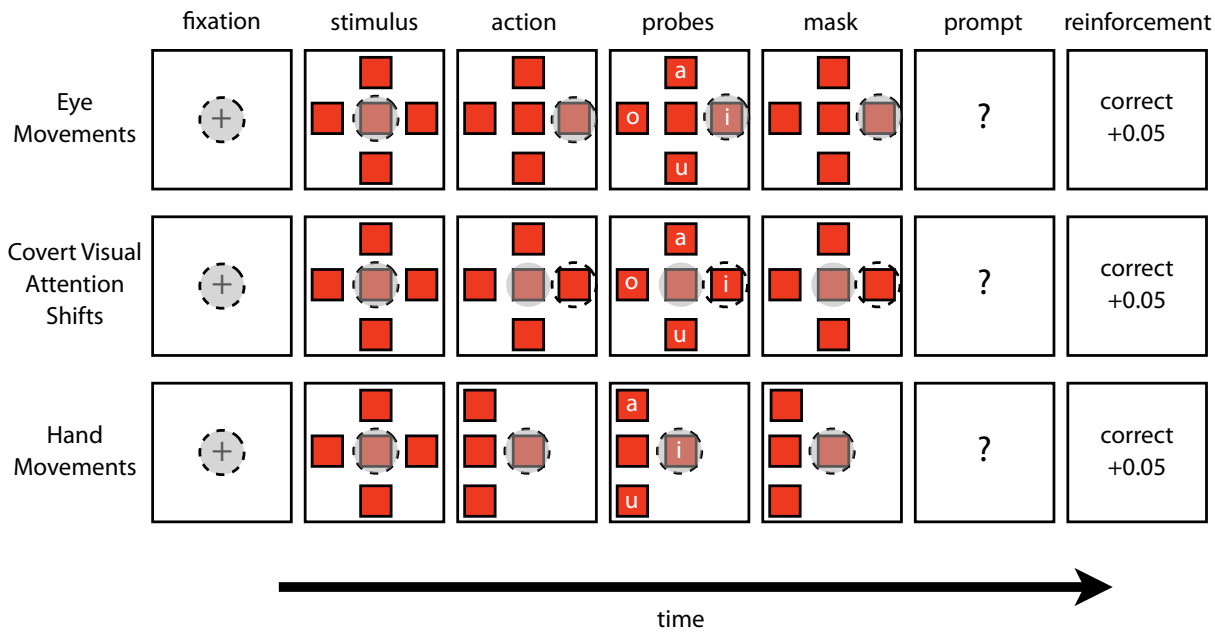

Figure 3.1: **Graphical Representation of the Multi-modal Reinforcement Learning Experiment.** This figure illustrates experiment procedure used in each trial of the behavioral and fMRI experiments. Time proceeds from the left to the right of the figure. Details of the display parameters and timing are included in the text. Note that action type was blocked, so that during each experimental run, participants performed only a single type of action (e.g., a series of eye movements) in one of four directions depending on the stimulus color (e.g., red [shown here], green, blue, purple). There were breaks between runs of each type of action. Behavioral Experiment 1 used eye movements during one run, and covert attention shifts during the other run. Behavioral Experiment 2 used eye movements during one run, and hand movements during the other run. The fMRI experiment used all three types of action.

Reward schedule. Reward was delivered contingent upon the participant making the appropriate action (i.e., moving the hand, eyes, or covert attention to the correct location) in response to each of the cue stimulus colors. The mapping between colors and rewarded actions were randomly determined at the beginning of the experiment and maintained until they were randomly changed

after a fixed number of trials (i.e., halfway through the run using a given task). Although one might be concerned this reversal might engage reversal-learning processes, prior research has shown that even reversal learning engages the dopamine system (Cools et al., 2009). In addition, there were two kinds of trials NORMAL versus INVERTED. On NORMAL trials (i.e., 90% of trials), participants were rewarded for moving to the reward location, as already described. On those trials, movement to any of the other three locations resulted in "punishment" (i.e., -$0.05). However, on the remaining INVERTED trials (i.e., 10% of trials), the contingency between color and reward location was inverted so that the rewarded location was actually punished if selected, and any of the three remaining locations were rewarded if selected.

3.3.4 EXPERIMENT 1: EYE MOVEMENTS VS ATTENTION SHIFTS

Experiment 1 compared the reinforcement learning of saccadic eye movements and covert shifts of visual attention.

Participants. 35 participants were recruited for the study through advertisements placed on the University of Pittsburgh campus. Participants had normal or corrected-to-normal vision using contact lenses. Participants were paid a base rate of $7 per hour, and were paid extra money according to the number of correct responses made during the experiment. No participants were eliminated because of their inability to perform the actions required by the experiment.

Apparatus. A desk tower-mounted SR Research EyeLink 1000 eye tracker was used in the study. Participants were seated in front of a 23-inch computer monitor with their chin on the eye-tracker s chin-rest so they were approximately 25 inches away from the screen. The eye tracker was calibrated to record the position of the participants' eye movements to within 0.7 degrees of visual angle. Gaze location was sampled every millisecond. Viewing was binocular, but only the right eye was tracked. Participants used the arrow keys on a keyboard to enter their responses to a probe screen at the end of each trial. The behavioral experiment was implemented in SR Research's graphical Experiment Builder version 1.5.

Design. Participants were instructed to learn which of four actions to perform (in response to each of the four color cue stimuli) to maximize their obtained reward. The experiment consisted of

360 trials: 180 trials for eye movements and 180 trials for covert attention shifts. The experiment lasted 1-1.5 hours. A message appeared on the screen between runs to inform participants that they could take a short rest break.

At the beginning of each run, participants were instructed about which task to perform and then given a small number (45) of practice trials. These practice trials ensured that participants were able to see the probe letters and perform the actions (i.e., either saccades or covert attention shifts without moving their eyes from the central fixation location) adequately. After the practice period, participants were encouraged to indicate whether they understood the task. Participants were then told that the actual task was going to begin, that there would not be any explicit arrows indicating the directions of the correct action s direction, and that they would have to learn the mappings between colors and actions. The participants were also reminded that one or more colors could map to the same action, so that they could not use the process of elimination to learn the mappings. Finally, participants were told that the correct action for each color could change at any time during the experiment, but that they should do their best to earn as much money as possible.

As shown in the first two rows of Figure 3.1 (page 41), participants quickly made an action (i.e., eye movement or attention shift) from the central box to one of the four peripheral boxes. After the stimulus had been displayed for 1 second, a single randomly selected probe letter (from the set B, D, O, and P, sampled without replacement) at 12-point font size was briefly displayed in each peripheral box for 100 milliseconds. The letters that were displayed in each peripheral box were randomized from trial to trial. To offset visual acuity limitations during the covert attention shift condition because eye movements were not permitted during the covert attention task, the probe letters were displayed in a larger (24-point) font. After a 1-second delay, a probe test screen then appeared asking participants which letter they had seen at the location where they had moved their eyes or attention. Participants indicated the letter by pressing the arrow key on the keyboard corresponding to the position of that letter on the probe test screen. If the correct letter was reported (which barring correct guesses would only occur following movements to the correct location), participants were shown a feedback screen indicating a +$0.05 reward. If participants reported an incorrect letter, they received a -$0.05 punishment. This feedback was displayed for 1 second and was followed immediately by the next trial. During the covert attention shift trials, a trial was

terminated (with neither reward nor punishment) if participants moved their eyes from the center of the screen. A region with a radius of approximately 2 inches had to be fixated during the entirety of the trial for the trial to be considered valid. The text "Do not move your eyes" was then displayed during the timeout (lasting 2 seconds) until the next trial began.

3.3.5 EXPERIMENT 2: EYE MOVEMENTS VS HAND MOVEMENTS

Behavioral Experiment 2 was identical to Behavioral Experiment 1, except that it compared the reinforcement learning of saccadic eye movements and hand movements using a joystick.

Participants. 40 participants from the same population using in Experiment 1 were recruited. Participants were paid exactly as in Experiment 1. None of the participants were eliminated because of their inability to perform the experiment.

Apparatus. The same eye tracker used in Experiment 1 was used in Experiment 2. To ensure consistent effector usage during the learning of hand movements, a video game joystick with a rubberized base (Logitech Attack 3 Joystick) was used, rather than arrow keys on a keyboard. The joystick was placed on the table in front of the participants so that it could be easily reached.

Design. Behavioral Experiment 2 was identical to Experiment 1 except for two changes that were made to accommodate the hand-movement condition. First, when the participant moved the joystick, the screen scrolled so that the peripheral box that was located in the direction of joystick movement scrolled to the center of the screen. Consequently, the probe letter that briefly flashed in this box was moved to the location of fixation. (As in the covert attention shift condition in Experiment 1, eye movements away from the central fixation region resulted in a timeout during which the text "Do not move your eyes" was displayed until the next trial began.) The second change was with respect to how participants reported probe letters: Rather than pressing keys on a keyboard to indicate the identity of the probe letter, participants were instructed to simply say the letter's name. This was done to avoid the use of the same effector for both learned actions (i.e., hand vs. eye movements) and the probe responses. A pilot study suggested that hand-movement learning was adversely affected by the use of hand movements to also indicate the random probe letter. The experimenter was in the room and entered the reported probe letters on a keyboard so

that the appropriate reinforcement feedback could be provided to the participant.

As in Behavioral Experiment 1, participants were given a small number of practice trials (here, 20 trials) prior to the beginning of the task.

3.3.6 EXPERIMENT 3: EYE MOVEMENTS VS HAND MOVEMENTS VS ATTENTION SHIFTS

An fMRI experiment was carried out to determine whether reward-prediction errors predicted by RL are located in anatomically distinct regions of the striatum for hand movements, eye movements and covert shifts of visual attention, in accordance with predictions from our RL framework. The experiment was similar to Experiment 2, but included all three action types, i.e., hand movements, eye movements, and covert attention shifts. Also, in order to better isolate fMRI signals during reinforcement learning, we included a run of a fixed number of CONTROL trials (described below) prior to each experimental run.

Participants. 19 participants were recruited for the study through advertisements placed on the University of Pittsburgh campus and through lists of participants who had participated in prior fMRI studies in affiliated laboratories. The requirements for the study were: normal or corrected-to-normal vision, right-handedness, and not taking drugs for any mental disorder. Participants were paid a base rate of $25 per hour, and were paid extra money according to the number of correct responses that they made during the experiment. None of the participants were eliminated because of their inability to perform the experiment.

Apparatus. A 30-minute behavioral practice session took place prior to the fMRI scan using the same room and equipment as Experiments 1 and 2. The fMRI scans then took place that day or the next at the Brain Imaging Research Center (BIRC) at the University of Pittsburgh. The MR scanner was a Siemens 3T Allegra scanner. A fiber-optic all-plastic joystick (model HH-JOY-4 from Current Designs: http://curdes.com) was used. The joystick movements were interpreted into keyboard button presses using the free software JoystickCursorTool (joycur.exe: http://www.deinmeister.de/joycur_e.htm). Eye tracking was done in the MR scanner (using an ASL Tilt-Pan Optics 500) to ensure that participants did not move their eyes during hand-movement and

covert-attention-shift trials. Eye movement data was streamed over a serial connection from the ASL Eye-Tracker interface to the experiment display computer. On trials where participants were required to maintain central fixation, the trial was terminated if participants fixated outside of a region surrounding the central stimulus for more than 500ms.

A microphone to detect the participants' oral responses was oriented towards the participant's head and taped to the head coil. A plastic insulator was inserted between the microphone and the head coil to reduce irregular noise caused by vibrations. To allow the experimenter to hear the participant responses, cancellation of the repetitive scanner noise was performed in real time by a Macbook Pro laptop running a script (see Appendix B) implemented in the strongly-timed audio programming language ChucK 1.2.2 (http://chuck.cs.princeton.edu/). The script simulated a virtual y-plug that combined the current incoming sound with a time-delayed inverted copy synchronized to the scanner repetition time (which was approximately 2 seconds).

Practice Session. Prior to the fMRI scan, all participants underwent a practice session to gain familiarity with and to ask questions about the task (because such an opportunity was not possible during the fMRI experiment). During practice, the eye-tracker was used to ensure that participants did not move their eyes during the hand-movement and covert-attention-shift conditions. Participants had 10 practice trials that, similar to the CONTROL condition, involved the presentation of an arrow that indicated the correct location for the current action (e.g., move the eyes left), followed by 20 learning trials. These practice (CONTROL) trials thus entailed no or little learning. Three participants had difficulty with the covert attention shifts, and were allowed additional practice trials.

fMRI Session. Prior to the fMRI scan participants were instructed to breathe regularly, and importantly, to not take deep breaths before reporting the probe letters because of the potential introduction of artifacts (Kwan-Jin Jung, personal communication). Participants were also told that their vocal responses would be processed using sound cancellation technology, so that speaking at a normal volume would be sufficient to enable us to hear their responses. In a few cases where it was difficult to hear the participant or distinguish between two letters because of participant's accent, the experimenter indicated this to the participant during the pause interval that occurred between the CONTROL and LEARN runs. Additionally, to maximize intelligibility, minimize head movement,

and minimize any sinus closures that might result in magnetic susceptibility artifacts, the vowels A, I, O, and U were used for the probe responses.

After a participant was moved into the scanner, the joystick was attached to the scanner bed or bore so that the participant could comfortably reach it. A screen was displayed to test whether the joystick was registering intended actions, and the experimenter adjusted the sensitivity threshold of the JoystickCursorTool software for detecting movement. After the joystick calibration, both T1 (MPRAGE) and T2-structural images were acquired. Then, seven minutes of resting state functional data was acquired during which the participant was instructed to lie still, with his/her eyes focused on the word "Relax" displayed in grey font at the center of the screen against a dark background. After this, the participants were instructed that the experiment would begin.

fMRI Experiment Design. The entire experiment consisted of six scanner runs, alternating between CONTROL runs, during which a triangle indicated the direction of the "correct" action and no reinforcement feedback was provided, and LEARN runs, during which the participant had to determine the direction of the correct action and received reinforcement feedback. The order of the action modalities (i.e., hand, eye, or attention movement) was pseudo-randomized across participants, but with the CONTROL run always preceding the LEARN run in the same action modality. At the beginning of each run, participants were informed (vocally by the experimenter and by a message on the screen) about the nature of the next run. Each CONTROL run contained 60 trials, and each LEARN run contained 120 trials. During the LEARN runs, a new random color-action contingency mapping was introduced after 60 trials (rather than 90 trials, as in Experiments 1 and 2). Thus, there were 540 trials: 180 CONTROL trials + 360 LEARN trials.

Finally, to adapt the basic paradigm to the fMRI experiment, two other design changes were made. First, because the fMRI experiment was a fast-event related design, jitter consisting of randomly selected periods lasting 1, 2, or 3 TRs (i.e., 2, 4, or 6 seconds) was introduced between the trials to improve the ability to deconvolve the fMRI signal. Approximately 20% of the experiment time was spent in jitter-introduced TRs. Second, to help ensure that attention was being directed to one of the peripheral boxes, the probe letters were replaced with a mask within 100 ms of their appearance by an ampersand character. This was done to ensure that participants were attending to the appropriate location at the appropriate time.

fMRI Image Acquisition. A 3-tesla MRI scanner (Siemens Allegra) was used to acquire structural T1-weighted MPRAGE images (35 slices, slice gap = 0mm, slice thickness = 1mm, TE = 3ms, TR = 1540ms, flip angle = 8 degrees, matrix = 256x256, FoV = 256mm), structural T2-weighted images (35 slices, slice gap = 0mm, slice thickness = 3.2mm, TE = 73ms, TR = 5530ms, flip angle = 150 degrees, matrix = 256x214, FoV = 210mm), and functional T2*-weighted echo planar images (38 slices, slice gap = 0mm, slice thickness = 3mm, TR = 2s, TE = 25ms, flip angle = 70 degrees, FoV = 200mm) to examine blood oxygen level-dependent (BOLD) contrasts.

fMRI Image Preprocessing. Functional images were preprocessed using AFNI (http://afni.nimh.nih.gov/). After reconstruction from DICOM format using to3d, time-slice correction was applied to the images to de-interlace them. The images were then corrected for head motion, i.e., volume registration, and then alignment using AFNI's align_epi_anat.py python script. Functional imaging signals were then restricted to within the brain using an auto-thresholded mask. For each run, the intensity values of each voxel were normalized to a common signal intensity range using its own mean within the run. Because of the highly specific predictions about the anatomical localization of the signals, no spatial smoothing was performed during preprocessing.

For the purposes of group-level analysis and standard coordinate reporting, structural images were warped to the TTN27 template brain provided by AFNI using the default settings of the nonlinear warping toolkit Advanced Normalization Tools version 1.6 (ANTS; http://www.picsl.upenn.edu/ANTS/). In the present study, the use of a high-dimensional nonlinear warp procedure was preferred to the linear affine transformations performed in AFNI because of the large degree of inter-subject variability we observed in the striatum, in particular, in the dorsal parts of the caudate nucleus. (The default settings used by ANTS were: transformation = SyN[0.25], maxIterations = 30x90x20, MultiResolution = 3, regularization = Gauss[2,0], metric="PR[".)

fMRI Data Analysis. The preprocessed functional data was subjected to a deconvolution analysis in AFNI's 3dDeconvolve program at the individual level. The resulting regression coefficient (i.e. "beta") maps were smoothed with a Gaussian kernel of 5mm full-width half-max. A *t*-test against zero was performed to derive group-level results. Nine regressors (consisting of binary events for movements, modeled scalars for positive reward-prediction errors, and modeled scalars for negative reward-prediction error, for each of the three action modalities) were then applied to

the preprocessed (see Appendix A) fMRI data using AFNI's 3dDeconvolve program. No hemodynamic shape was assumed, because of prior findings showing deviations from the standard hemodynamic response function in the striatum reward and punishment signals (Delgado et al., 2003, 2000). The impulse response function was estimated for TRs 1 through 11 (i.e., from 0 to 22 seconds from the event of interest). Given the length of the scanner runs, AFNI s recommended baseline polynomial of order 6 (i.e., "polort = 6") was included in the statistical model to capture trends in the data due to scanner drift and periodic signals. The coefficients of the fitted functions (i.e., the "beta weights") from 3dDeconvolve were then transformed to a standard template brain in Talairach space, TTN27, using the ANTS WarpImageMultiTransform command. Group-level *t*-tests were then performed to determine whether the coefficients were reliably greater than 0.

The reward-prediction error regressors were obtained by computing using a Rescorla-Wagner rule (Rescorla and Wagner, 1972) to update a stored reward-prediction value associated with the combination of the stimulus viewed (e.g., "green boxes"), action taken (e.g., "shifted attention to left box"), and reinforcement received (e.g., "-$0.05") on every trial **i** by each human participant.

$$V_{i+1}(\text{color \& action}) = V_{i}(\text{color \& action}) + \text{learningRate} * \text{rewardPredictionError}_{i}$$

where

$$\text{rewardPredictionError}_{i} = [\text{reinforcement}_{i} \quad V(\text{color \& action})_{i}]$$

In this equation, "V" indicates the value associated with a particular combination of action and stimulus. To determine the learning rate parameter for the group of participants, we performed 10 Monte Carlo RL agent simulations for each condition (i.e., independent agents for hand movements, eye movements, and covert attention shifts). The learning rate (i.e., the rate at which the agent updated the value function that guided its action selection) was varied from 0.1 to 1.0 in steps of 0.1. For each learning rate, the output behavior of the each agent was compared to that of each human participant. The learning rate that allowed the agents to best fit the actions selected by all 14 participants (e.g., the rate that minimized the mean absolute deviation between the simulated and observed values) in all conditions simultaneously was 0.4. For all subsequent analyses the learning rate was therefore fixed to this value, 0.4, and reward-prediction errors generated by each agent on each trial were then used to derive regressors for each participant. The positive and negative components of action-specific reward-prediction errors were separated for each of the three types

of action under the assumption that the hemodynamic response shapes for reward and punishment differ (Delgado et al., 2003, 2000).

3.3.7 BEHAVIORAL RESULTS

Asymptotic Performance and Learning Rates. Because we were interested in studying neural mechanisms underlying the learning of different types of actions, we first sought to carry out a behavioral study to obtain performance data. In particular, we wanted to determine whether participants would be able to learn each of the three actions to an asymptotic, optimal level within a single experimental session, or whether the different actions are more or less difficult to learn. If it turned out that a particular action was fundamentally more difficult to learn through reinforcement, then this would provide justification for making specific differential predictions about the underlying neural effects of reinforcement on that action.

We therefore quantified three aspects of human behavioral data as they were rewarded for performing hand movements, eye movements, and covert attention shifts: (1) asymptotic performance levels; (2) fitted curve learning rate parameters; and in the next section (3) fits between human and RL agent action selection. First, we measured participants' asymptotic performance levels by quantifying the percent correct responses achieved by the end of each run. The maximum expected asymptotic performance level was 90% correct because 10% of the trials were randomly inverted (see Methods). Minimum (i.e., chance) performance was expected to be 25% because there were four possible actions on each trial. We therefore excluded participants that were over two standard deviations away from the 90% performance level. (These excluded participants were also within one standard deviation of chance performance.)

To test the hypothesis that different actions could be learned equally well, we compared the asymptotic behavioral performance levels between the different conditions to see whether they were equal. Asymptotic performance levels were defined to be the average number of rewarded responses during the last 25 trials. The trials were binned into groups of five because participants made no responses on some of the trials (e.g., trials that required participants to maintain central fixation were terminated early if participants moved their eyes). Thus, if there was no response for a particular trial, it was counted as neither correct nor incorrect according to this measure.

For example, if participants complete 4 of the 5 trials in a bin, and got 1 response incorrect, then the percent correct was computed with respect to 4 responses, rather than 5 trials (i.e., 4 - 1/4 = 75% correct). We first examined the 35 participants in Experiment 1, which compared RL of eye movements and covert attention shifts. As indicated, the data from participants whose final asymptotic performance at the end of the task in any condition was more than 1.96 standard deviations away from the theoretical optimal performance level of 90% were eliminated from further analyses. This cutoff was 90% - 1.96 x 0.2275 = 45.4% for the eye-movement condition, and 90% - 1.96 x 0.2787 = 35.4% for the covert-attention-shift condition. Using these cutoffs, we eliminated eight participants and then tested for differences in asymptotic performance levels in the remaining 27 participants. Binned performance data for these participants is shown in the top portion of Figure 3.3. We found that, at the halfway point in each experimental condition (i.e., before the first contingency switch), asymptotic performance levels were not significantly different between eye-movement and covert-attention-shift conditions (difference = 5.14%; $t(26) = 1.4768$, $p = 0.1517$, n.s.), nor were these conditions different at the end of the experiment (difference = 2.3%; $t(26) = 0.6484$, $p = .5224$, n.s.). To maximize our power to detect any such differences, we then pooled the asymptotic performance data from the halfway point and the final asymptotic levels from both conditions and again found no significant differences between eye movements and covert attention shifts (difference = 1.41%; $t(53) = 0.5600$, $p = 0.5779$, n.s.). A power analysis (using G*Power; Faul et al., 2007) indicated power of $p = 0.86$ to detect a difference of 10% (i.e., an effect size of 0.5851 computed using a pooled standard deviation of 0.1694); thus, the failure to find any such differences is unlikely to be due to inadequate statistical power.

The results of Experiment 1 suggest that participants were able to learn to move both their eyes and covert attention through reinforcement, and that they were able to do so equally well. Furthermore, eye movements and covert attention shifts themselves do not seem to require learning processes that would significantly affect the ability of participants to learn such behaviors from reinforcement.

Next, we carried out the same analysis for the 40 participants in Experiment 2. Participants whose final asymptotic performance at the end of the task in either condition was more than 1.96 standard deviations away from the theoretical optimal performance level of 90% were eliminated

from further analyses. These cutoffs were 90% - 0.4435 = 45.6% and 90% - 0.4077 = 49.2% for the hand- and eye-movement conditions, respectively. We then tested for differences in asymptotic performance in the remaining 32 participants, whose binned performance data is shown in the bottom portion of Figure 3.3. At the halfway point in each experimental condition, asymptotic performance levels were not significantly different between the hand-movement and eye-movement conditions (difference = 4.01%; $t(31) = 0.7535$, $p = 0.4569$, n.s.), nor were they different at the end of the experiment (difference = 2.34%; $t(31) = 0.9012$, $p = 0.3744$, n.s.). To maximize our power to detect differences, we pooled the asymptotic performance data from the halfway point and final asymptotic levels from both conditions, and found no significant differences between eye movements and hand movements (difference = 3.18%; $t(63) = 1.0806$, $p = 0.2840$, n.s.). A power analysis indicated power of $p = 0.844$ to detect a difference of 10% (i.e., an effect size of 0.5296 using a pooled standard deviation of 0.1888).

As with Experiment 1, these results indicate that participants were able to learn eye movements and hand movements through reinforcement, and suggest that they were able to learn both equally well. Hand movements also do not require learning processes that significantly affect the ability of participants to learn from reinforcement. Taken together, the results from Experiments 1 and 2 suggest that these three types of actions hand movements, eye movements, and covert attention shifts can be learned equally well by participants.

In a second analysis, we evaluated the rate of learning for the different types of action. We used MATLAB (The Mathworks, Inc., Natick, MA.) Curve Fitting Toolbox (i.e., cftool) to fit the coefficients of a power law curve describing the relationship between behavioral performance binned in time (i.e., the proportion correct) to the bin number. The coefficients that were fit were the variables a and b in the equation $y = at^b$, where y was the binned performance and t is time, in bins containing 5 consecutive trials each. In Experiment 1, eye movements were fit with $y = 0.3676t^{0.2859}$ ($r^2 = 0.9331$), and covert attention shifts were fit with $y = 0.3245t^{0.3329}$ ($r^2 = 0.9501$). In Experiment 2, eye movements were fit with $y = 0.3866t^{0.2754}$ ($r^2 = 0.8991$) and hand movements were fit with $y = 0.3398t^{0.3083}$ ($r^2 = 0.9725$). The fits, along with the 95% confidence intervals for the fits, are plotted in Figure 3.4 (page 70). The high degree of overlap within the fits provides additional evidence consistent with the hypothesis that the different types of actions are likely learned using either the

same or a similar algorithm.

Together, the behavioral results strongly suggest that participants are able to learn these different behaviors equally well. Furthermore, these results suggest the possibility that widely different actions are learned using a similar algorithm, although they do not suggest any particular algorithm. Therefore, we next proceeded to test whether the data were consistent with an algorithm based on learned reward predictions such as RL.

RL Agent Fits. Although the analyses described in the preceding section suggested that a similar learning algorithm might be used to reinforce different types of actions, they do not test this claim with respect to a particular learning algorithm. To address this limitation, we used RL simulations to test the hypothesis that RL could explain the patterns of learning observed with each of the actions. Specifically, we tested whether RL agents are capable of fitting the participants behavior equally well, regardless of the type of action being reinforced. If there is a difference between the fits as a function of the behavior being simulated, then this might provide a basis to suggest that other learning processes (e.g., error-driven motor learning) are involved instead of or in addition to RL. We simulated RL agents on the same sequence of stimuli, actions, and reinforcement that were experienced by each participant, comparing the output of the agent on each trial to the action selected by the human participant. The rate at which positive and negative reinforcement updated the value function (i.e., the learning rate) was varied from 0.1 to 1.0 in steps of 0.1. There were 10 Monte Carlo agents simulated per participant and learning rate, and a randomized tie-breaking procedure was used when competing actions had identical values. The percentage of the number of actions selected by both the human participant and the simulated RL agent was used to assess the fits.

In terms of the proportion of the same action directions selected by both humans and RL agents, human eye movements and covert attention shifts were fit equally well by RL agents in Experiment 1 using a single learning rate (0.3) to simultaneously fit for both actions. Humans and RL agents selected the same action (e.g., move left) on 67.59% and 67.72% of the trials in the eye-movement and attention-shift conditions, respectively [$t(26) = 0.0529$, $p > 0.95$, n.s.]. In a similar manner, human eye and hand movements were fit equally well by RL agents in Experiment 2 trained with a single learning rate (0.5) for both actions. Humans and RL agents selected the same action on

70.06% and 66.65% of the trials in the eye- and hand-movement conditions, respectively [$t(32) = 1.4320$, $p > 0.16$, n.s.]. We also directly compared eye movements in the first and second behavioral experiments by performing a two-sample *t*-test assuming equal variances and found no significant difference [$t(57) = 0.8839$, $p > 0.38$, n.s.].

These simulation results and analyses suggest that RL agents that choose actions based on learning from reward-prediction error fit the three kinds of action equally well. They also suggest that the same kind of learning occurs for physical actions (e.g., eye and hand movements) and cognitive "actions" that have no motor component. Taken together, these results and the results of the analyses of asymptotic performance that were reported in the previous section indicate that all three actions can be used in a task that is designed to investigate whether reward-prediction error computations are being computed in the striatum for different kinds of action. The results of the fMRI study and analyses that are reported in the next section provide additional evidence about the implementation of the RL algorithm that supports the learning of these actions.

3.3.8 FMRI RESULTS

Striatal Activations. The main neuroscientific hypothesis of this paper is that reward-prediction errors associated with each type of action (i.e., hand movements, eye movements, and attention shifts) are computed in distinct parts of the striatum that are the targets of those actions' neocortical representations. To test this hypothesis, we performed a multiple regression analysis for action-specific reward-prediction error signals in the striatum of participants as they received reinforcement for performing those actions. As described in the Methods, these reward-prediction error signals were derived from RL agents that simulated the data from each participant's LEARN-condition runs. Positive and negative reward-prediction errors were set equal to values of zero during the CONTROL-condition portion of the experiment, under the assumption that no reward predictions were being made during the control condition, and thus that no reward-prediction errors were hypothesized to be computed by participants during these trials. In addition to the action-specific reward-prediction error regressors, action-specific movement regressors were also included in the model because participants actually performed an action of the instructed type in each trial. This allowed us to test the additional hypothesis that reward-prediction errors were being computed in

the same general striatal regions that are responsible for controlling the movements themselves.

The table in Figure 3.2 (page 68) lists the largest and most significant regions of interest (ROIs) in the striatum that correlated significantly with positive and negative reward-prediction error, and with movement. As the table indicates, the parts of the striatum that correlated significantly with reward-prediction error and movement were the regions known to receive afferent information from neocortical regions that control the actions that were being reinforced. Specifically, we found that when hand movements were being reinforced, positive reward-prediction error correlated with activity in the left putamen and the anterior caudate (Figure 3.5 on page 71, Panels E & F). Negative reward-prediction error for hand movements correlated with smaller regions in bilateral putamen (Figure 3.6 on page 72, Panels A & B). When eye movements were being positively reinforced, positive reward-prediction error correlated with activity in both the anterior caudate and the dorsal caudate (Figure 3.5, Panels C & D). In addition, negative reward-prediction error for eye movements correlated significantly with activity in the anterior caudate, dorsal putamen and nucleus accumbens (Figure 3.6, Panels C & D). Finally, when covert attention shifts were being reinforced, correlated activity was localized to the anterior caudate for both positive reward-prediction error and negative reward-prediction error (Figures 3.5-3.6, Panels A & B).

These results suggest that reward-prediction error computations for different actions (i.e., right hand movements, saccadic eye movements, and covert attention shifts) are localized to different parts of the striatum (i.e., the anterior caudate, dorsal caudate, and left putamen, respectively). Furthermore, because the parts of the striatum that correlated with reward-prediction error were in locations to which neocortical representations for those actions are known to project, these results are consistent with the hypothesis that reward-prediction errors are computed in the cortico-basal ganglia loop circuits.

We also asked whether striatal activation involved in the control of the movements themselves were located in the same parts of the striatum that computed reward-prediction errors. To obtain a map of movement without the influence of reinforcement, we performed a second deconvolution analysis involving only the CONTROL trials during which no reward feedback was given to the participants. The *t*-values associated with the peak voxel intensities of the ROIs are also included in the table (Figure 3.2) as "movement-related activity", and the ROIs themselves are shown visually

in Figure 3.7 (page 73).

As can be seen, the results are generally consistent with this hypothesis. The striatal regions correlating with hand movements (i.e., the left putamen) were involved in the reinforcement of hand movements (compare Figures 3.5-3.7, Panels A & B). The striatal regions correlating with eye movements (i.e., the anterior caudate) also are involved in the reinforcement of eye movements, although the dorsal caudate was not active above threshold for saccadic eye movements (compare Figures 3.5-3.7, Panels C & D). Striatal regions correlating with covert attention shifts (e.g., anterior caudate) overlapped with the regions involved in the reinforcement of attention shifts (e.g., compare Figures 3.5-3.7 Panels E & F).

Delegation of Control: Decreased Involvement of the PFC. We hypothesized that although the control of particular actions might initially be effortful and voluntary, rewarding those actions would allow them to shift into a more automatic mode of control that would no longer require the prefrontal cortex. Thus, the prefrontal cortex is hypothesized to be significantly more active when reward learning has started compared to after it is well underway. This hypothesis is in line with our proposed framework, which suggests that with sufficient reinforcement, the control of actions can been delegated to cortico-striatal loops that are able to perform these actions independently from prefrontal control. To test this idea, we examined whole brain activity for regions that exhibited such a temporal profile. The data in each LEARN run was de-trended by removing a sixth-order linear least squares-fit polynomial on a voxel-wise basis (using AFNI's 3dDetrend), and then means were computed for each quarter into the run (first, second, third and fourth). These means were subjected to a one-factor between-runs voxel-wise ANOVA, with the quarter (i.e., time) into the run as the factor; action type was considered to be a random effect. The largest and most significant region in the resulting statistical map was located in the prefrontal cortex (Middle Frontal Gyrus / Brodmann area 8; peak coordinates x=-14mm, y=23mm, z=36mm, $p < 0.0001$ corrected for multiple comparisons), shown in Figure 3.8 (page 74). This result suggests that the prefrontal cortex ceases to be involved in the control of actions once they have been sufficiently reinforced. Pairwise post-hoc *t*-tests showed that the peak of this cluster was more active during the first and third quarters (i.e., when a new random mapping between colors and directions was introduced) compared to the second and fourth quarters (first versus second, $t(13) = 3.10$, $p < 0.005$; third versus

fourth, $t(13) = 2.97$, $p < 0.006$.) This result suggests that the change in action-outcome mapping that occurred halfway through each LEARN run reengages the prefrontal cortex simultaneously with the basal ganglia.

3.3.9 GENERAL DISCUSSION

This thesis presented behavioral, fMRI, and computational evidence in support of the idea that a global reinforcement signal (as is hypothesized to be mediated by phasic dopamine), combined with a particular pattern of connectivity (the cortico-basal ganglia loop) allows the brain to learn to control its own actions and thereby generate adaptive behaviors in spite of its very limited connectivity. Importantly, these behaviors can consist of physical actions (i.e., involving muscle effectors, such as hand movements) and/or cognitive "actions" (i.e., with no motor component, such as covert shifts of visual attention). More specifically, the behavioral experiments demonstrated that all three types of action could be learned through reinforcement at similar learning rates, and to similar asymptotic levels of performance. The present fMRI experiment is the first instance to show that reward-prediction errors are computed in separate regions of the striatum for different types of actions. Furthermore, it showed these computations occur in regions that correspond to anatomical and functional inputs into cortico-basal ganglia loops previously known to be involved in the control of those actions. Specifically, metabolic activity related to the reinforcement of hand movements was localized to the putamen, activity related to the reinforcement of eye movements was localized to the dorsal and anterior caudate, and activity related to the reinforcement of covert attention shifts was localized to the anterior caudate.

One particularly interesting and somewhat unanticipated result was the activation of the anterior caudate during the reinforcement of hand movements. This result was particularly interesting because the anterior caudate was not involved in the control of hand movements themselves. One explanation for this activation may be due to the nature of the experimental task: when participants performed hand movements, there was scrolling of the stimulus on the screen, while participants had to maintain attention and fixation in the center of the screen. Consequently, on trials when hand movements were reinforced, participants must have succeeded in maintaining their attention centrally to make a correct response. This explanation also fits with the idea that attention shifts are

a prepotent response, and that they occur by default. Note that in the present experiment, the other two conditions (eye movement and covert attention shifts) always involved attention shifts – undesired perseverance of attention shifts from one condition to the next would explain why attention shifts would be prepotent in this particular experiment context.

Another possible explanation for this result is that the anterior caudate is involved in multiple forms of attention, i.e., not only involved in visual-spatial attention, but also, e.g., in attention to motor plans and/or decisions. Indeed, the caudate locus of activation for reinforcement of hand movements is ventral and anterior in the head of the caudate, whereas the loci of activations for the reinforcement of both eye movement and covert attention shifts is most posterior within the caudate head. However, this conclusion is not directly supported by the data because, as just mentioned, hand movements themselves do not involve the anterior caudate during their control.

In light of these results, we proposed that the brain solves the problem of controlling itself in spite of very limited connectivity by using a global reinforcement signal, i.e., the reward-prediction error transmitted by dopamine. A diagram illustrating the framework is shown in Figure 3.9 (page 75). At this point it is important to be clear on why the basal ganglia are required for the present task: presumably, for example, the cortico-cortical projections from sensory areas that detect "red boxes" have very sparse and weak connections to the motor areas that, for example, generate hand movements in various directions. Early in learning, the prefrontal cortex is likely involved in generating a pattern of activation to select various hand movements. However, as particular hand movements are rewarded, the cortico-striatal projections that happen to be active are reinforced. Thus, the dopamine signal is hypothesized to have local effects at the level of cortico-striatal inputs to the basal ganglia loops to adjust the likelihood of performing particular actions in the future. Phasic dopamine therefore plays the role of a conductor that provides a low-dimensional signal to all "players" (i.e., brain regions) involved in performing the behaviors. By broadcasting a positive reward-prediction error signal, the conductor can encourage players to perform the same behaviors in the same context in the future, given their immediate environment. The basal ganglia loops thus serve as a mechanism through which particular "players" can integrate reinforcement from the conductor into their individual performance.

As mentioned earlier, it is also the case that dopamine has effects on cortico-striatal projections

in real-time (Bamford et al., 2004). Thus in addition to its effects via plasticity, dopamine may also serve as a "conductor" in a very real sense, because it can act through D2 receptors on cortico-striatal projections to inhibit the least active synaptic terminals, while presumably increasing the activity of the previously reinforced neocortical action representation.

Reinforcement Learning Theories and Basal Ganglia Function. The framework proposed in this paper is similar to proposals by Schultz (1998; see Fig. 11), Houk et al. (1995), and other contemporaries that described how the conjunction of cortico-striatal and nigro-striatal inputs at the level of the striatum could allow stimuli (and by extension in the current dissertation, action representations) to acquire dopamine responses. These proposals are theoretically important in that the link reward-prediction errors to the phasic activity of dopamine neurons. According to Schultz, the cortico-striatal synapses that represented stimuli that preceded reward can be potentiated or depotentiated based on positive and negative phasic changes in dopamine, respectively. However, Schultz noted two apparent limitations of an apparently global reward signal (e.g., like dopamine) for explaining RL. The first limitation was that dopamine responses "do not explicitly code rewards as goal objects, as they only report errors in reward prediction." The second limitation was the fact that the dopamine signal "lack[s] clear relationships to arm and eye movements, [and this would] disfavor [its] role in directly mediating the behavioral responses that follow incentive stimuli." However, our finding that reward-prediction errors for actions are computed in striatal regions that are otherwise involved in the control of those actions suggests that Schultz' original idea might indeed come into play for neocortical representations of action. That is, our results support the conclusion that dopamine also influences the synapses receiving cortical input from motor and cognitive control areas. With this addition to Schultz' original proposal, the function of cortico-striatal projections of action representations may serve a dual purpose: first, to provide input to a circuit that recurrently amplifies the selection of actions that have previously led to reward, and second, to inhibit the release of dopamine in accordance with the amount of expected reward.

The action-specificity of reward-prediction error computations has important implications for learning theories based on the Rescorla-Wagner learning rule (Rescorla and Wagner, 1972) including RL, suggesting a modification which allows it not only to better explain the present results, but also to account for phenomena such as downshifting during unblocking (e.g., Holland, 1984).

The fact that prediction errors are computed separately for each action suggests that learning is not based on the difference between the global predicted reward and the obtained reward. Instead, the results suggest that learning is based on the difference between reward predictions for *specific actions* and the obtained reward. This move from global reward-prediction error to local reward-prediction error can be implemented as a simple extension to the standard Rescorla-Wagner rule. In this extension, although a global reward-prediction error is still present, separate reward predictions are generated for each stimulus and/or action. Each of these action-specific reward predictions is updated depending on the difference between it and the global reward. As illustrated in an example below, this accounts for how previously unconditioned stimuli do not obtain aversive values during downshifting.

Schultz (1998) also points out that dopamine neurons transmit signals other than reward-prediction error, such as novelty and saliency signals. The intermixing of these signals could be problematic for temporal-difference learning performed using that signal as a reward-prediction error signal (Kakade and Dayan, 2002). However, if the dopamine neurons are involved in RL, i.e. the control of action through reinforcement, it can be shown that novelty and saliency responses of the dopamine neurons do not conflict with reward-prediction error signals, and may even emerge from them (Laurent, 2008, see also Chapter 4, page 124).

The fMRI results in this paper can be directly compared with other similar studies involving reward-prediction error in the basal ganglia. Key studies include those by O'Doherty et al. (2004), McClure et al. (2003), Tanaka et al. (2004), and Haruno and Kawato (2006). These studies provide some of the earliest evidence of BOLD signals in the striatum that were qualitatively and quantitatively related to reward-prediction error. However, in those studies, a puzzling result is that the BOLD signals were often localized to distinct regions within the striatum (e.g., in specific parts of the caudate or putamen). In an early instrumental conditioning task, Pagnoni et al. (2002) found activation correlating with reward-prediction error in the ventral striatum (nucleus accumbens). In a Pavlovian task, McClure et al. (2003) found activity in dorsal striatum. Interestingly, O'Doherty et al. (2004) also found striatal activity in a Pavlovian conditioning task, but that activity was in the ventral striatum. Unfortunately, these researchers did not include a manipulation in their study to determine why particular regions of the striatum would exhibit a BOLD response correlating with

reward-prediction error, a question that is especially salient in view of the fact that dopamine neurons carrying reward-prediction error project to the entire striatum. The results from the present study suggest that the localization of the striatal activation may relate to the task that was being performed. Although in the case of Pavlovian conditioning, it is unclear what "action" was being performed. However, since it is clear that non-motoric "actions" (i.e., covert attention shifts) can be reinforced, perhaps the notion of a cognitive action could be extended to include the act of explicitly assigning values to specific stimuli, e.g., whose representations may be in working memory. Given that anterior striatum is known to be involved in attention, perhaps McClure's results could be explained in terms of attention shifts in working memory.

Cognitive Control Theories. The proposed theoretical framework can be considered a class of cognitive control theory because it suggests how the brain may achieve the selection and control of perceptual input, internal cognitive actions, and motor output. There are a number of other such theories that explain how the brain generates behaviors. We will briefly review some of these and compare them to the current proposal.

Minsky's *Society of Mind* theory is a theory of human cognition that proposes that the mind is composed of independently acting agents that interact (Minsky, 1986). The theory posits a number of basic cognitive operations that can be controlled by each of these independent agents, as well as methods for the agents to intercommunicate. There are also suggested mechanisms for how these agents can compose their behaviors (e.g., hierarchically) into more complex behaviors, and communicate amongst themselves to coordinate their behaviors. However, there was no attempt on the part of Minsky to map the various aspects of the Society of Mind onto specific brain structures.

The Society of Mind theory is similar in spirit to the theoretical framework proposed in this thesis, the main similarity being in claiming the existence of multiple distinct "agents" (i.e., neocortical regions that can be recurrently excited via a cortico-basal ganglia loop). We have connected our framework more directly with existing ideas in RL, and have also made specific claims about what brain structures are involved. Our proposal is agnostic as to the various types of cognitive operations that can be controlled by the various "agents", and just states that the basal ganglia can apply the proposed RL mechanism to any action that can be represented in neocortex. Another difference between the current proposal and the Society of Mind theory is that we do not require

communication between the various "agents" to obtain their apparent coordination, although we do not deny that such communication is ultimately possible. In particular, after large numbers of trials (i.e., much larger than the number of trials in the presented experiments), it is likely that direct cortico-cortical synapses could potentiate sufficiently so as not to require recurrent basal ganglia amplification for the post-synaptic actions to continue to occur. However, the basal ganglia would likely be involved again in any switch from that more "automated" cortico-cortical behavior.

Another class of cognitive control theory focuses specifically on the role of neocortex, particularly on the prefrontal cortex, as part of a cognitive control network (e.g., Hazy et al., 2007; O'Reilly and Frank, 2006). In these theories, the prefrontal cortex is hypothesized to maintain goal and state information in working memory during a task (e.g., Funahashi et al., 1990), but is also thought to be involved in the selection and control of actions. However, we lack a clear explanation for how the prefrontal cortex is able to achieve the selection of specific actions. The relegation of action selection to prefrontal cortex has been termed the homunculus, or a "little man" problem. The basal ganglia mechanism we propose is a solution to the homunculus problem because it does not require that the prefrontal cortex be involved in the direct control of motor and cognitive actions occurring in other regions of neocortex. This is because, according to the framework, the control of those actions is delegated to regions of cortex that have local control of the actions. These regions are then able to act independently using locally available information in order to maximize overall reward for the organism.

This proposal therefore raises interesting questions with regards to how prefrontal cortex could exert direct control over actions, in particular for actions controlled by regions of neocortex to which the prefrontal cortex does not have particularly dense projections. Although it is clear that the prefrontal cortex is able to control dopamine inputs to itself (Carr and Sesack, 2000), it is possible that the prefrontal cortex is also able to control dopamine release in the striatum. For example, the striatum has inhibitory projections to its own source of dopamine innervation (the substantia nigra pars compacta), but the prefrontal cortex targets neurons in the nearby ventral tegmental area that release GABA onto those inhibitory projections. Such a projection would in principle allow prefrontal cortex to cause a basal ganglia-specific release of dopamine when a desired action occurred, thereby reinforcing its occurrence in future contexts (of which prefrontal cortical activity would be

a part). Thus, by acting as a "conductor" wielding a dopamine "baton", the prefrontal cortex could, through trial-and-error, generate patterns of activity that give rise to particular desirable actions or behaviors (e.g., among neocortical brain regions serving as the "players" in an "orchestra") without requiring powerful direct connections to those corresponding neocortical regions. Or, to use another metaphor, the brain can be viewed as a corporation in which a prefrontal cortex "executive" distributes bonuses to all employees in order to generate policies that are in line with overarching corporate objectives, yet irrelevant to the day to day activities of individual employees.

A third class of cognitive control theory is ACT-R, which has been used to explain behavior in a number of cognitive tasks, and the components of which have been recently mapped to specific brain structures (Anderson et al., 2004). ACT-R conceptualizes cognitive control as interactions between separate functionally specialized modules in order to maximize the probability of completing a goal. For example, a "manual module" controls hand movements, a "goal module" maintains information related to current goals, and a visual module identifies objects in the world. These modules are controlled through "production rules", which operate on active subsets of the information available in each of the modules. The productions themselves take the form of if-then statements, and are hypothesized to be implemented by the basal ganglia: the striatum performs a pattern-matching operation (i.e., the "if" part of the production), and then the pallidum and thalamus respectively perform the selection and execution of the resulting actions (i.e., the "then" part of the production). Finally, when actions co-occur either simultaneously or in a sequence with sufficient frequency, so long as the outcome of one action does not require the completion of the preceding action, they can be compiled into a single "production".

Although ACT-R and our framework both attempt to map cognitive control to brain structures in similar ways, there are also a number of points of difference between the architectures. Similarly to ACT-R, we propose that the selection of actions, as well as the competition between actions in different neocortical regions, occurs within the basal ganglia, although we propose that the mechanism is similar to that proposed by Leblois et al. (2006). However, there are a number of important differences between our framework and ACT-R. Because our framework is based on RL, the selection of actions is based on maximizing positive reinforcement for the organism (rather than the probability of successfully completing a goal). This difference is important for two reasons: First,

once reward predictions are equal to the obtained reward, learning effectively stops. In contrast, in ACT-R, learning continues because the probability of success equation keeps accumulating evidence in favor of successes. This leads to differential predictions in learning behavior (e.g., RL models can change their learned reward predictions more quickly because they don t have to work against a large amount of accumulated prior evidence from past experience). Second, the numerical comparison between obtained and received reward is important because it can lead to improved coordination in actions even when a goal can be completed successfully (e.g., this can lead to improved temporal coordination between actions like attention shifts and eye movements to yield improved performance in a reading-like task).

ACT-R proposes that the primary role of the striatum is to perform a pattern recognition function, and this differs markedly from our proposal. The striatum is hypothesized to act as the "reward-weighted" entry to a recurrent feedback system that selects among competing actions in neocortex, i.e., where the recurrent feedback for each action is weighted by reward-prediction error history. A further function of the striatum is to inhibit dopaminergic activity once the corticostriatal projections have be trained to predict the correct amount of reinforcement that is obtained by performing that action (i.e., in that context). Thus, unlike ACT-R, we propose that the pattern recognition that takes place to trigger behaviors occurs at the level of the sparse cortico-cortical connections, and that the basal ganglia are involved in amplifying the effects of those cortico-cortical inputs so that action can be generated.

A final difference between ACT-R and our framework is that ACT-R proposes that sequences of actions can be compiled into productions if they occur together with sufficient frequency. Our framework does not include a mechanism for this capability and therefore cannot explain how, for example, training on one task can lead to improved performance on similar tasks. However, as just mentioned, the use of scalar reward predictions for reinforcement that is based on performance does allow the coordination of actions that are interdependent. For example, the simulations by Rayner et al. (2006) showed how the timing of saccadic eye movements became temporally coordinated with visual attention shifts to the next target to optimize reading rate (an instance of interdependence between vision and attention in that task). Nevertheless, the problem of how to benefit from transfer of strategies across tasks is an unaddressed issue in our framework. One possibility is that the brain

possesses additional machinery (perhaps in neocortex and/or the cerebellum) that contributes to the generalization of particular combinations of actions to novel but similar problems.

Computational Insights. The present study demonstrates that reward-prediction error computations are localized to distinct regions of the striatum that are known to be involved in particular anatomical loops. This piece of information provides additional constraints in forming a mapping between RL computations and the basal ganglia. That is, because signatures of reward-prediction error computations are spatially localized to each loop depending on the type of action being reinforced, it is reasonable to conclude that the loops themselves are somehow important in the relevant computations. This leads to a number of theoretical points. The first is that the use of eligibility traces to mark particular actions is a less tenuous idea, because those traces may be differentially spatially localized depending on the action that just occurred. Large neural networks like the brain may use these eligibility traces to mediate the reinforcing effects of the reward-prediction error on synaptic plasticity. Eligibility trace mechanisms have been shown to be sufficiently powerful to give rise to action-outcome learning in a "plain vanilla" (i.e., randomly-connected) neural network (Izhikevich, 2007). However, Izhikevich's network required many trials to learn. It would be interesting to investigate whether the neural architecture suggested by the parallel cortico-basal ganglia connectivity pattern might accelerate such learning. One would predict that because action representations are relatively segregated in these loops, the effects of reinforcement on the simulated synapses in a relatively "quiet" (i.e., low baseline activity, less interference from other action inputs) will accumulate more rapidly, and thereby increase the rate of learning.

The second neurocomputational point is that, even though the proposed "globally-reinforced, locally-controlled" mechanism is relatively simple, it is powerful enough to gives rise to efficient control of complex behaviors because it can result in the coordination of perceptual, cognitive, and motor processes. One example of this can be found in work by Reichle and Laurent (2006), in which RL agents were tasked with learning to control lexical processing, attention shifts, and saccadic eye movements in order to "read" as efficiently as possible in a simplified reading-like task. Although lexical processing, attention shifts, and saccadic eye movements were implemented as independent actions in these simulations, the RL agents gradually learned to perform the actions to the point where they became strongly coordinated. In particular, through learning, the comple-

tion of a specific amount of lexical processing triggered the initiation of a saccadic eye movement programming, so that the eyes would move precisely when attention shifted to the next word. This emergence of coordination between independently controlled actions came about purely as a function of task demands and a non-specific (i.e., global) reinforcement signal related to the RL agent's behavioral performance. These simulation results thus provide a demonstration of how a single reinforcement signal can lead to the coordination of perceptual (e.g., word length), cognitive (e.g., lexical processing, attention shifting), and motor (e.g., saccade programming) processes. Therefore, one can conclude that the reinforcement of actions in particular cortico-striatal loops as is functionally demonstrated in the experimental portion of this paper is a powerful mechanism that could lead to the coordination of arbitrary actions into intelligent behaviors.

The third theoretical point is that, by mapping RL to the basal ganglia, one gains an understanding of the function of the striatum in both reward-related and novelty/saliency-related processing. A number of studies have suggested that striatal activity (and positive phasic dopamine activity) conveys information that is apparently unrelated to reward, such as novelty and saliency: in particular, positive signals are seen for salient, novel, and/or unexpected stimuli, surprisingly regardless of whether those stimuli lead to positive or negative reward. These signals appear to be different from the reward-prediction error signal because they also appear for aversive stimuli (Horvitz, 2000; Salamone, 1994), and often appear very rapidly before the stimuli that triggered the signal have been fully identified (Redgrave and Gurney, 2006; Redgrave et al., 1999). Generally, these responses accompany orienting behavior to the stimulus (Schultz, 1998), and subside when the orienting behavior no longer occurs. Given the apparent discrepancy in information content between reward and saliency, it seems perplexing that the same system might represent both. However, recent simulations demonstrate that these novelty/saliency responses can in fact be shown to emerge from RL principles (Laurent, 2008). These simulations show that an agent that must orient towards a sudden-onset stimulus in order to determine the rewarding or punishing nature of that stimulus will always exhibit a positive reward-prediction error, even if, half of the time, the stimulus onset predicts the appearance of a negatively-valued object. The explanation for this paradoxical result was that the dopamine signal is related to the actual reinforced behaviors: In cases where the agent later perceives that the object was negatively valued, it does not approach it to collect and receive negative reinforcement (see Laurent, 2008, Figure 2). Thus, the reward-prediction when the stimulus

appears is the average of positive outcomes (collected rewards) and near-zero outcomes (avoided punishments). These computational findings suggest that the apparently non-reward-related signals conveyed by dopamine are in fact compatible with RL, and may in fact emerge from RL principles operating in basal ganglia brain circuits.

Regressor	Observed Location	Number of Voxels	Voxelwise Threshold	Peak Coord.	Peak Statistic
Covert Attention					
Positive RPE	Anterior caudate*	60	$p < 0.01$	-8,7,7	t=8.65, p<0.001
		55	$p < 0.01$	11,7,16	t=7.72, p<0.001
Negative RPE	Anterior caudate*	21	$p < 0.04$	-8,13,0	t=-3.82, p<0.002
Movement	Anterior caudate/putamen[1]*	19	$p < 0.03$	14,13,4	t=3.30, p<0.003
		28		17,10,4	t=3.60, p<0.002
Saccades					
Positive RPE	Anterior caudate*	19	$p < 0.01$	-5,4,4	t=3.34, p<0.003
		10	$p < 0.01$	11,7,16	t=2.34, p<0.018
	Dorsal caudate*	6	$p < 0.01$	17,-6,20	t=3.56, p<0.002
		4	$p < 0.01$	-17,1,20	t=2.92, p<0.006
Negative RPE	Anterior caudate*	25	$p < 0.2$	17,16,7	t=1.72, p<0.055
	Nucleus accumbens	12	$p < 0.2$	20,16,-6	t=2.17, p<0.025
		11	$p < 0.2$	-23,16,-3	t=1.66, p<0.061
	Dorsal putamen*	7	$p < 0.2$	23,4,13	t=1.66, p<0.061
Movement	Anterior caudate*	21	$p < 0.08$	-17,4,-16	t=2.71, p<0.009
Right Hand					
Positive RPE	Anterior caudate	22	$p < 0.2$	-5,13,7	t=2.25,p<0.022
	Left putamen*	6	$p < 0.2$	-20,1,13	t=1.76,p<0.051
		10	$p < 0.2$	-23,-6,-3	t=-2.11,p<0.027
Negative RPE	Left putamen*	5	$p < 0.12$	-20,-2,-3	t=-2.55,p<0.013
	Nucleus accumbens	8	$p < 0.12$	-5,13,0	t=-2.29,p<0.02
	Right putamen	5	$p < 0.12$	27,-15,-3	t=-2.38,p<0.017
Movement	Left putamen*	30	$p < 0.05$	-27,-18,-3	t=3.51,p<0.002

Figure 3.2: **Coordinates of the peaks of largest and most significant clusters correlating to positive and negative reward-prediction error and movement regressors in the striatum.** Generally, the pattern of activity supports the hypothesis that reward-prediction errors for particular actions are computed in striatal regions to which neocortical regions controlling those actions project. Predicted activations are indicated by an asterisk (*). Although reward regions are similar to those involved in movements, positive reward-prediction error signals were generally more dorsal compared to the movement-related signals. Interestingly, the anterior caudate was involved in positive reinforcement of hand movement, suggesting that some aspect of covert attention, although not necessarily a covert attention movement, was reinforced. This may correspond to reinforcing the maintenance of covert attention (and fixation) to the center of the screen during the scrolling motion that results from the hand movement. Coordinates are with reference to the Talairach-Tournaux atlas. (Abbreviations: RPE=reward-prediction error; Note (1): The two ROIs are likely a single caudate ROI centered in the white matter between the putamen and the caudate (Feekes and Cassell, 2006).

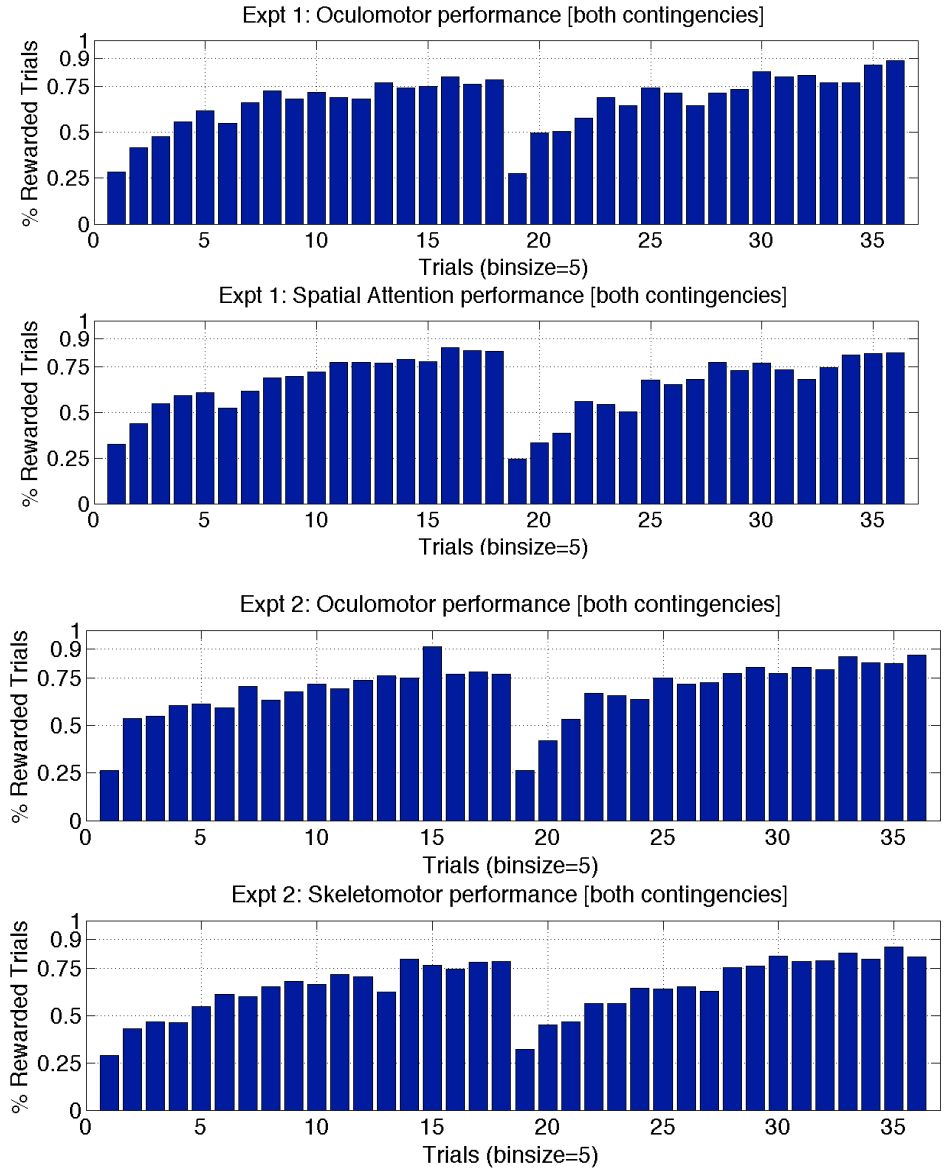

Figure 3.3: **Behavioral Performance of Participants as they Learn Eye Movements, Covert Attention Shifts and Hand Movements.** The following two plots show participants learning to perform the different actions during the experiment. Each bar represents the average number of trials rewarded, across participants, for five trial bins. Chance performance is 25%; theoretical optimal performance in this task is 90% (indicated on the x-axis) because rewarded and unrewarded directions were stochastically inverted during 10% of the trials.

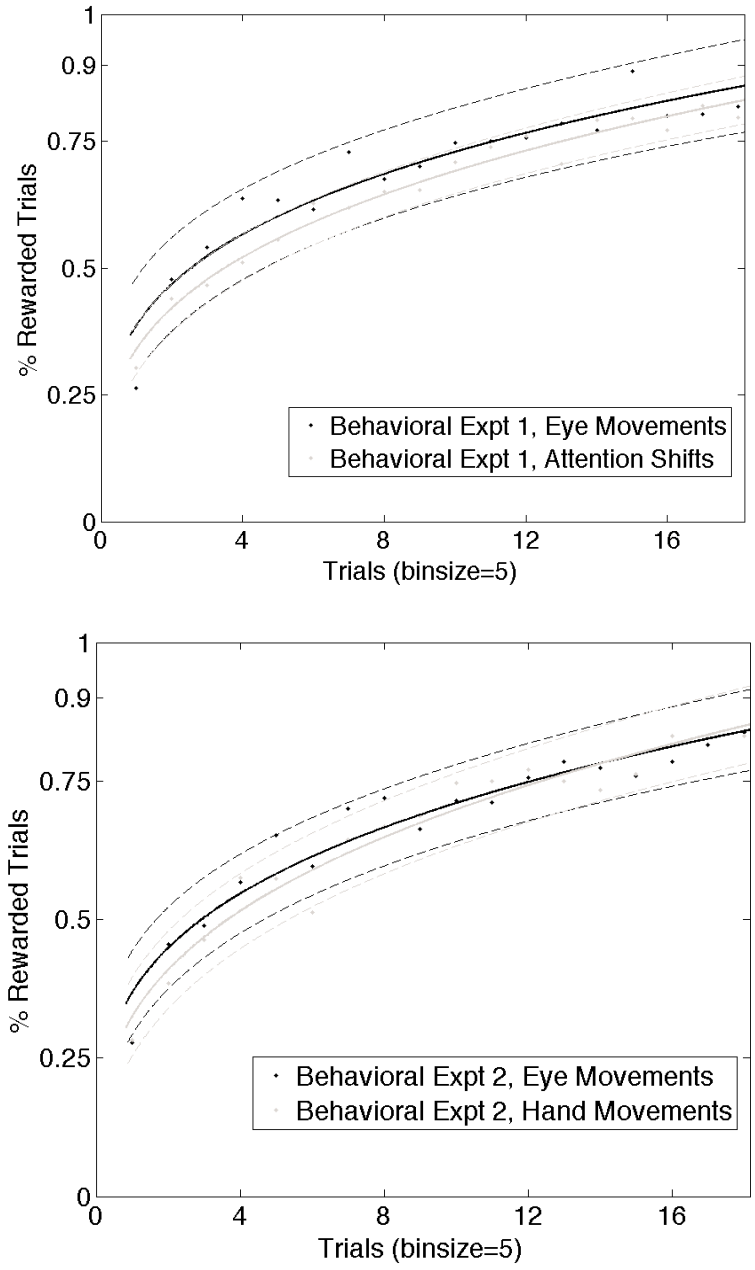

Figure 3.4: **Power Law Curve Fits to Pooled Behavioral Performance Data.** Data from the first and second halves of each behavioral experiment were pooled and a power law curve was fit to the percent correct. The fitted curve is plotted in solid, and 90% confidence bands are plotted in dotted lines. Note that the curves are very close to each other and the confidence bands overlap. (Refer to the text for the curve parameters and their confidence intervals.)

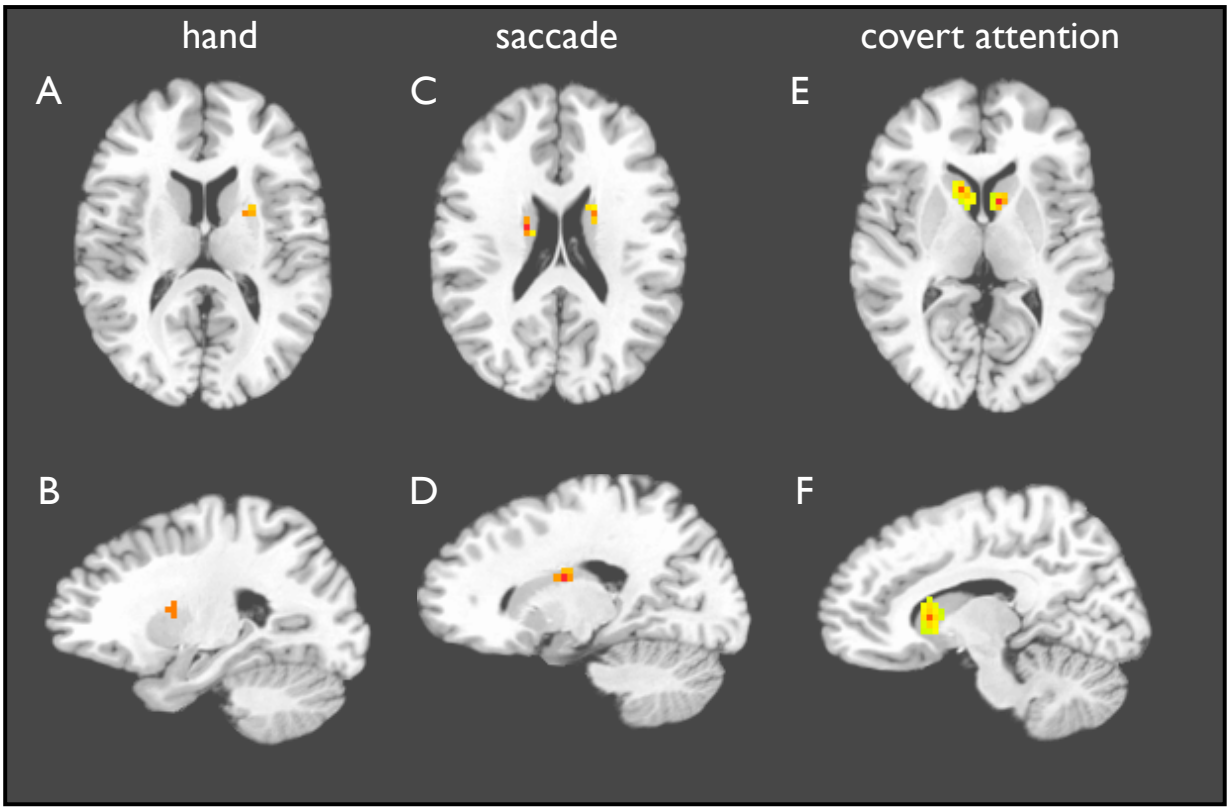

Figure 3.5: **Striatal Regions Correlating Positively with Positive Reward-Prediction Error while Reinforcing Hand Movements, Saccadic Eye Movements, and Covert Attention Shifts.** Panels A-B show a region in the left putamen that correlated with positive reward-prediction error while hand movements were being reinforced. Panels C-D show bilateral regions in the anterior caudate that correlated with positive reward-prediction error while saccadic eye movements were being reinforced. Panels E-F show bilateral regions in the anterior caudate that correlated with positive reward-prediction error while covert attention shifts were bring reinforced. See Table 1 for complete list of striatal regions correlating with reward-prediction error and movement events.

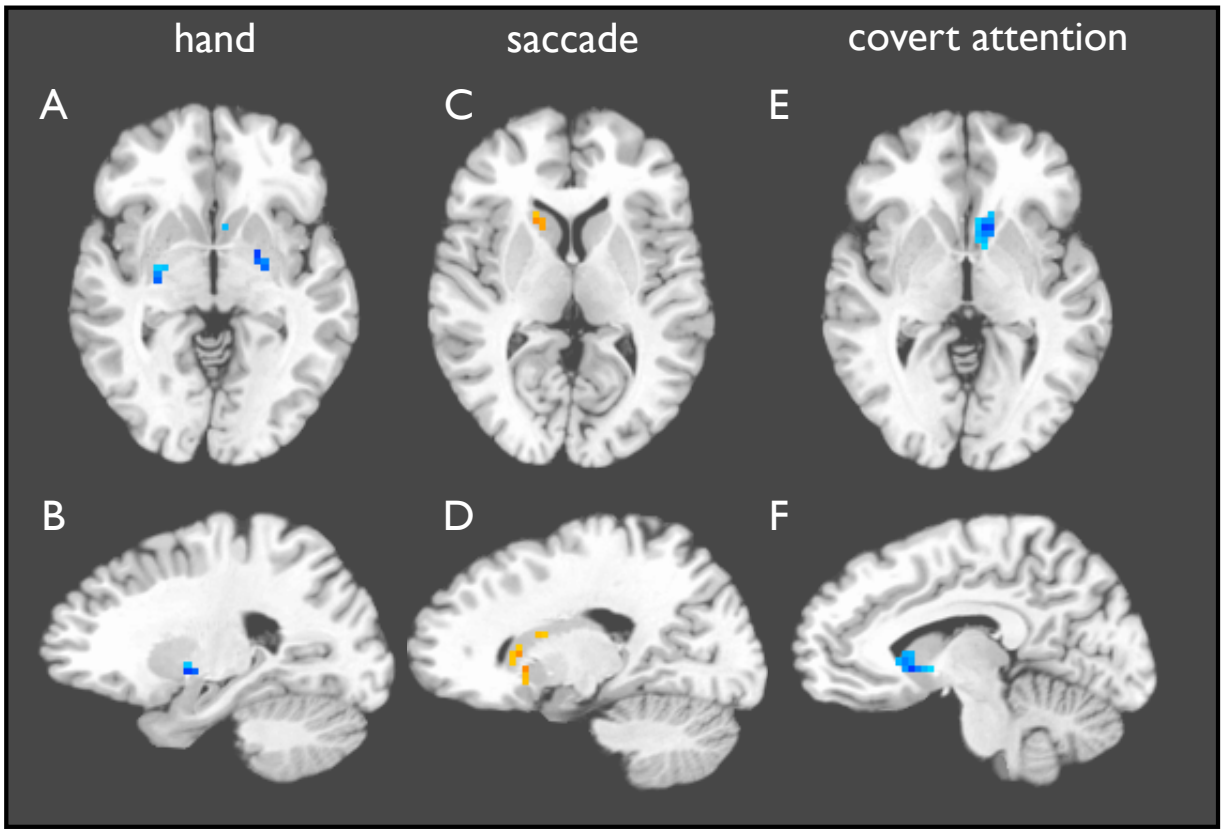

Figure 3.6: **Striatal regions correlating with negative reward prediction error.** Regions correlating with negative reward-prediction error were generally smaller, less statistically significant, and more diffuse compared to regions correlating with positive reward-prediction error. Panels A-B on the left show bilateral regions in the putamen that correlated with negative reward-prediction error while hand movements were being reinforced (note that the correlation was negative). Panels C-D show regions of the anterior caudate that correlated with negative reward-prediction error while saccades were being reinforced. Panels E-F show the region in the anterior caudate that correlated with negative reward-prediction error when covert attention shifts were being reinforced.

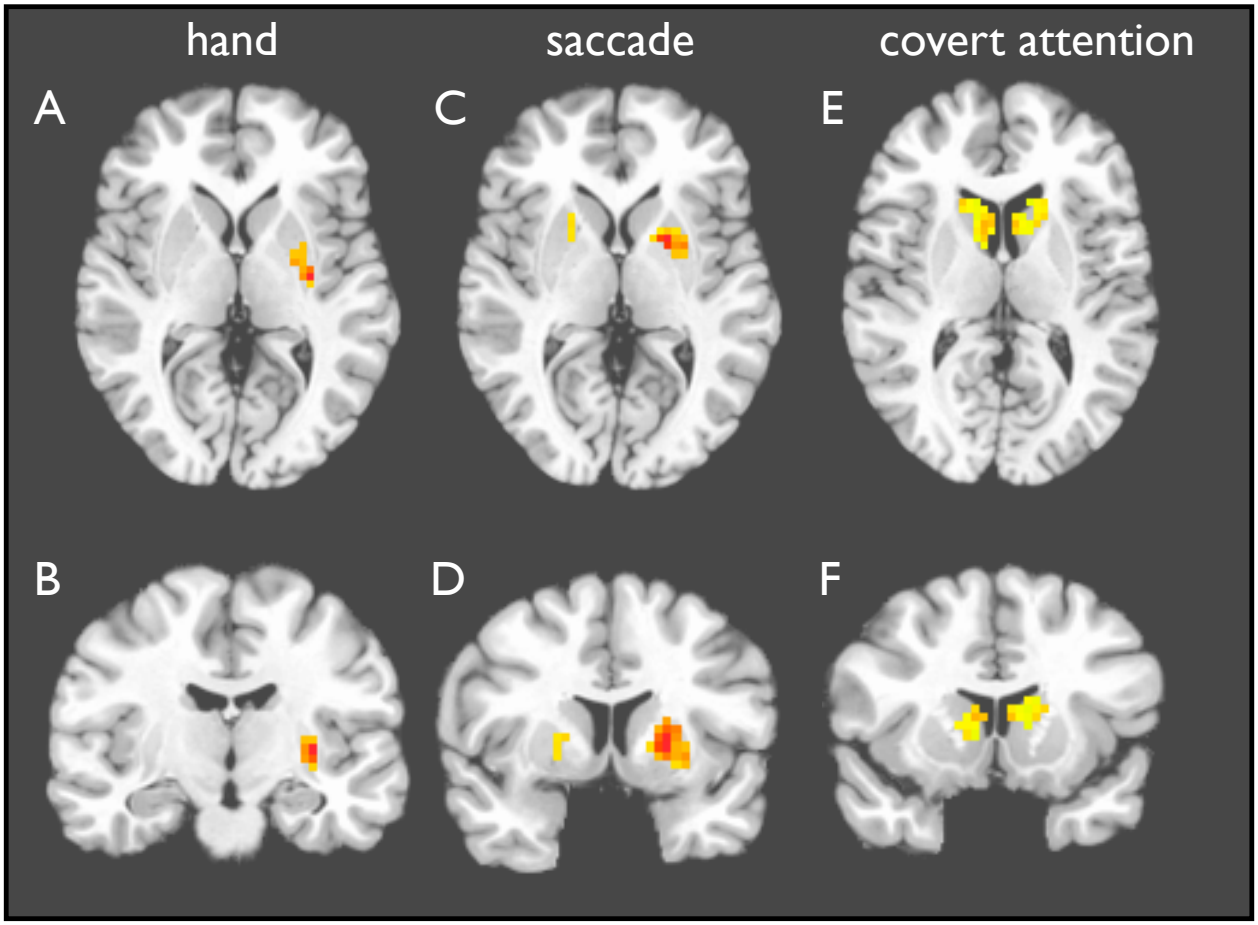

Figure 3.7: **Striatal Regions Correlating Positively with Movements for Attention Shifts, Saccades, and Hand Movements.** These figures show data from the CONTROL runs of the experiment only, during which no reward was present. Panels A-B show the striatal region (left putamen) that correlated with right hand movements. Panels C-D show regions of the striatum (anterior putamen) that correlated with saccadic eye movements. Panels E-F show regions of the striatum (anterior striatum, dorsal caudate) that correlated with attention shifts.

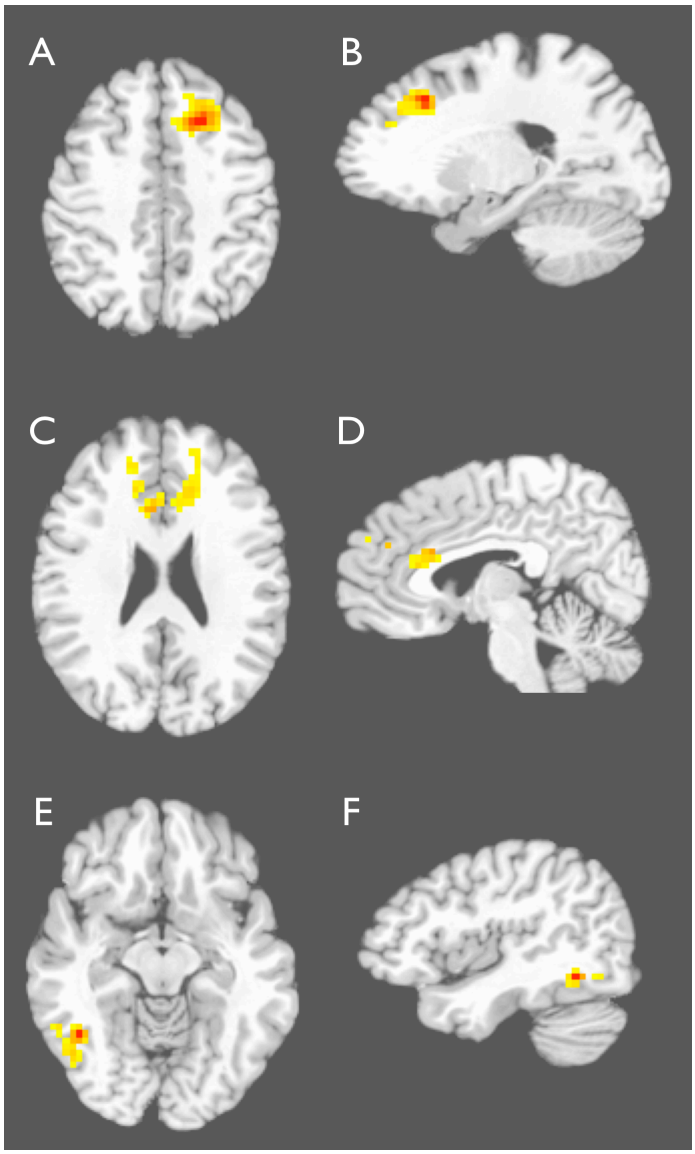

Figure 3.8: **Prefrontal Cortex Activity Decreases with Reinforcement.** This figure shows the peaks of the regions resulting from an ANOVA to which the means of each quarter of each learning run (hand movements, eye movements and covert attention shifts) were submitted. The image is thresholded at $p < 0.001$ and corrected for multiple comparisons at $p < 0.0001$. The regions are (A) dorsal-medial prefrontal cortex (BA8 peak at -14mm, 23mm, 36mm [$F(1,55) = 13.0$]; BA9 peak at 5mm, 41mm, 26mm, not shown [$F(1,55) = 8.73$]), (C,D) the anterior cingulate cortex (peak at 5mm, 19mm, 23mm [$F(1,55) = 9.78$]), and (E,F) the right middle temporal gyrus (peak at 42mm, -52mm, -9mm [$F(1,55) = 12.86$]).

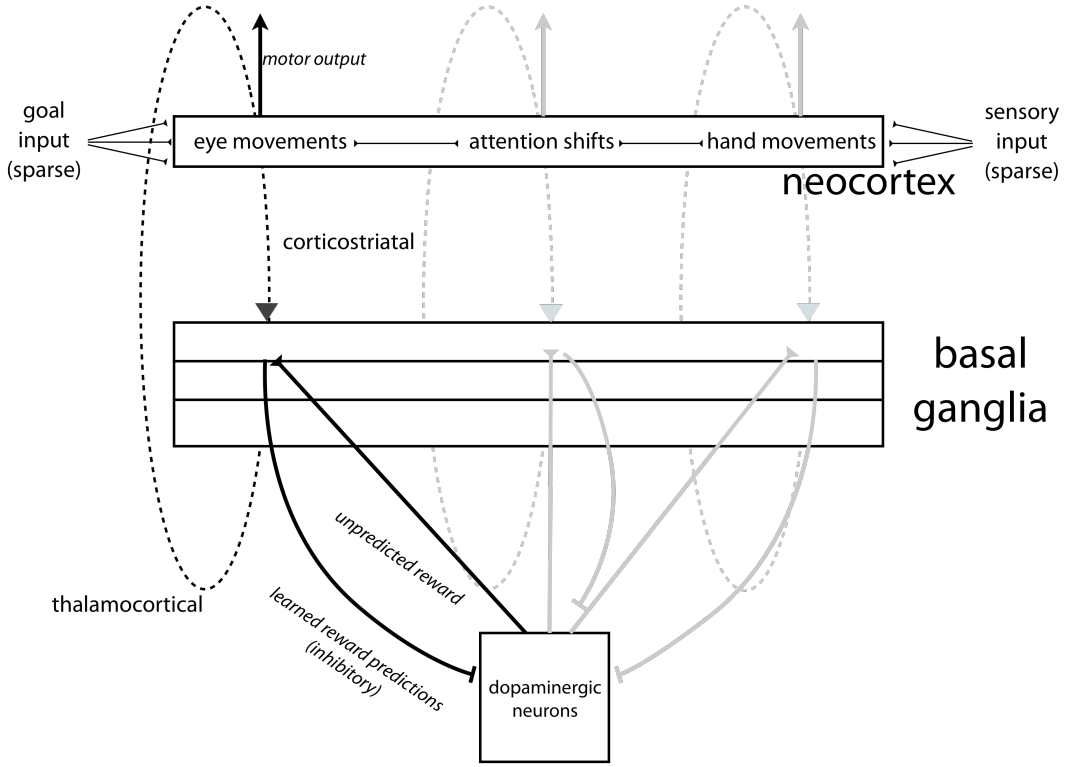

Figure 3.9: **Diagram of proposed framework.** Sparse cortico-cortical connectivity allows goal (from the left) and/or sensory (from the right) information to propogate through neocortex. However, because this connectivity is very sparse, it is assumed that, in many cases, this afferent activity is not sufficient to initiate action (i.e., not sufficient strong or temporally coordinated to reliably depolarize the desired quantity of neurons). Some subset of the desired configuration of neurons are activated and project to the striatum. To the degree that those connections had been strengthened through prior reinforcement, they will engage a spatiotemporal feedback loop. Through this loop, associated neocortical neurons will also be excited, eventually generating the desired action. Other feedback loops that may have been reinforced, but perhaps less so, will also begin to excite themselves. Local organization within the neocortex could provide efficient competition among competing actions. When unexpected reinforcement is received, dopamine potentiates corticostriatal projections for recently activated actions using an eligibility-trace mechanism (e.g., as used by Izhikevich, 2007). The striatum also has inhibitory efferents to the dopamine-releasing neurons. Plasticity at the level of the striatum is hypothesized to store the reward prediction associated with the neocortical action representation, and striatonigral inhibitory connections can inhibit dopamine release in proportion to the predicted reward, thereby computing reward-prediction errors.

3.3.10 CONCLUSION

We presented a theoretical framework, along with supporting experimental evidence that explains how the brain is able to control itself and generate intelligent, adaptive behavior in spite of constraints on connectivity. The framework proposes that the control of particular actions is delegated to local regions of cortex, but that a global reinforcement signal reinforces those actions when they lead to favorable outcomes for the organism as a whole. The implementation of this framework involves the basal ganglia, which receive information from regions of the neocortex, and then outputs information back to those regions of the neocortex to trigger specific actions. The proposed framework is that the basal ganglia apply the same computation to all afferent information, namely, the computation of reward-prediction errors for the neocortical patterns of activity that generate "action". According to the framework, if neocortical activity patterns generate actions leading to more or less reward than was predicted, then the projections from those patterns to the striatum were altered in proportion to the reward-prediction error.

We also presented data in support of this framework. In particular, we found fMRI BOLD signals indicative of the computation of reward-prediction errors were localized to distinct regions of the striatum depending on the type of action that was being reinforced. Most importantly, we found that these regions corresponded generally to regions of the striatum known to be involved in the control of the actions being reinforced. Furthermore, we found that this result was true whether the actions being rewarded are physical (i.e., hand movements and eye movements), or cognitive (i.e., covert attention shifts). By demonstrating that reward-prediction errors are computed in distinct parts of the striatum for different actions, these findings provide a functional explanation for the anatomical finding that the basal ganglia have parallel connections with both motor and non-motor regions of neocortex. In particular, the results suggest that reward-prediction error computations can be mapped to the cortico-basal ganglia loops, and that these loops mediate a pattern of connectivity through which a global reinforcement signal can be applied to locally-controlled actions in the brain.

These findings have implications for models of learning based on the Rescorla-Wagner learning rule. In particular, these findings suggest that although global reward prediction errors are important, action-specific (and by generalization, stimulus specific) reward-prediction errors are

computed separately and in parallel. Critically, learning can be done based on the difference between received reward and action-specific reward expectations, rather than global reward expectations. Adopting this extension to the Rescorla-Wagner rule provides a better explanation for the present data (because it explains how reward-prediction errors are located in different parts of the striatum), and it explains previously problematic phenomena in behavioral RL such as unblocking and downshifting. This is discussed in more detail in the section below entitled **Extending Rescorla-Wagner-based Learning Rules**, Section 3.6.1 on page 113.

These findings also have a number of implications for future research. Neurophysiologically, the present results suggest that reward-prediction error computations are not a property of a specific region of the dorsal versus ventral striatum, but rather suggest that evidence of reward-prediction error should be observable in all parts of the striatum depending on the type of neocortical activity that is being reinforced. Multiunit recording in the striatum, coupled with tasks that reinforce specific types of action that engage reinforcement learning, in particular cortico-basal ganglia loops, may generate insight into how plasticity due to reward-prediction errors at the level of the striatum is used to ultimately recurrently "boost" neocortical activity patterns that lead to reward.

From the standpoint of future computational model development, the present findings suggest how to solve two types of computational problems. The first problem is that of performing RL in large state spaces. RL algorithms are afflicted by the curse of dimensionality the state space grows dramatically as the dimensionality of the problem to be solved increases (Sutton and Barto, 1998). Although researchers have applied a number of strategies to address this problem, such as sampling only parts of the space (i.e., trajectory sampling) or by attempting to use function approximation, the solution suggested by the current results is that some measure of success could be achieved by dividing the state space into sub-spaces that can be tackled by (largely) independent agents. Unfortunately, our data do not suggest how to optimally divide state spaces into sub-state spaces. However, organizing sensory and motor representations based on the statistics of the environment could be an important function of neocortex itself (e.g., Földiák, 1990; Oja, 1982) by collecting related and competing actions into the same cortico-basal ganglia loops. Thus, the independent parallelized RL agents can operate on subsets on the state space to decide among competing actions rather than on the entire state space faced by the organism. The second problem to which these

findings propose a solution was mentioned in the General Discussion: how to achieve control in large simulated neural networks. As computer technology improves and researchers study increasingly large neural networks, attempts to successfully control them may benefit from biomimicry at this level.

From a clinical standpoint, the finding that the reinforcement of actions is mediated through changes in different parts of the striatum suggest a potential means for early diagnosis of basal ganglia disorders, especially if those disorders happen to disproportionately affect one region of the striatum more than another (e.g., Huntington's disease; Vonsattel et al., 1985). An inability to reach asymptotic performance, or a lack of fit by RL agents for a particular type of action compared to others could serve as an early indication of the loss of cells in, or of dopamine innervation to, a particular striatal region. Another potential diagnostic could be based on the possibility that tasks that engage single parts of the striatum (e.g., hand movements or covert attention shifts, in the current results) are more sensitive to the loss of striatal dopamine function than those that possibly recruit several striatal areas simultaneously (e.g., saccadic eye movements, in the current results.)

3.4 SUPPLEMENTAL RESULTS

Two main analysis methods are used in this section. The first method is deconvolution analysis, which is equivalent to multiple regression analysis with coefficients for each point in time. Group-level statistics are then computed based on the sum (over time) of the resulting regression coefficients. To provide the reader with a qualitative understanding of deconvolution analysis, a brief example in which two simulated signals are convolved into a single "voxel" and then recovered through deconvolution is presented in Appendix C.

The second analysis method is a direct application of analysis of variance (ANOVA) to the mean voxel activity (after normalization) from particular trials. This latter approach is quite common in fMRI analysis and is not demonstrated with an example.

3.4.1 MOVEMENT-ONLY MODEL

The data from the CONTROL (i.e., non-reinforced) portion of the experiment were examined for signals correlating to hand movements, saccadic eye movements, and covert attention shifts. Binary event time series were created for each of the three types of action performed by participants. These time series were used as inputs for a deconvolution analysis performed for 11 time points (i.e., TRs) in the time range -4...16s around the time of the movement event. The first 6 beta coefficients were then summed to create a voxel-wise beta map for each type of movement event, and this map was then spatially smoothed with a 5mm full-width half-max Gaussian filter. (For an illustration of the summated coefficients with respect to the events of interest, see Figure 3.10 on page 81.) These maps were then transformed using ANTS (as detailed in Appendix A) to bring them into a common space. Finally, a group-level *t*-test was performed across the participants' maps to reveal areas whose activity correlated to the occurrence of the modeled movement actions.

IN THE STRIATUM

Striatal involvement in the three types of movements was examined. To restrict analysis to the striatum, a hand-drawn striatal mask shown in Figure 3.11 (page 84) was applied to the group-level *t*-map. The results of this analysis are tabulated in Table 3.1 and illustrated in Figures 3.12-3.14. The results show that three separate regions in the striatum were involved in the control of the three different kinds of movements. Figure 3.12 shows that right hand movements involved the left (i.e., contralateral) posterior putamen. Figure 3.13 shows that saccadic eye movements involved the anterior portions of the putamen and parts of the caudate. Finally, Figure 3.14 shows that covert attention shifts involved the head and the most anterior portions of the caudate.

Although the activated regions were generally in line with the aforementioned hypotheses (briefly, that right hand movement would involve the left putamen, that saccadic eye movements would involve anterior and dorsal striatum, and that covert attention shifts would involve the anterior caudate), the main unexpected result was that the dorsal caudate was not apparently involved in saccadic eye movements, as would be predicted given anatomical projections there from the frontal eye fields. One possible explanation for this is that the dorsal caudate is a very small and narrow part of the striatum, and that activity there could have been obscured by noise from the nearby ventricles. Another possibility is that the dorsal caudate is involved in rewarding voluntary saccades

rather than in the cued execution of saccades. This latter possibility will be addressed in the subsequent analyses which examine reward-prediction error signals during saccadic eye movements.

Event-related averages (i.e., impulse response functions) were computed for the movement-related regressors. These are plotted in Panel D in Figures 3.12-3.14. Although the regions depicted in those figures were regions of maximal activity during the first 6 beta-weights (i.e., time range -4...6 seconds), these plots show that the signal is maximal during the time range of 0...6. Thus in the subsequent deconvolution analysis, that restricted time period (i.e., TR2 - TR5) was used as the time-range during which movement-related signals were hypothesized.

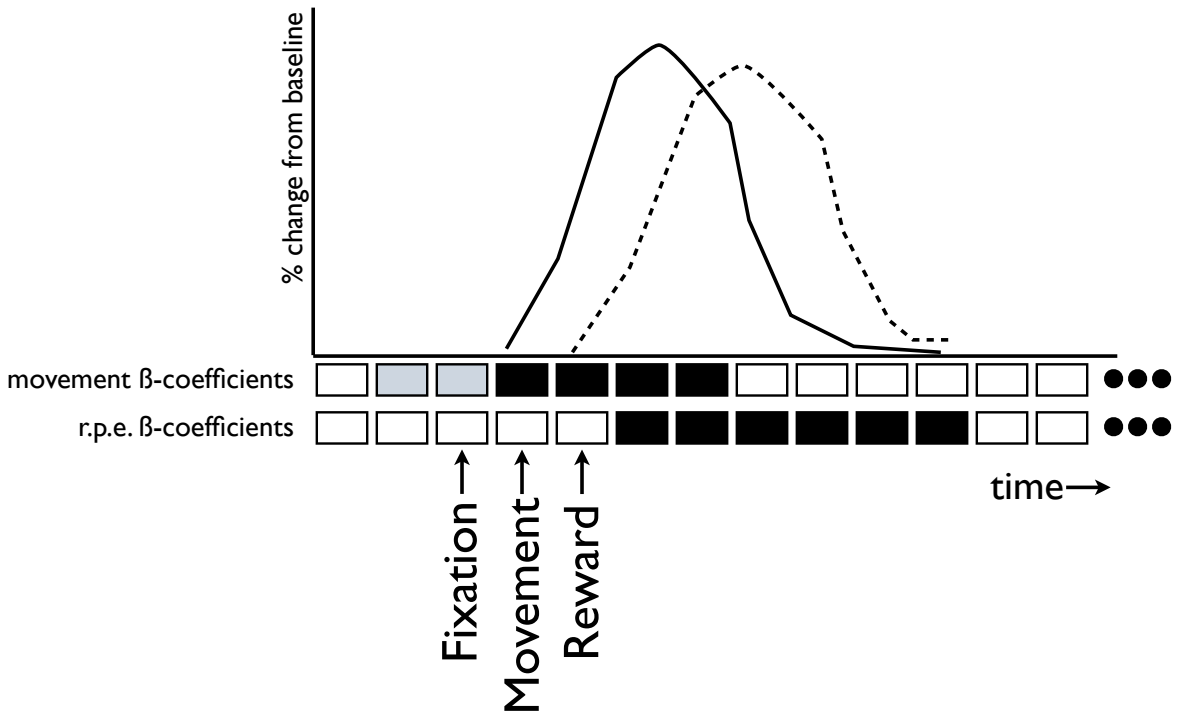

Figure 3.10: **Illustration of Regressors in the Deconvolution Analyses**. Deconvolution involves specifying multiple regression (beta) coefficients, one for each data point of interest (e.g., scan image or TR). This figure illustrates those regression coefficients with respect to the events of interest (i.e., the movement event and the reinforcement event, indicated by arrows). Six (6) regression coefficients were summated for the movement signals in the movement-only model analysis (i.e., the two gray and 4 black boxes in the row labeled "movement β-coefficients"). In the movement+reinforcement model analysis, only the regression coefficients indicated by the black boxes were summated. As indicated in the text, deconvolution itself spanned 12 TRs starting from the onset of each event. Note that although the hypothesized hemodynamic responses from the events of interest overlap in time, deconvolution analysis should allow the successfully separation of their individual contributions. A demonstration of successful separation is shown in Appendix C with simulated data possessing similar properties to the real data set (i.e., movement signals with assumed constant amplitude, but reinforcement signals with variable amplitudes).

Region	Size	Peak Coord.	Peak t-value	Corrected p-value
Hand Movements				
(*voxelwise p < 0.05*)				
Posterior Putamen	40 voxels	-30,-12,4	t=3.66	p<0.0001
Eye Movements				
(*voxelwise p < 0.0045*)				
Anterior Putamen	71 voxels	-20,4,4	t=6.32	p<0.0001
	27 voxels	23,13,0	t=4.75	p<0.0001
Covert Attention				
(*voxelwise p < 0.02*)				
Anterior Caudate	44 voxels	8, 10, 10	t=4.31	p<0.0001
	42 voxels	-5, 10, 10	t=4.15	p<0.0001

Table 3.1: **Striatal Regions active during CONTROL trials.** Complete list of striatal regions involved in hand movements, saccadic eye movements, and covert visual-spatial attention. These regions survive correction for multiple comparisons at a threshold of $p < 0.0001$.

Region	Size	Peak Coord.	Peak t-value	Corrected p-value
Visual cortex	782 voxels	-23,-77,-12	t=7.09	p<0.0001
	145 voxels	27,-87,16	t=4.33	p<0.0001
Primary motor cortex	172 voxels	-55,-24,39	t=5.34	p<0.0001
	35 voxels	52,-15,39	t=3.95	p<0.001
Inferior Temporal Cortex	36 voxels	55,-9,-22	t=4.42	p<0.0007

Table 3.2: **Regions Correlating with Hand Movements during CONTROL trials.** Complete list of regions active during hand movements at a voxelwise threshold of $p < 0.01$, extent threshold 39 voxels for a corrected threshold of $p < 0.0001$.

Region	Size	Peak Coord.	Peak t-value	Corrected p-value
Visual cortex	1780 voxels	-45,-46,-16	t=5.54	p<0.0001
Anterior putamen	262 voxels	-20,4,4	t=6.32	p<0.0001
	192 voxels	27,-2,0	t=4.86	p<0.0001
Frontal Eye Fields	74 voxels	-39,-9,36	t=4.39	p < 0.0001
	60 voxels	36,-12,32	t=4.17	p < 0.0001
Inferior Parietal Lobule	66 voxels	39,-59,42	t=4.94	p < 0.0001
Posterior Cingulate	52 voxels	-8,-46,29	t=3.72	p < 0.0001

Table 3.3: **Regions Correlating with Eye Movements during CONTROL trials.** Complete list of regions active during eye movements at a voxelwise threshold of $p < 0.015$, extent threshold 51 voxels for a corrected threshold of $p < 0.0001$.

Region	Size	Peak Coord.	Peak t-value	Corrected p-value
Visual Cortex	296 voxels	23, -74, -3	t=9.92	p<0.0001
	62 voxels	-23, -87, 0	t=5.08	p<0.0001
Anterior Insula	160 voxels	-23, 29, 7	t=6.33	p<0.0001
Superior Parietal Lobule	104 voxels	-23, -56, 32	t=5.34	p<0.0001
	60 voxels	17, -65, 39	t=4.87	p<0.0001
Right Thalamus	72 voxels	20, -21, -3	t=6.254	p<0.0001
Culmen	63 voxels	-27, -52, -28	t=5.17	p<0.0001
	49 voxels	36, -52, -35	t=4.98	p<0.0001

Table 3.4: **Regions Correlating with Covert Attention Shifts during CONTROL trials.** Complete list of regions active during covert attention shifts at a voxelwise threshold of $p < 0.006$, extent threshold 31 voxels for a corrected threshold of $p < 0.0001$.

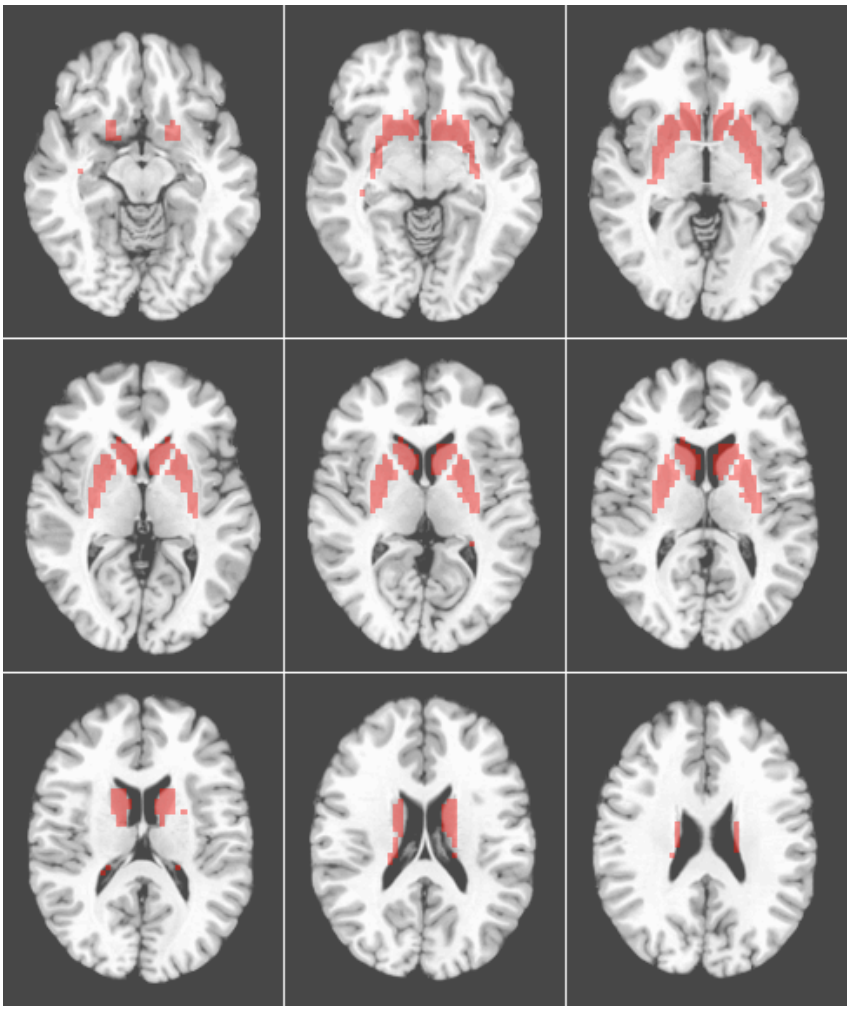

Figure 3.11: **Striatal Mask.** A mask based on the caudate and putamen masks built into AFNI's drawing tools were used for group analyses restricted to the striatum.

IN THE NEOCORTEX AND OTHER SUBCORTICAL STRUCTURES

Hand movements. The entire brain was then examined for regions correlating significantly with movements during CONTROL trials. As can be seen in Figure 3.15 (page 90, an extensive region in the contralateral (i.e., left) primary motor cortex was observed to be correlated with hand movements. This activity reflects muscular control of hand movements as participants moved the joystick during the experiment. The figure also shows that there was extensive visual activation during hand movements, likely due to the scrolling of the screen when hand movements occurred.

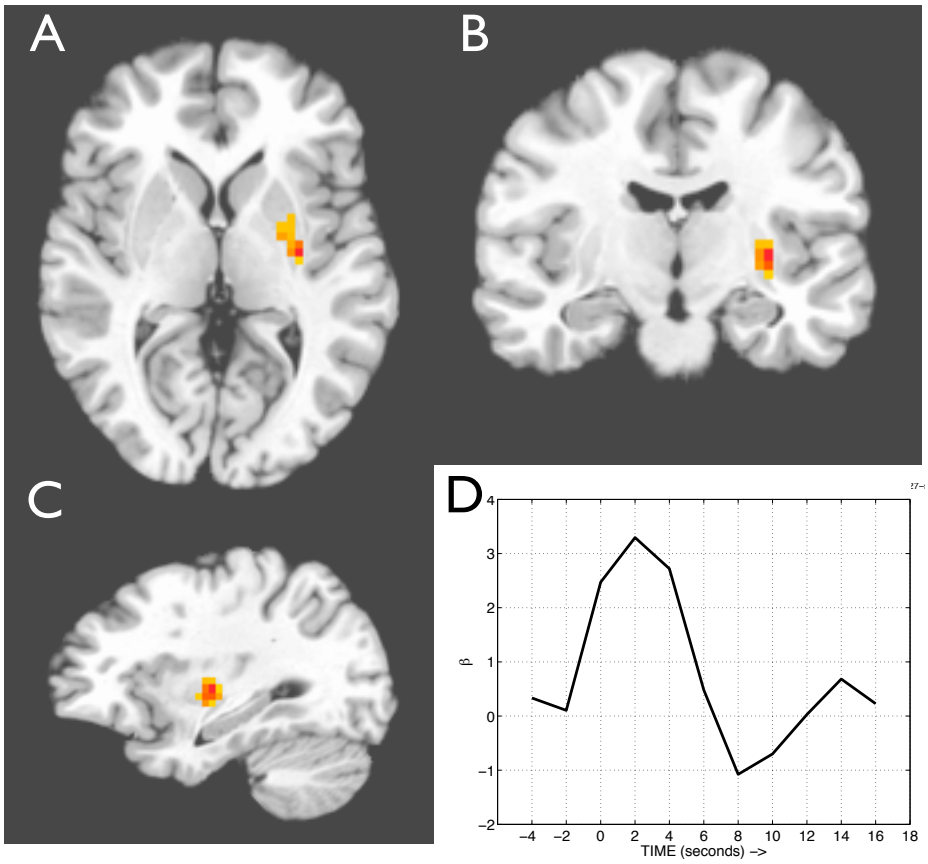

Figure 3.12: **Striatal regions involved in hand movements during CONTROL trials.** Panels **(A)**, **(B)**, and **(C)** show axial, coronal, and saggittal views of the region in the left putamen that was strongly correlated with hand movements. **(D)** Average deconvolved impulse response function, i.e., regression coefficients, for all voxels depicted in Panels A-C (y-axis: percent change from average voxel activity for all modeled runs; x-axis, seconds; t=0 is the time of the color cue stimulus onset). [Voxelwise threshold $p < 0.05$, extent threshold of 40 for a corrected threshold of $p < 0.0001$]

Although it was predicted that there would be distinct activation of a region analogous to region MT during this task, such activation was not observed distinct from general visual activation. One possible reason for this was insufficient power – and thus, this hypothesis will be tested again when the full data set is analyzed (i.e., CONTROL and LEARNING trials combined).

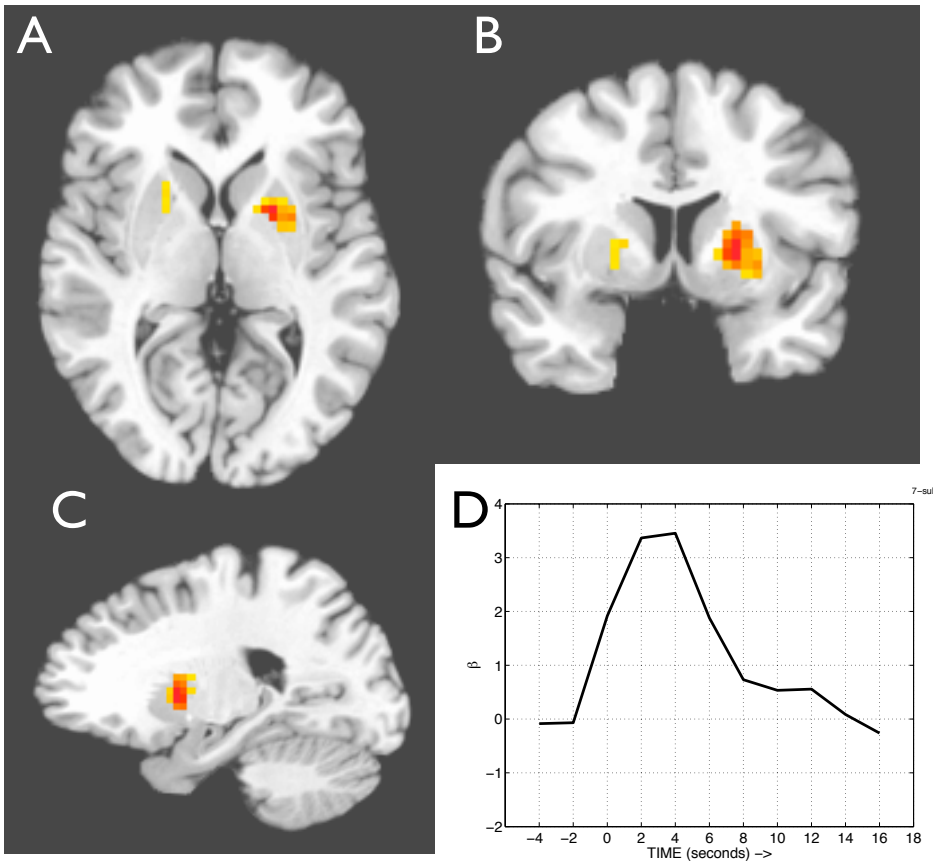

Figure 3.13: **Striatal regions involved in eye movements during CONTROL trials.** Panels **(A)**, **(B)**, and **(C)** show axial, coronal, and saggittal views of the region in the bilateral anterior putamen regions that were strongly correlated with hand movements. **(D)** Average deconvolved impulse response function, i.e., regression coefficients, for all voxels depicted in Panels A-C (y-axis: percent change from average voxel activity for all modeled runs; x-axis, seconds; t=0 is the time of the color cue stimulus onset). [Striatal mask; Voxelwise threshold $p < 0.0045$, extent threshold of 13 for a corrected threshold of $p < 0.0001$]

Continuing with examining hand movement-related activity, activation of the contralateral (left) thalamus was observed at a lower threshold. This activation is shown in Figure 3.16 (page 91). These results suggest that the control of hand movements involves activation through the thalamus.

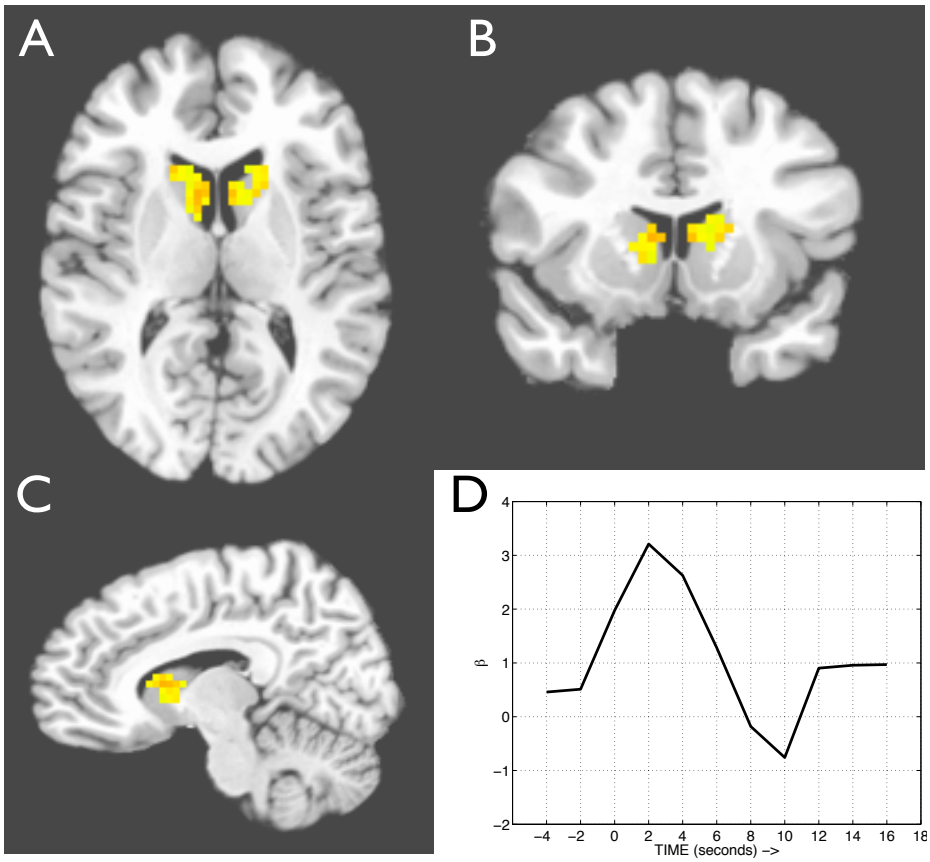

Figure 3.14: **Striatal regions involved in covert attention shifts during CONTROL trials.** Panels **(A)**, **(B)** and **(C)** show axial, coronal, and saggittal views of the bilateral anterior caudate regions that were strongly correlated with covert shifts of visual-spatial attention. **(D)** Average deconvolved impulse response function, i.e., regression coefficients, for all voxels depicted in Panels A-C (y-axis: percent change from average voxel activity for all modeled runs; x-axis, seconds; t=0 is the time of the color cue stimulus onset). [Striatal mask; Voxelwise threshold $p < 0.02$, extent threshold of 25 for a corrected threshold of $p < 0.0001$]

Saccadic eye-movements. Next, whole-brain activity was examined for saccadic eye movements. As can be seen in Figure 3.17 (page 92), there was bilateral activations of a lateral frontal cortical region on both sides of the brain which is a likely human analogue of the frontal eye fields (i.e.,

FEF). Also shown in panel D of that figure is weak activation that was observed in the substantia nigra. Activity in the substantia nigra is consistent with prior neuroscientific knowledge showing that the substantia nigra pars reticulata is involved in the control of saccadic eye movements by the basal ganglia. The resolution of the present scans is not sufficient to distinguish between activations in pars reticulata and pars compacta .

Covert attention shifts. Finally, whole-brain activity was examined for covert attention shifts. As can be seen in Figure 3.18 (page 93), covert attention shifts activated bilateral extrastriate cortex, BA 7 (medial superior parietal lobule) and precuneus. A smaller frontal eye field region (i.e., smaller compared to saccadic eye movements) was also active during attention shifts at a lower voxelwise threshold of $p < 0.01$ uncorrected. This 49-voxel cluster survived a whole-brain correction for multiple comparisons at $p < 0.0001$.

3.4.2 MOVEMENT + REINFORCEMENT MODEL

The data from both the CONTROL (i.e., non-reinforced) and LEARNING (i.e., reinforced) portions of the experiment were examined for signals correlating to movements as well as action-specific positive and negative reward-prediction errors. To complete this analysis, the sequence of binary events for hand movements, saccadic eye movements, and covert attention shifts were generated for the entire length of the experiment. Then, as described in the text of the preceding article, an RL agent was provided the same sequence of stimuli, movements, and reinforcement seen by each human participant.

These time series were used as inputs for a deconvolution analysis performed for 11 time points (i.e., TRs) in the time range -4...16s around the time of the movement event. Given shapes of the impulse-response functions obtained in the movement-only model analysis, only 4 beta-weights in the time range 0..8s relative to the onset of the event were summed to create a voxel-wise beta map for each type of movement event. For positive and negative reward-prediction error events, 6 beta-weights in the time range 2s..14s were summed to create voxel-wise beta maps for action-specific positive and negative reward-prediction error. (For an illustration of the summated coefficients with respect to the events of interest, see Figure 3.10 on page 81.) This analysis therefore yielded 9 maps, crossing three action modalities with three responses: (hand, saccade, attention) X (movement,

positive reward-prediction error, negative reward-prediction error). Each map was again spatially smoothed with a 5mm full-width half-max Gaussian filter. These maps were then transformed using ANTS (as detailed in Appendix A) to bring them into a common space. Finally, a group-level *t*-test was performed across the participants' maps to reveal areas correlated to the movements.

IN THE STRIATUM Although the previous (movement-only) analysis examined signals correlating with movement (i.e., Section 3.4.1 on page 79), the Movement+Reinforcement analysis allowed an examination of movement-related activation as well. There were several reasons to examine movement-related activity in this analysis. One reason is that it might reveal a clearer picture of movement related activity than was possible during the CONTROL portions of the experiment because more trials were used. The results of this analysis are displayed in Figure 3.19 (page 95). These results show that movement-related activity during CONTROL+LEARNING was relatively consistent, although more spatially focused, compared to the results of the analysis during CONTROL only. The form of the hemodynamic response functions for each region of interest were computed from the regression (beta) coefficients from the deconvolution analysis. These are plotted in Figure 3.20 (page 96). Note that in this figure, the units of the y-axis are in percent change from baseline for each event of interest, i.e., signal change for each movement or for one cent ($0.01) of reward-prediction error.

Striatal regions correlating with positive reward-prediction error signals were then examined. The results of this analysis are shown in Figures 3.21, 3.22, and 3.23. As can be seen in Figure 3.21, the positive reinforcement of hand movements correlated with the left putamen and the anterior caudate. Figure 3.22 shows that the positive reinforcement of saccadic eye movements correlated both the dorsal caudate and the anterior caudate. Figure 3.23 shows that the positive reinforcement of covert attention shifts correlated with the anterior caudate.

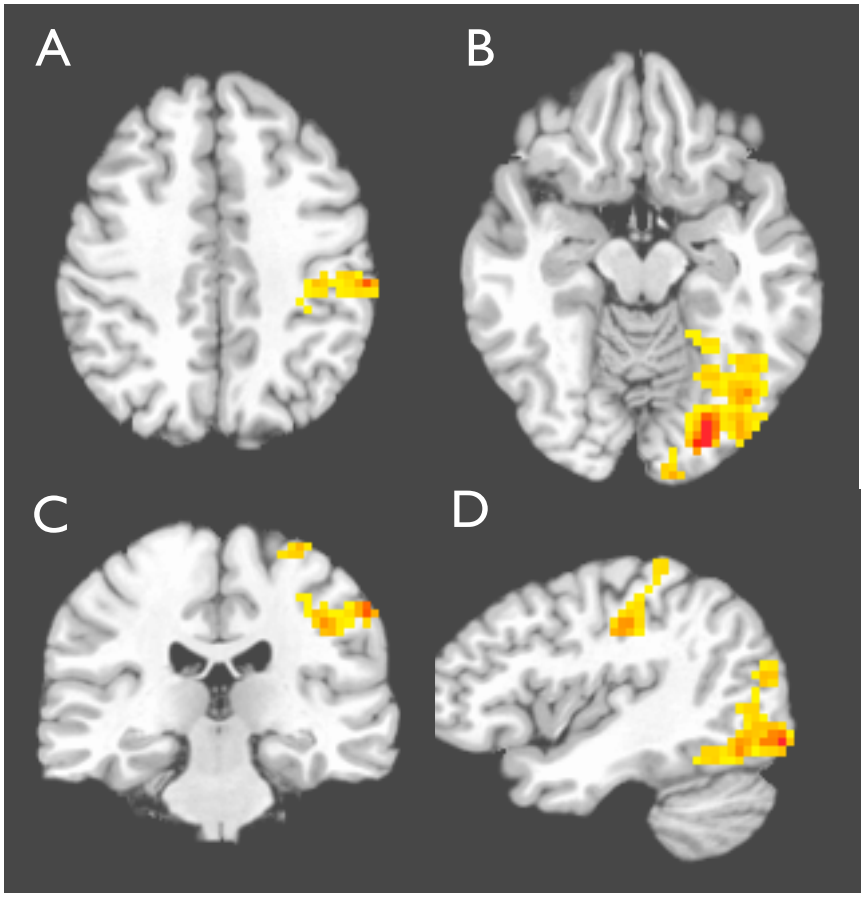

Figure 3.15: **Whole-brain correlations for hand movements during CONTROL trials. A.** An axial view through the dorsal cerebral cortex supporting the hypothesis that there was significant, extensive activation in the contralateral left primary motor cortex correlating with right-hand movements (172 voxels, peak *t*-value=5.34, coordinate=-55mm, -24mm, 39mm). **B.** An axial view through ventral-occipital cerebral cortex showing that there were extensive visual activations likely due to visual input during the task (782 voxels on left, peak *t*-value=7.09, coordinates=-23mm, -77mm, -12mm; 145 voxels on right, peak *t*-value=4.33, coordinate=27mm, -87mm, 16mm). **C.** A coronal view of the activations shown in panel A. **D.** A saggittal view showing both motor (top) and visual (bottom) activations. [Voxelwise threshold was $p < 0.01$ and cluster extent threshold was 39 for a corrected threshold of $p < 0.0001$.]

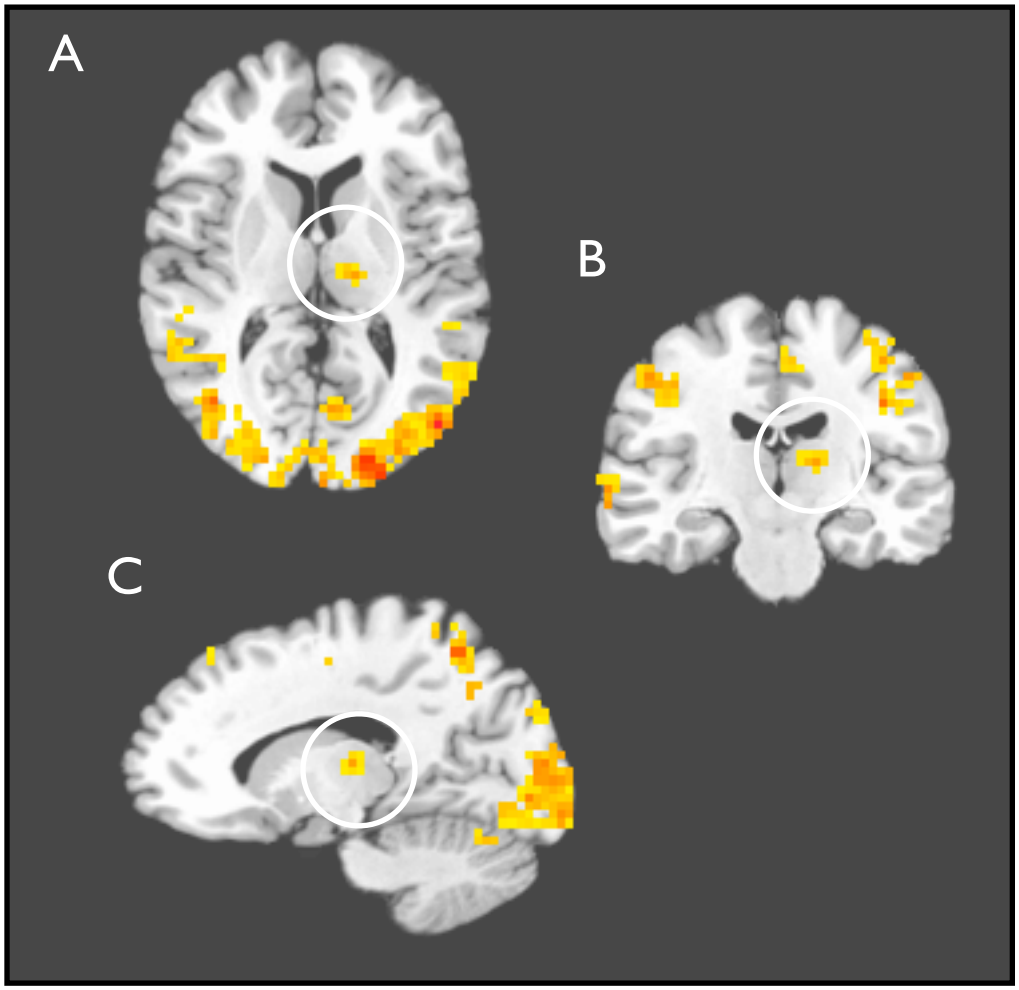

Figure 3.16: **Left Thalamus Activation for Right-Hand Movements.** At a less conservative threshold the left thalamus was correlated with right hand movements, as indicated in these **(A)** axial, **(B)** coronal, and **(C)** saggittal images by a white circle. The peak voxel of this 17-voxel region was located at -14mm, -18mm, 10m with t = 3.681 and did not survive corrections for multiple comparisons. [Voxelwise threshold was $p < 0.02$ and cluster extent threshold was 17.]

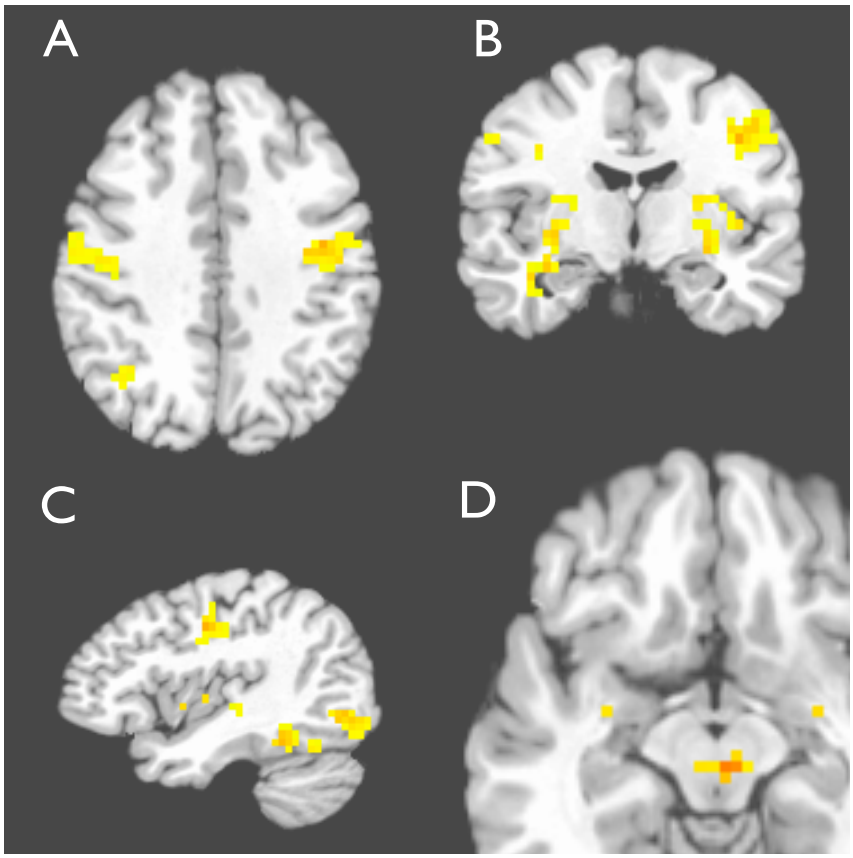

Figure 3.17: **Whole-brain correlations for eye movements during CONTROL trials.** There were extensive visual activations likely due to visual input during the task (not shown). **(A,B,C)** As hypothesized, bilateral frontal eye field activations were observed (74 voxels on the left, peak *t*-value=4.39, coordinate=-39mm, -9mm, 36mm; 60 voxels on the right, peak *t*-value=4.17, coordinate=36mm,-12mm,32m. In addition, likely reflecting the involvement of covert attention in saccadic eye movements, a region in the superior parietal lobule was also active (66 voxels on the right, peak *t*-value=4.94, coordinate=39mm,-59mm,42mm. Regions shown in A-C survived corrections for multiple comparisons at $p < 0.0001$. [For Panels A-C, voxelwise threshold was set to $p < 0.015$ and cluster extent threshold was 51 voxels for a corrected threshold of $p < 0.0001$] **(D)** Activation in substantia nigra was also observed for eye movements (17 voxels at $p < 0.0062$, peak *t*-value=5.12) but this region did not survive correction for multiple comparisons.

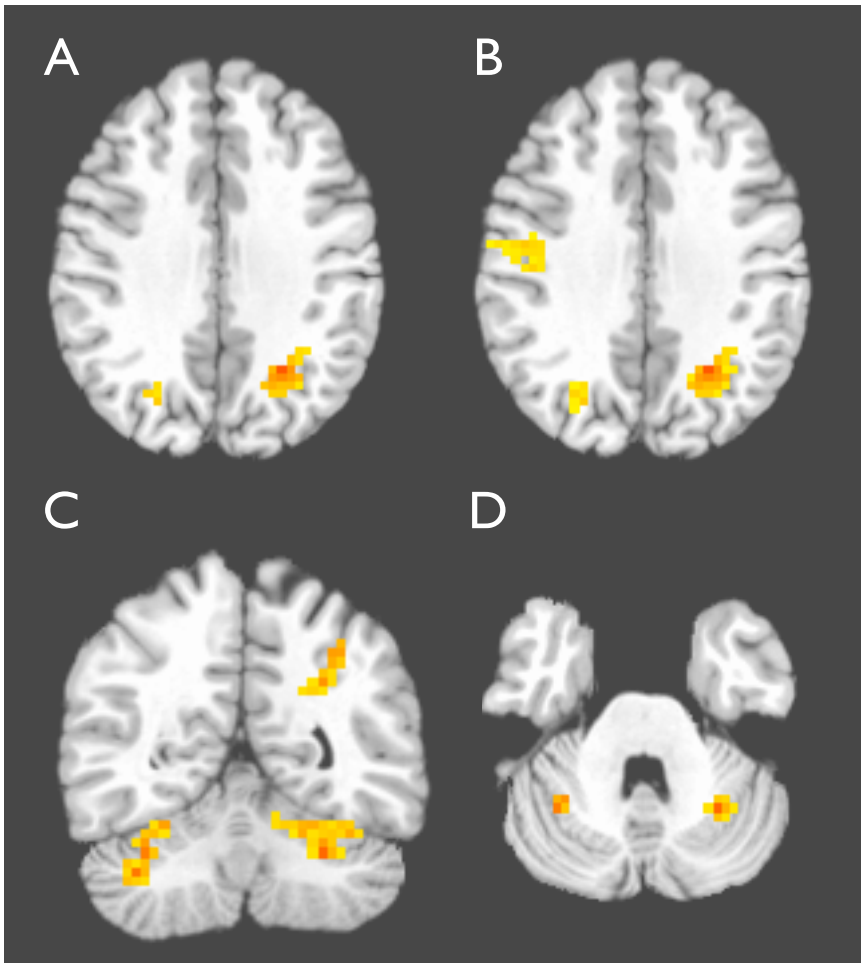

Figure 3.18: **Whole-brain correlations for covert attention shifts during CONTROL trials. (A)** An axial view through dorsal cerebral cortex shows regions in the superior parietal lobule (i.e., precuneus) that were involved in covert shifts of visual-spatial attention reported in table 3.4 (page 83). **(B)** Same as (A), but at a less conservative voxelwise threshold of $p < 0.01$ some activation of the frontal eye fields (FEF) in the right hemisphere was also seen for covert attention shifts. **(C-D)** Coronal and axial views showing bilateral cerebellar activation correlating with attention shifts. In addition, there were extensive visual activations likely due to visual input during the task (not visible in these particular views). [Voxelwise threshold for A, C and D was set to $p < 0.006$ and cluster extent threshold was set to 31 voxels for a corrected threshold of $p < 0.0001$.]

Region	Size	Peak Coord.	Peak t-value
Hand Movements			
(*voxelwise p < 0.05*)			
Left Putamen*	30 voxels	-27,-18,-3	t=3.51
Eye Movements			
(*voxelwise p < 0.08*)			
Anterior Striatum*	21 voxels	-17,4,-16	t=2.71
Dorsal Caudate*	-	-	-
Attention Shifts			
(*voxelwise p < 0.03*)			
Anterior Putamen*	28 voxels	17,10,4	t=3.60
Anterior Caudate*	19 voxels	14,13,4	t=3.30

Table 3.5: **Movement Signals in the Striatum.** Complete list of striatal regions involved in hand movements, saccadic eye movements, and covert visual-spatial attention during both CONTROL and LEARNING trials. Asterisks (*) indicate regions expected to be involved given prior anatomical knowledge.

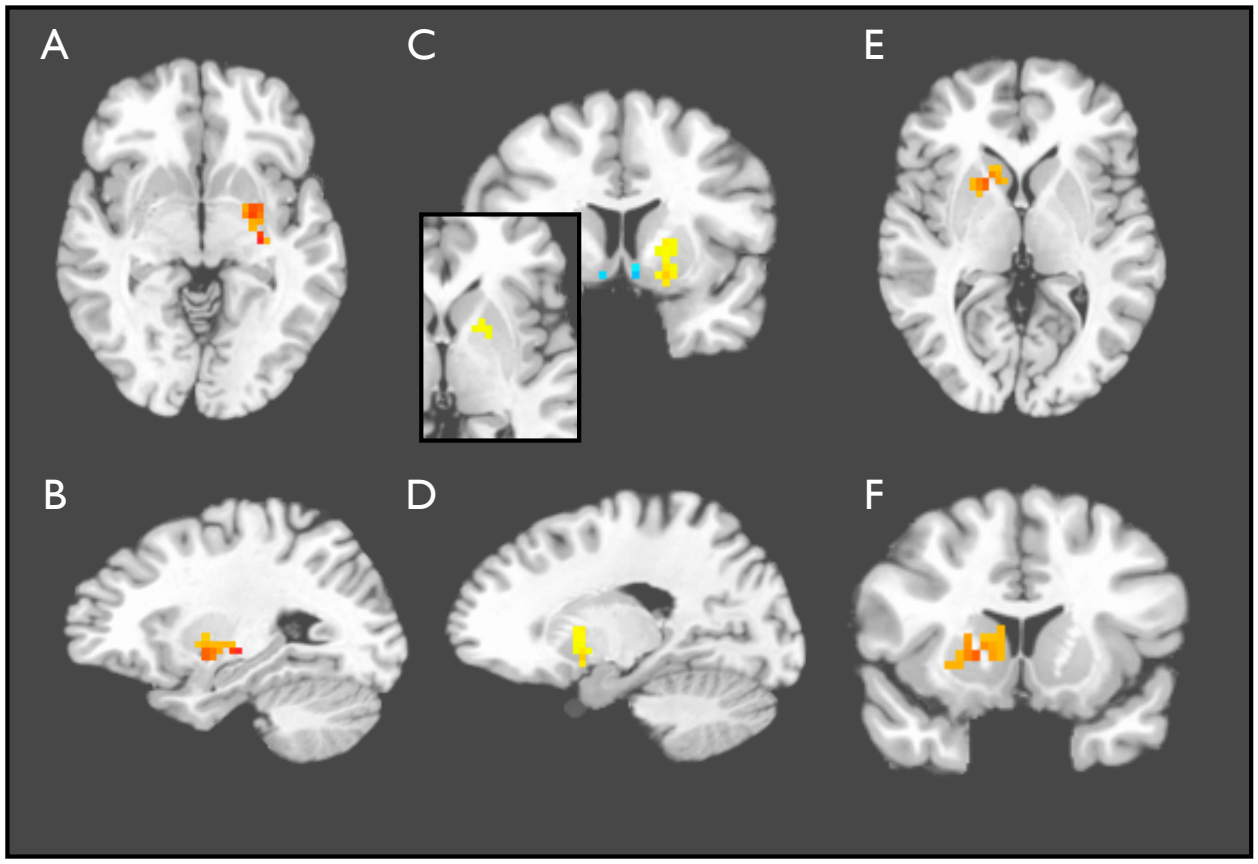

Figure 3.19: **Movement-related Striatal Activity during entire experiment.** Although Figures 3.12 - 3.14 show movement signals without the presence of reinforcement, it was interesting to also compare those results to movement signals deconvolved across the entire experiment including trials with reinforcement. Interestingly, the results are qualitatively similar, although not identical. The discrepancies may indicate mechanistic differences in basal ganglia function depending on whether they are controlling movements based on cues or based on reward predictions. **(A-B)** Axial and sagittal views showing right-hand movement signals in the left putamen. **C-D** Coronal and sagittal views (axial view inset) showing anterior-dorsal extent of a putamen region involved in saccadic eye movements. **E-F** Axial and coronal views of caudate and putamen regions involved in the control of covert attention shifts.

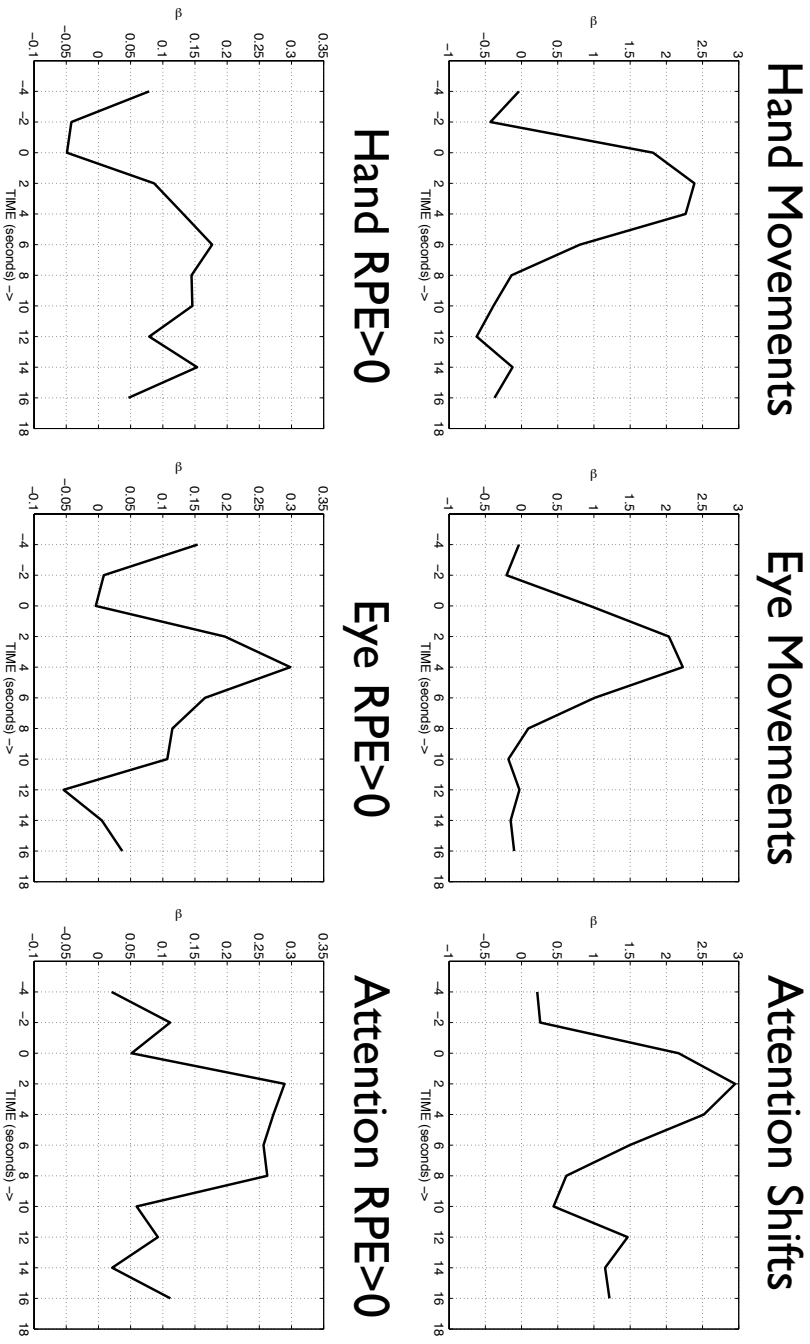

Figure 3.20: **Deconvolved Impulse Response Functions.** This figure plots the computed shapes of the BOLD responses for movements and positive reward-prediction errors for the voxels in the ROIs that correlated with the signal of interest. X-axis is time (0 marks the onset of movement or reinforcement). Y-axis units are beta, i.e., percent change from baseline for one movement event, or for one cent ($0.01) in the case of reward-prediction error. See text on page 89.

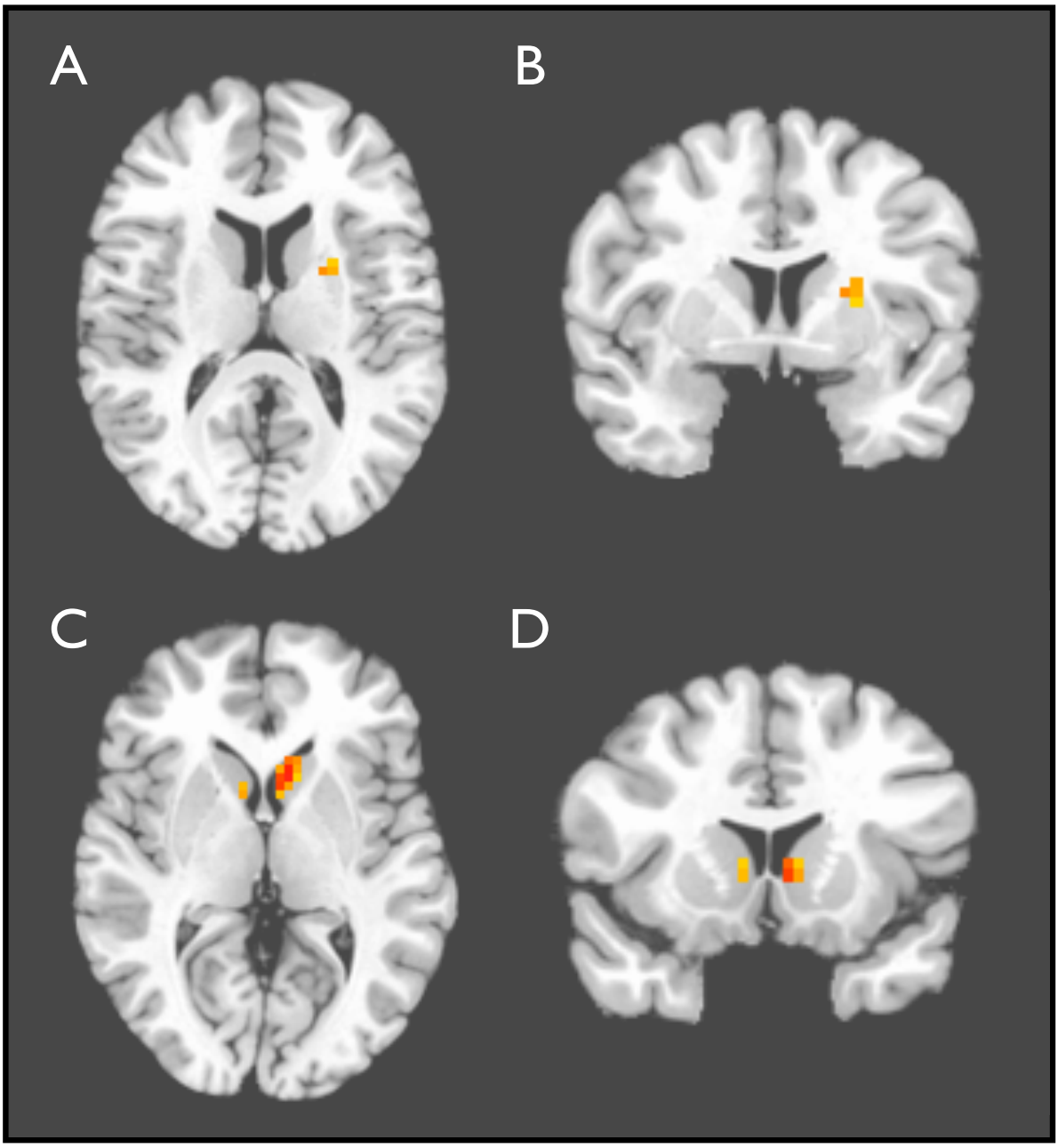

Figure 3.21: **Striatal Regions Correlating with Positive Reward-Prediction Error for Reinforced Right-Hand Movements. (A-B)**The left putamen was correlated with the unpredicted positive reinforcement of right hand movements. **(C)**. Some negative correlation was also observed in ventral putamen. **(D)** Activation in anterior caudate. This activation suggests that some aspect of attention was reinforced when hand movements were reinforced, i.e., maintaining covert attention at the center of fixation. Data tabulated in Table 3.6 on page 104.

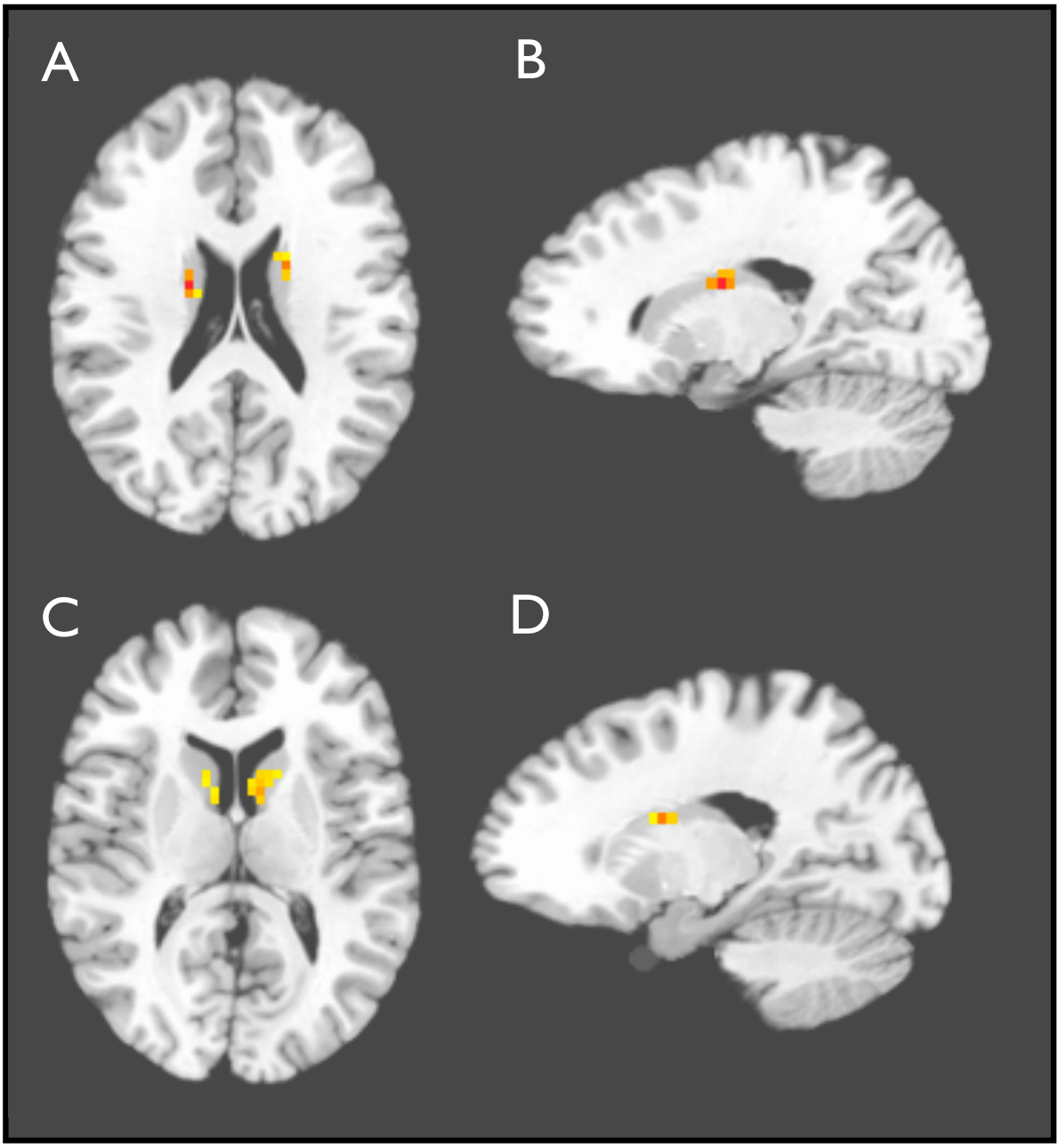

Figure 3.22: **Striatal Regions Correlating with Positive Reward-Prediction Error for Saccadic Eye Movements. (A-B, D)** Axial, left, and right saggittal views showing dorsal caudate activation when saccadic eye movements were reinforced. **(C)** Axial view showing anterior caudate activation. See Table 3.6 on page 104.

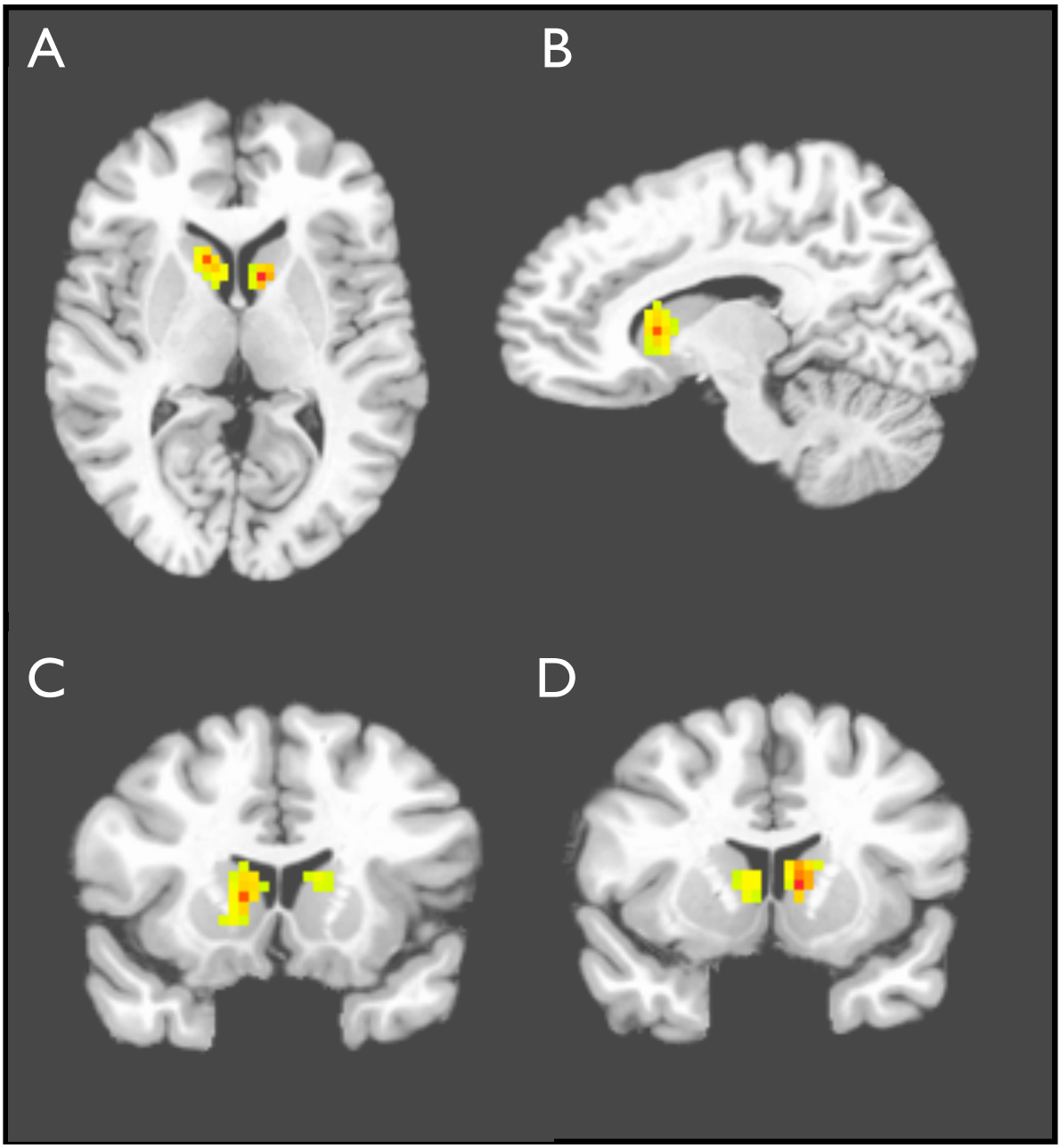

Figure 3.23: **Positive Reward-Prediction Error for Covert Attention Shifts.** Anterior caudate was positively correlated with both positive and negative reward-prediction errors when covert attention shifts were reinforced. See Table 3.6 on page 104.

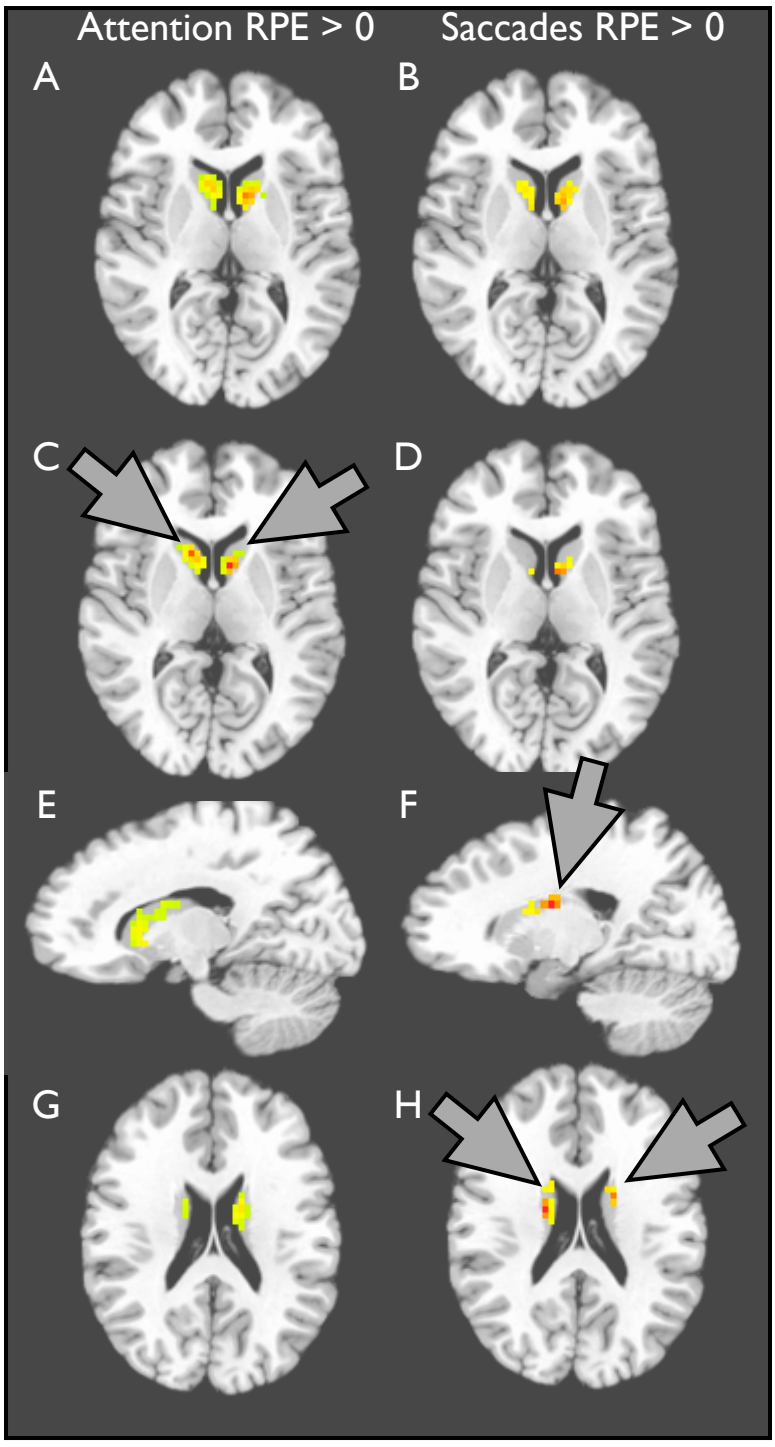

Figure 3.24: **Comparing Peak Activations for Positive Reward-Prediction Error of Covert Attention versus Saccadic Eye Movements which are in Anterior versus Dorsal Caudate, Respectively**. Peak voxels are indicated by grey arrows. See text for further discussion (page 112).

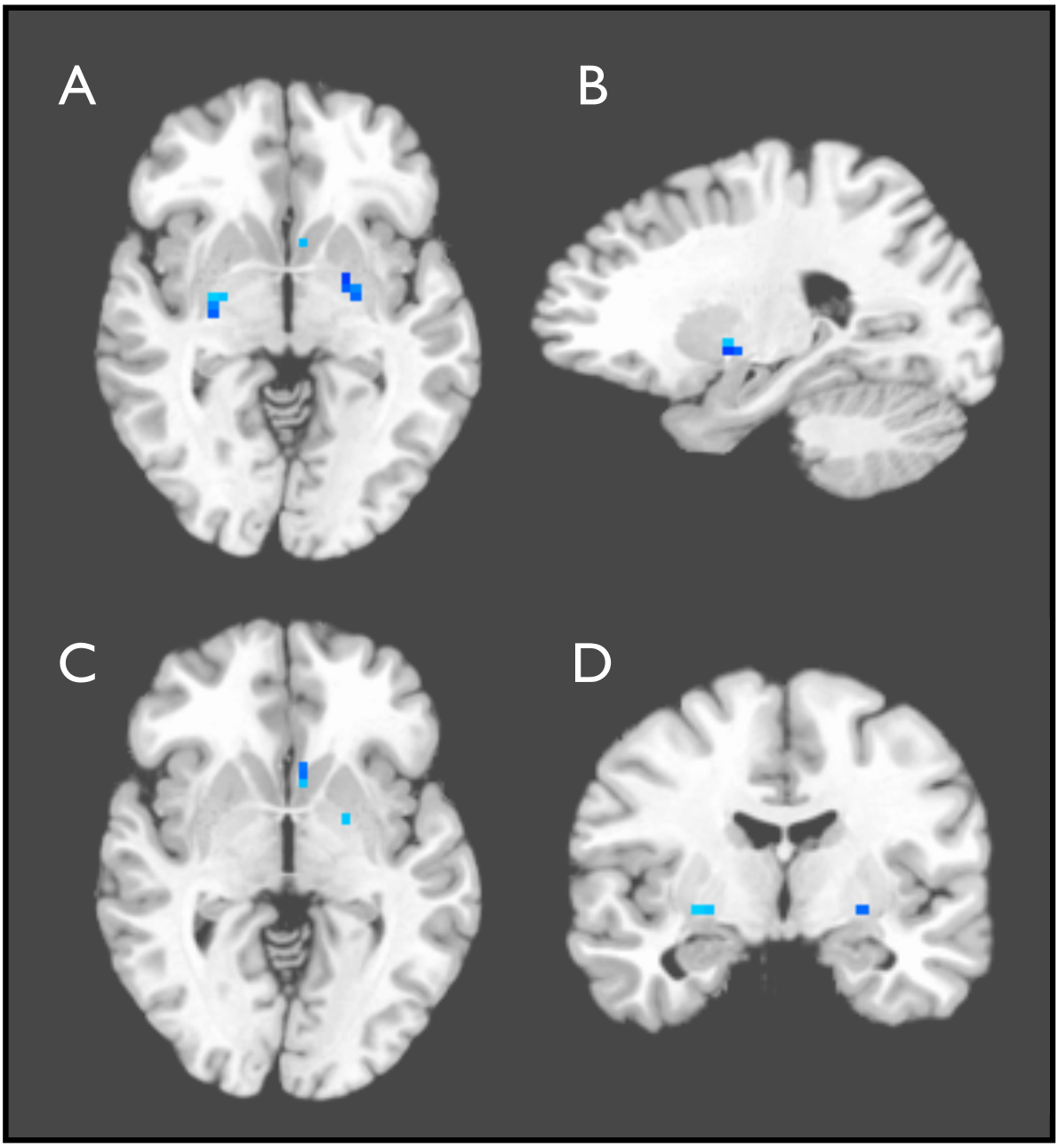

Figure 3.25: **Striatal Regions Correlating with Negative Reward-Prediction Error Signals for Right-Hand Movements. (A-D)** Axial, Saggittal, Axial and Coronal views showing regions that were correlated with negative reward-prediction error signals when right-hand movements were reinforced. Note that the putamen regions are very close to, and appear to overlap with the globus pallidus (particularly on the right/ipsilateral side). A small region was located in the anterior caudate. For numerical details, see Table 3.7 on page 106.

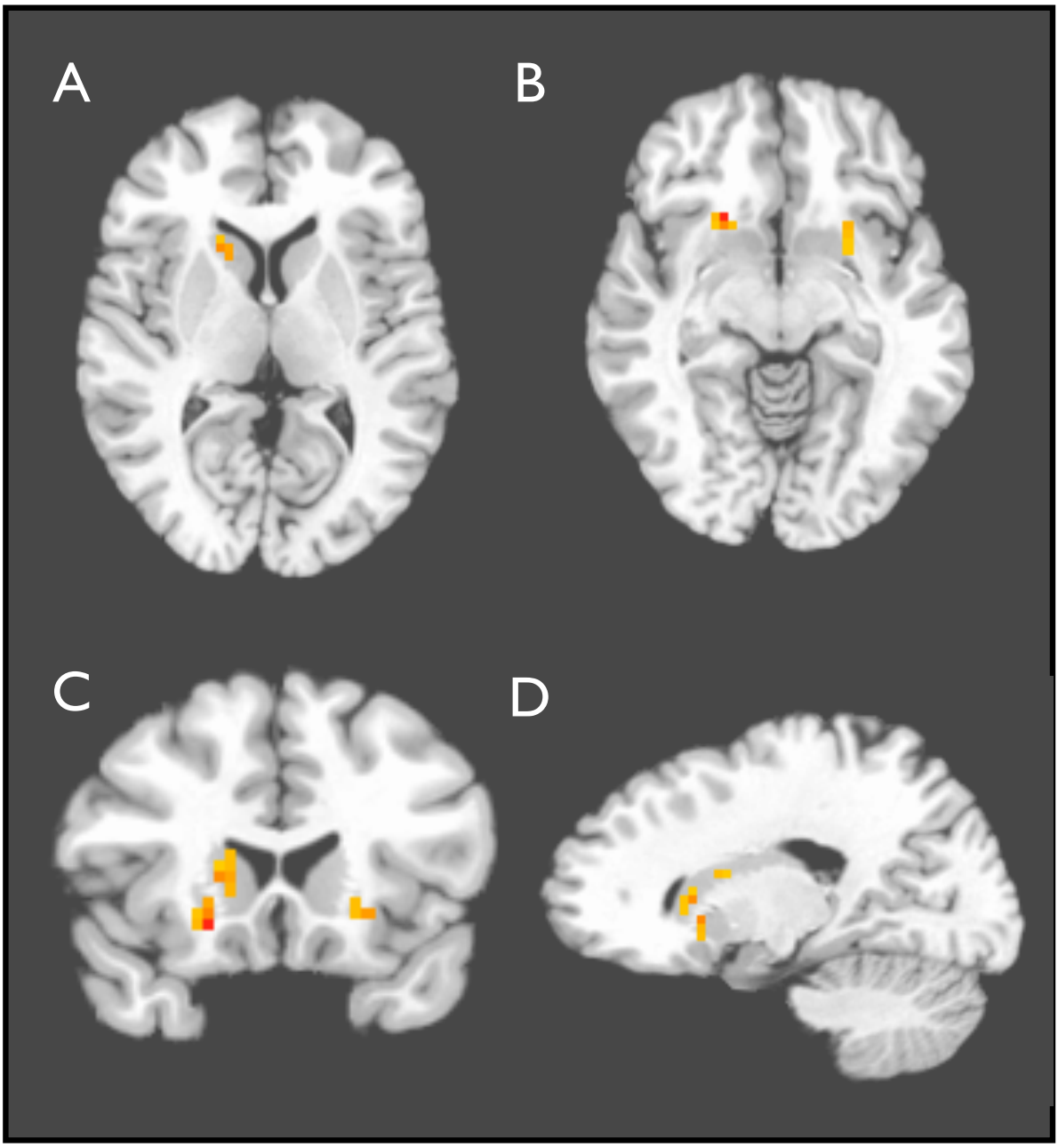

Figure 3.26: **Striatal Regions Correlating with Negative Reward-Prediction Error for Saccadic Eye Movements.** A diffuse set of regions were correlated with negative reward-prediction error signals when eye movements were reinforced. These regions were located in the anterior caudate, nucleus accumbens, and dorsal caudate. See Table 3.7 on page 106.

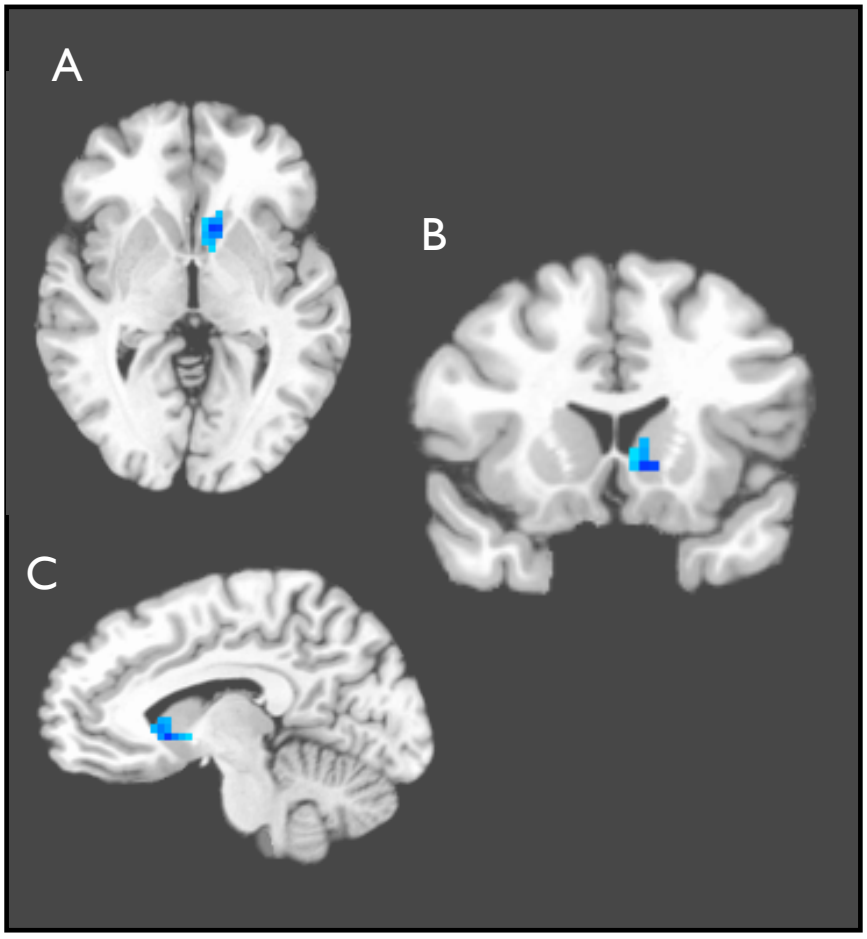

Figure 3.27: **Striatal Region Correlating with Negative Reward-Prediction Error for Reinforced Covert Attention Shifts.** A single region in the head of the caudate was correlated strongly with negative reward-prediction error signals. See Table 3.7 on page 106.

Region	Size	Peak Coord.	Peak t-value
Positive Hand Mvt.			
(*voxelwise p < 0.2*)			
Anterior Caudate	22 voxels	-5,13,7	t=2.25
	6 voxels	8,7,4	t=1.64
Left Putamen*	6 voxels	-20,1,13	t=1.76
Left Putamen*	10 voxels	-23,-6,-3	t=-2.11
Positive Eye Mvt.			
(*voxelwise p < 0.1*)			
Anterior Caudate*	19 voxels	-5,4,4	t=3.34
	10 voxels	11,7,16	t=2.34
Dorsal Caudate*	6 voxels	17,-6,20	t=3.56
	4 voxels	-17,1,20	t=2.92
Positive Att. Shift			
(*voxelwise p < 0.01*)			
Anterior & Dorsal Caudate*	60 voxels	-8,7,7	t=8.65
Anterior Caudate*	55 voxels	11,13,7	t=7.72

Table 3.6: **Positive Reward-Prediction Errors in Striatum.** Complete list of striatal regions involved in the positive reinforcement of hand movements, saccadic eye movements, and covert visual-spatial attention shifts during both CONTROL and LEARNING trials. Asterisks (*) indicate regions expected to be involved given prior anatomical knowledge. Extent thresholds for hand movements, eye movements and covert attention shifts were 6, 4, and 10 voxels, respectively. Note that because the dorsal caudate is a much smaller region, a smaller extent threshold was used to test for predicted activation there. Activation when attention shifts were reinforced extended into the dorsal caudate but did not create a focus there that could be separated by thresholding.

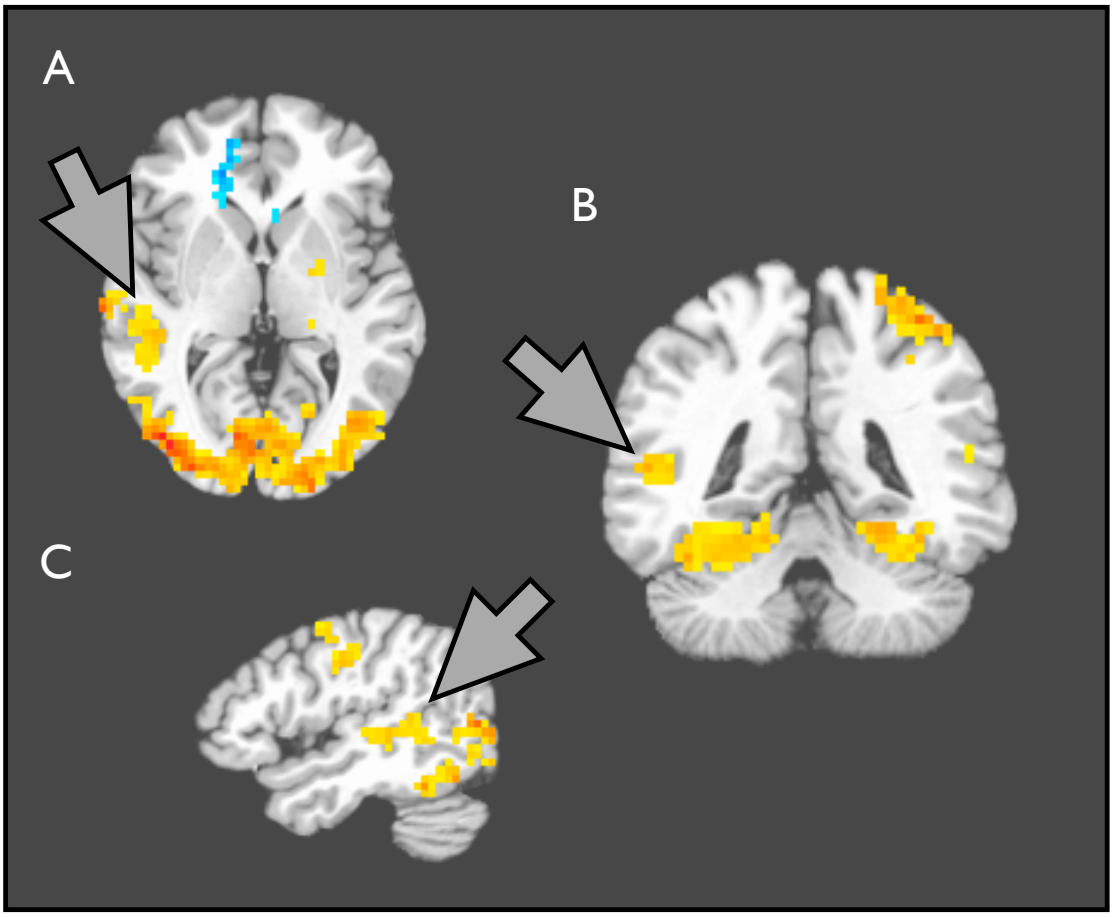

Figure 3.28: **MT Activation Correlating with Movement during Right-Hand Movement Trials**. The data in this figure confirm the hypothesis that the scrolling of the stimulus during hand-movement trials would result in visual motion signals in putative human area MT in the middle temporal cortex. **(A)** Axial view showing a distinct region in right middle temporal gyrus (arrow) correlating with hand movement regressors. This region contained 67 voxels (peak at 45mm, -31mm, 0mm, *t*-value=3.47). There was also a smaller region on the contralateral side containing 20 voxels (peak at -52mm, -46mm, 13mm, *t*-value=3.40). Smaller hand movement-related activations corresponding to right-hand movements can still be seen at this conservative threshold in the left putamen and thalamus. There was also extensive activation in visual cortex generally. **(B)** Coronal view showing putative area MT. Left primary motor cortex activation can also be seen at the top of the brain. **(C)** Saggittal view. (See text on page 107.) [Voxelwise $p < 0.025$.]

Region	Size	Peak Coord.	Peak t-value
Negative Hand Mvt.			
(*voxelwise p < 0.12*)			
Left Putamen*	5 voxels	-20,-2,-3	t=-2.55
Nucleus Accumbens	8 voxels	-5,13,0	t=-2.29
Right Putamen	5 voxels	27,-15,-3	t=-2.38
Negative Eye Mvt.			
(*voxelwise p < 0.2*)			
Anterior Caudate*	25 voxels	17,16,7	t=1.72
Nucleus Accumbens	12 voxels	20,16,-6	t=2.17
	11 voxels	-23,16,-3	t=1.66
Dorsal Putamen*	7 voxels	23,4,13	t=1.66
Negative Att. Shifts			
(*voxelwise p < 0.04*)			
Anterior Caudate*	21 voxels	-8,13,0	t=-3.82

Table 3.7: **Negative Reward-Prediction Errors in Striatum.** Complete list of striatal regions involved in the negative reinforcement of hand movements, saccadic eye movements, and covert visual-spatial attention shifts during both CONTROL and LEARNING trials. Asterisks (*) indicate regions expected to be involved given prior anatomical knowledge. Extent thresholds for hand movements, eye movements, and covert attention shifts were 4, 4, and 10 voxels, respectively.

MIDDLE TEMPORAL CORTEX ACTIVATION As mentioned earlier in section 3.4.1 (page 79), when only the CONTROL trials were analyzed, there was no distinct region in the middle temporal cortex that correlated with the on-screen scrolling due to joystick/hand movements, although such a region was hypothesized to be active. With the increased number of trials, and therefore power, obtained by analyzing both the CONTROL and LEARNING trials, whole-brain activity was again examined during hand movements to re-test the hypothesis. The pattern of activation is shown in Figure 3.28 on page 105. Examining these data showed a bilateral region of middle temporal cortex that correlated with hand movements. These data support the hypothesis, as this region likely reflects perception of visual motion when the stimulus scrolled on the screen.

3.4.3 TASK-RELATED ACTIVATIONS

In addition to performing model-based analyses for detecting the regions involved in movement and reward-prediction errors, it was also interesting to examine brain regions that were involved in the task generally. To that end, the data from each quarter of the experiment were averaged and a whole-brain voxelwise ANOVA was performed. As tabulated in Table 3.8 and shown in Figure 3.29, there were three regions whose variance was explained by quarter (i.e., by dividing the total number of trials into quarters) as a factor. The middle frontal gyrus, i.e., Brodmann area 8, was the largest significant region. This region borders regions of the dorsal-lateral prefrontal cortex (i.e., Broadmann area 9) and the cingulate gyrus, and appears to be involved in uncertainty during early decision-making tasks (Volz et al., 2005). It is also possible that this region is involved in maintaining the possible options that have been recently selected or those that should be explored in the near future. If so, it should be the case that this region is more active earlier in learning compared to later in learning. A post-hoc examination of the present data confirms this idea: post-hoc pairwise comparisons confirmed that mean activity was higher during the first quarter compared to second quarter (peak t-value = 3.10, $p < 0.005$) and third quarter compared to the fourth quarter (peak t-value = 2.97, $p < 0.006$).

Another area that followed a similar pattern to the prefrontal cortical areas was the right fusiform gyrus (first quarter greater than second, peak t-value = 2.63, $p < 0.02$; third quarter greater than fourth, peak t-value = 2.96, $p < 0.006$). The increased activation of the right fusiform gyrus, known

to be involved in processing of visual configural information such as faces and houses, may reflect increased attention to the spatial configuration of the on-screen stimuli to help encode and/or map the stimuli to the actions that are being reinforced (Harris and Aguirre, 2008).

Although the ANOVA also found the anterior cingulate cortex to be involved, the pattern in the anterior cingulate cortex was different compared to the prefrontal and fusiform areas. There was a only significant decrease in activation from first quarter to the second quarter of the task (peak *t*-value = 2.69, $p < 0.01$), but not from the third to the fourth (activation did not survive correction for multiple comparison). This pattern may reflect increased response conflict at the beginning of each run, i.e., between action choices early in learning, possibly because not enough evidence from reward-prediction errors has accumulated to make one response clearly better (or worse) than the others (Holroyd and Coles, 2008; Yeung et al., 2004) whereas during the second half participants did not have strong response conflict: they either selected actions based on previous reinforcement history, or based on newly received reinforcement.

Region	Size	Peak Coord.	Peak t-value	Corrected p-value
Middle Frontal Gyrus / BA8	220 voxels	-14,23,36	F(1,4*14-1) = 13.0	$p < 0.0001$
Right Middle Fusiform Gyrus	62 voxels	42,-52,-9	F = 12.86	$p < 0.0001$
Anterior Cingulate	62 voxels	5,19,23	F = 9.78	$p < 0.0001$
Medial Frontal Gyrus / BA9	27 voxels	5,41,26	F = 8.73	$p < 0.0001$

Table 3.8: **Whole-Brain ANOVA as a factor of Quarter of Learning Task** [Voxelwise $p < 0.001$.]

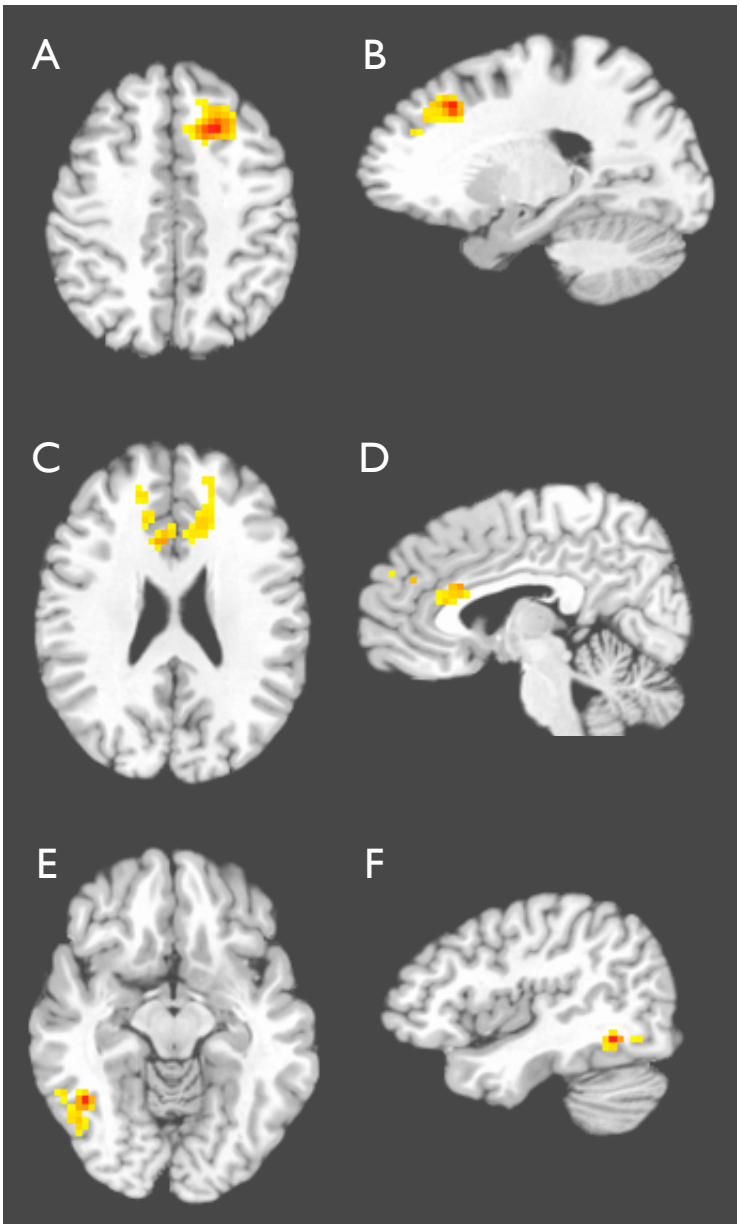

Figure 3.29: **Whole-Brain ANOVA with Quarter of Learning Task as a Factor**. Activation shows regions active when new contingencies are being learned compared to after performance reached asymptote. **(A-B)** Activation in medial front gyrus may reflect increased memory demands or uncertainty. **(C-D)** Activation in anterior cingulate cortex may reflect increased response conflict. **(E-F)** Activation in right fusiform gyrus which may reflect increased attention to the spatial arrangement of the stimuli. See text on page 107 for details. [Voxelwise $p < 0.001$; all regions corrected at $p < 0.0001$.]

3.5 SUPPLEMENTAL DISCUSSION

The results of the experiments showed that the basal ganglia are involved in not only the control of physical actions, but more generally in the control and RL of both physical and cognitive actions. The results also demonstrated that action-specific reinforcement signals are localized differentially depending on the action that is being reinforced. A comparison of these activations with movement-related activations demonstrated that they were roughly co-localized to the same regions of the striatum, regions known to receive input from the cortical areas controlling those actions.

As mentioned in the introduction chapter there are likely many forms of learning going on during any task. How then can one hope to isolate learning that is due to reinforcement? First, predictions from RL are continuously-valued and highly dependent on the history of rewards– this is unlikely to correlate strongly to signals due to other forms of leaning. Nevertheless, the experiments were designed to attempt to minimize the effects of other forms of learning as much as possible so that intact measurements of RL signals can be observed.

The two main forms of learning that are being controlled for are error correction learning and associative forms of learning. Error correcting learning is assumed to occur during the CONTROL trials as participants learn to correct errors in their hand movements on the joystick as well as errors in their saccadic eye movements that might both arise because of the posture while the participant is in the scanner. (It is unclear whether it is possible for participants to make "errors" in the allocation of visual-spatial attention.) Associative learning of the form of associating stimulus colors to movements should be ongoing during both the CONTROL and LEARNING phase of the experiment, and is thus also controlled in the experiment design.

One particularly interesting and somewhat unanticipated result was the activation of the anterior caudate during the reinforcement of hand movements. This result was particularly interesting because the anterior caudate was not involved in the control of hand movements themselves. One explanation for this activation may be due to the nature of the experimental task: when participants performed hand movements, there was scrolling of the stimulus on the screen, while participants had to maintain attention and fixation in the center of the screen. Consequently, on trials when hand movements were reinforced (i.e., trials with better than predicted outcomes), participants must have

succeeded in maintaining their attention centrally to make a correct response. This explanation also fits with the idea that attention shifts are a prepotent response, and either that they occur by default, or that they occur because of the other two experiment conditions – covert attention shifts and eye movements – also involved attention shifts.

Another possible explanation for this result is that the anterior caudate is involved in multiple forms of attention, i.e., not only involved in visual-spatial attention, but also, e.g., in attention to motor plans and/or decisions. Indeed, the caudate locus of activation for reinforcement of hand movements is ventral and anterior in the head of the caudate, whereas the loci of activations for the reinforcement of both eye movement and covert attention shifts is most posterior within the caudate head. However, this conclusion is not directly supported by the data because, as just mentioned, hand movements themselves do not involve the anterior caudate during their control.

Taken together, these data have important implications for a strong version of the premotor theory of attention, and for learning theories based on Rescorla-Wagner, and for computational theories of basal ganglia function. As will be discussed in the dissertation's conclusions, these data also provide novel information which can serve to constrain future theoretical and computational models of basal ganglia function.

3.5.1 EVIDENCE AGAINST A STRONG PREMOTOR THEORY OF ATTENTION

The present experiment provides an opportunity to test the *premotor theory of attention* (Rizzolatti et al., 1987). According to this theory, eye movements to a particular location require a preceding shift of attention to that location. A strong version of the theory states that covert visual attention shifts to particular locations are an early stage of saccade programming, rather than a distinct action directed to those locations (for a review, see Shipp, 2004). When a covert attention shift occurs that is not followed by a subsequent saccadic eye movement, the premotor theory of attention holds that the saccade program has been interrupted before completion – it can be said to be 'canceled'. Part of the evidence in support of this view is that visuomotor neurons in superior colliculus neurons gradually increase their activity until they reach a threshold at which time a saccade occurs, and this activity is also present during covert shifts of attention (Hoffman and Subramaniam, 1995; Kowler et al., 1995). Thus a covert attention shift is thought to simply be 'sub-threshold' saccadic

eye movement activity, identical to the act of programming the saccadic eye movement. Additional evidence for this view is that microstimulation of the superior colliculus can result in covert attention shifts (Müller et al., 2005).

However, although there is much data to support the view that visual attention shifts are equivalent to saccade programming, there is evidence in support of a competing hypothesis which maintains that attention shifts and eye movements are separate actions. A number of prior studies have attempted, with varying degrees of success, to dissociate attention from saccade programming through recordings in frontal eye fields (e.g., Murthy, Thompson & Schall, 2001).

In the present data set we tested the hypothesis that covert attention shifts and the attention shifts that accompanied saccadic eye movements in the task, could be separately reinforced (i.e., as separate actions). This hypothesis was tested by examining the location of reward-prediction error signals for eye movements versus covert attention shifts. Data from the two regressors are plotted side-by-side in Figure 3.24. These results show that the anterior versus dorsal caudate are differentially involved in the positive reinforcement of covert attention shifts and saccadic eye movements. For attention shifts, activity correlating with positive reward-prediction error was bilaterally focused in the anterior caudate. In contrast, for saccadic eye movements, activity correlating with positive reward-prediction error was bilaterally focused in the dorsal caudate. These differences may reflect the fact that the anterior caudate receives projections from parietal cortical regions involved in attention shifts whereas dorsal caudate receives projections from the frontal eye fields, which are predominantly involved in saccadic eye movements. Note that although the anterior caudate region that correlated with positive reward-prediction error extended in the dorsal direction, attention shifts did not activate distinct foci in the dorsal caudate. These results suggest that reward-prediction errors are computed in distinct regions of the striatum for shifts of visual attention that are not accompanied by eye movements, compared to eye movements themselves.

The present analysis was carried out to test the hypothesis that covert attention shifts and saccadic eye movements have separate neural substrates for reinforcement. If they did, this would be evidence against a strong form of the premotor theory of attention, which holds that covert attention shifts are exactly the programming of saccadic eye movements to a location. The results suggest that covert attention shifts were reinforced in the anterior part of the caudate head, whereas saccadic

eye movements were reinforced at the posterior part of the caudate head. In addition, saccadic eye movements were also reinforced in the dorsal caudate and the nucleus accumbens, whereas covert attention shifts were not. These results support the hypothesis and suggest that a strong view of the premotor theory of attention may be false.

The results thus suggest a weaker view of the relationship between saccades and covert attention: in normal vision, covert attention and eye movements are both directed to locations in visual space, and because attention shifts are faster than saccades, attention shifts tend to precede the eye movements themselves (e.g., Posner et al., 1980). Through this highly repeated process, it is even possible that covert shifts of visual attention and saccades become associated through large numbers of trials. In the majority of saccades, to the extent that the goal of vision is to perceive stimuli at the target of the saccade, covert attention shifts are beneficial to vision, they will tend to precede the saccadic eye movements themselves.

The fact that our behavioral and fMRI data support the independent reinforcement of eye movements and covert attention shifts does not preclude the possibility that *all eye movements to a location require a covert attention to that location*. Therefore, to test the hypothesis that covert attention shifts need not precede saccadic eye movements, we are currently preparing an experiment that aims to dissociate saccade eye movements from covert attention shifts by rewarding participants for keeping their attention at fixed locations while moving their eyes to various targets on the screen.

3.6 THEORETICAL IMPLICATIONS

3.6.1 EXTENDING RESCORLA-WAGNER-BASED LEARNING RULES

The results from the fMRI study suggest that reward-prediction error computations are performed not in a single circuit, but rather in separate parallel circuits. This conclusion has profound implications for formulations of learning rules such as the Rescorla-Wagner learning rule. Importantly, these data suggest a modification to that style of learning rule that help it explain the phenomena ob-

served in this dissertation. A consequence of this extension is that phenomena that were previously beyond the scope of Rescorla-Wagner learning rules, such as **blocking**, can now be explained.

The standard Rescorla-Wagner rule suggests that learning should be done based on the difference between received and predicted reward. This quantity, delta, can be computed as:

$$\delta = reward - V(s) \tag{3.1}$$

where V(s) is the amount of reward that is currently predicted from the state preceding the reward. This can then be integrated into a learning rule which describes the new reward prediction quantity after learning has occurred:

$$V_{new}(s) = V(s) + learningRate \times \delta \tag{3.2}$$

By substituting Equation 3.1 into Equation 3.2, we have:

$$V_{new}(s) = V(s) + learningRate \times [reward - V(s)] \tag{3.3}$$

An extensive body of literature has shown that this learning equation works very well for many cases and turns out to be quite powerful. Importantly, note that when reward has been predicted correctly $V(s) = reward$, and learning stops because the difference $[reward - V(s)]$, and therefore δ, is zero.

However, the current experimental results suggest an alternate formulation in which reward predictions are not simply computed as a function of a single global difference. Instead, the results suggest the that this difference is *permissive* for learning, but the learning itself is performed in a piecewise fashion depending on the reward predictions errors computed separately for component neocortical activations (in this case, those neocortical activations representing actions). That is, **reward prediction errors are also computed separately for each neocortical activation or action**

representation.

$$\begin{aligned} \delta_{a1} &= reward - V_{old}(a1) \\ \delta_{a2} &= reward - V_{old}(a2) \\ \delta_{a3} &= reward - V_{old}(a3) \\ &\ldots \end{aligned}$$

where s1, s2, s3... are patterns of neocortical activation corresponding to separate component stimuli or actions. A local reward-prediction error can therefore be computed to allow for learning for each stimulus present or action performed when the reward was received:

$$\text{When } reward - V(s) \text{ is not = 0} \begin{cases} V_{new}(s) = V(s) + learningRate \times [reward - V(s)] \\ \text{and} \\ V_{new}(a1) = V(a1) + learningRate \times [reward - V(a1)] \\ V_{new}(a2) = V(a2) + learningRate \times [reward - V(a2)] \\ V_{new}(a3) = V(a3) + learningRate \times [reward - V(a3)] \end{cases} \tag{3.4}$$

The above allows learning to occur for global reward predictions simultaneously with independent reward predictions for each stimulus component or neocortical action representation. Critically, when only a single action is being performed or reinforced, this extension of the Rescorla-Wagner reduces exactly to the standard Rescorla-Wagner rule.

One possibility that certainly warrants investigation is that the global reward-prediction error could be based on the difference between received reward and the **sum of all currently-predicted rewards**, i.e., $\delta = reward - V_{old}(s) = reward - [V_{old}(a1) + V_{old}(a2) + V_{old}(a3) + \ldots$ If indeed reward predictions are additive, the new global reward prediction could be computed as the sum of the independent component reward predictions: $V(s) = V(a1) + V(a2) + V(a3) + \ldots$ Note that there may be a also be a bias term to compensate for the possibility that two neocortical patterns predict the same amount of reward (rather than each of those patterns predicting half the amount of reward, i.e., $V(s) = V(s\_bias) + V(a1) + V(a2) + V(a3) + \ldots$).

This extension of the Rescorla-Wagner rule explains the current data better than the standard Rescorla-Wagner rule because it captures the aspect of the data showing that reward-prediction errors are computed independently, i.e., in separate brain circuits.

Besides helping to explain the current data, this extension to the Rescorla-Wagner rule can also explain downshifting during unblocking. For example, a rat is trained to predict positive reinforcement (e.g., 3 pellets of food) given a single conditioning stimulus (e.g., a light). After the reward prediction has been learned for the light stimulus, another stimulus (e.g., a noise) is paired with that conditioning stimulus during reward trials. Blocking refers to the fact that no learning appears to take place for that second stimulus: when the rat is tested with the noise alone, no appetitive behavior ensues. Learning of reward predictions for that new stimulus has been "blocked".

However, learning for the second, or new, component of the CS does occur if the amount of reward is changed. For example, if the amount of reward is changed from 3 food pellets to 1 food pellet, learning to the noise is no longer blocked. In particular, the rat now learns that the noise predicts 1 food pellet.

As pointed out by Holland (Holland, 1984), the standard Rescorla-Wagner learning rule predicts that the reward prediction for the noise should be aversive during unblocking – because the reward-prediction error was 3 - 1 = -2. The fact that an appetitive response was learned to the noise suggests that the standard Rescorla-Wagner rule is inadequate for explaining learning in this situation.

In contrast, the present parallelized learning rule *does* capture this aspect of blocking, as follows: Initially, all reward predictions are zero:

$$\begin{aligned} V(global) &= 0 \\ V(light) &= 0 \\ V(noise) &= 0 \end{aligned}$$

Then, a reward of 3 food pellets is paired with a light CS until reward-prediction is adequately learned so that the global reward prediction and light predictions both are 3:

$$\begin{aligned}
\delta_{global} &= 3 - V(global) \quad \ldots \Rightarrow \ldots V(global) = 3 \\
\delta_{light} &= 3 - V(light) \quad \ldots \Rightarrow \ldots V(light) = 3 \\
\delta_{noise} &= \text{(not trained)} \quad \ldots \Rightarrow \ldots V(noise) = 0
\end{aligned}$$

Once this has been learned, administration of the same amount of reinforcement, i.e., 3 pellets, will result in blocking for reinforcement of V(noise) presented, because the predicted reinforcement, V(global), is equal to 3, and therefore no learning will occur as per Equation 3.4, i.e.,

$$\text{Because } reward - V(S) = 3 - V(s) = 0, \begin{cases} V_{new}(global) = unchanged \\ \text{and} \\ V_{new}(a1) = unchanged \\ V_{new}(a2) = unchanged \\ V_{new}(a3) = unchanged \end{cases} \tag{3.5}$$

Now, in the downshifting case, the *noise* stimulus is presented simultaneously with the *light* stimulus and reinforcement is reduced from 3 to 1. Because of this reduction, the global reward-prediction error will become non-zero, and therefore permissive for learning. During these downshifting training trials, $V(global)$ and $V(light)$ begin to drop rapidly in their reward predictions from 3 to 1, while V(noise) rises from 0 to 1.

$$\begin{aligned}
\delta_{global} = 1 - V(global) \quad &\ldots \Rightarrow \text{decreases from 3} \ldots \Rightarrow \quad V(global) = 1 \\
\delta_{light} = 1 - V(light) \quad &\ldots \Rightarrow \text{decreases from 3} \ldots \Rightarrow \quad V(light) = 1 \\
\delta_{noise} = 1 - V(noise) \quad &\ldots \Rightarrow \text{increases from 0} \cdots \Rightarrow \quad V(noise) = 1
\end{aligned}$$

(Predictions of Proposed 'Parallelized' Extension of Rescorla-Wagner Rule)

This increase for $V(noise)$, is predicted by the proposed "parallelized" variant of the Rescorla-Wagner rule, and is more consistent with observed behavioral changes. In contrast, the standard

Rescorla-Wagner rule detailed earlier ends up predicting a different set of learned reward predictions. This is because the standard Rescorla-Wagner rule bases learning on a global reward-prediction error, rather than a local (i.e., action-specific) reward-prediction error. As a result, the reward prediction associated with the each stimulus is predicted to change by -2 units. In particular, the noise stimulus is predicted – incorrectly, according to behavioral results – to decrease from 0 to -2:

$$\delta_{global} = 1 - V(global) \quad \ldots \Rightarrow \text{decreases from } 3 \ldots \Rightarrow \quad V(global) = 1$$

$$\delta_{light} = 1 - V(global) \quad \ldots \Rightarrow \text{decreases from } 3 \ldots \Rightarrow \quad V(light) = 1$$

$$\delta_{noise} = 1 - V(global) \quad \ldots \Rightarrow \text{decreases from } 0 \cdots \Rightarrow \quad V(noise) = -2 * * *$$

(Predictions of Standard Rescorla-Wagner Learning Rule)

The "incorrect" prediction of the Rescorla-Wagner rule therefore, as indicated by the asterisks, is that the noise takes on an aversive value.

Thus, the proposed extension of the Rescorla-Wagner rule – a simple parallelization of reward-prediction error computations that is suggested by the present imaging results – is not only a more adequate explanation of the present neuroimaging results, but also is better able to capture at least this unblocking/downshifting phenomenon that is observed in reinforcement learning in rats.

It is important to note that this proposal is somewhat at odds with the idea of an exclusively global reward-prediction error signal. However, so are the neuroimaging results in this dissertation. If the reward-prediction error signal were effectively global at the level of the striatum, one would not expect differentially-localized activation depending on which action was being performed: rather, the entire striatum should be active. Instead, it appears that the striatum can mediate local computations reflecting the difference between action-specific reward predictions, and the administered reward.

Although this extension of Rescorla-Wagner was suggested by the neuroimaging results, the extension also makes new fMRI predictions that are not tested in the current data set. In particular,

once reward prediction errors for actions have been learned separately, administering unexpected reward for the conjunction of those actions (i.e., performed simultaneously) should generate positive and negative reward prediction error signals in multiple striatal zones. This kind of experiment would provide strong evidence for the simultaneous parallel reinforcement of actions in separate circuits.

3.6.2 COMPUTATIONAL THEORIES OF BASAL GANGLIA FUNCTION

The results from the fMRI experiment have a number of implications for understanding the kinds of computations that are performed in the basal ganglia. Two of these implications have already been mentioned in detail (i.e., control of large neural networks and separate local computation of reward-prediction errors.) There are several additional implications that will be discussed below.

The first implication of the fMRI results is that they license the assumption that the basal ganglia are the locus of the link between action selection and reward. Although the basal ganglia have long been implicated in go/no-go signals for the initiation of acton, a number of existing computational models have more recently proposed that **action selection** given particular reward history to be one of the primary functions of the basal ganglia (Bar-Gad et al., 2003; Gurney et al., 2001a,b; Leblois et al., 2006). These models imply that the basal ganglia monitor reward-related signals and use these signals to select particular actions. Although there is evidence that distinct regions of the basal ganglia are involved in the control of different kinds of actions, there has not been evidence that reinforcement signals also followed this pattern of localization. Indeed, as mentioned previously, most studies of the basal ganglia focus on the nucleus accumbens (i.e., the ventral striatum) as the locus of reinforcement learning. Thus, the present results license this core assumption of a number of basal ganglia models, that is, that action selection in cortico-basal ganglia loops can be modulated by localized reinforcement signals.

The second implication of the fMRI results is the prefrontal cortex and the basal ganglia appear to collaborate in control during RL. Specifically, the present results suggest that the prefrontal cortex neither leads or lags basal ganglia activation during reward-related learning. The topic of whether the prefrontal cortex or basal ganglia "learns first" has been a matter of debate (e.g., Pasupathy and Miller, 2005). The data in the present study suggest that prefrontal cortex and basal

ganglia activations tracked with each other: both decreased in activity over time as learning progressed. This result suggests that neither the prefrontal cortex or the basal ganglia are the originating sources of control during learning, but rather that the two structures continuously interact. What is the purpose of this interaction? The proposed "cognitive orchestra" framework suggests one possible answer: the basal ganglia may allow the prefrontal cortex to control specific actions, overcoming limited cortico-cortical connectivity. Another possible answer is that the prefrontal cortex activation is important in keeping track of the different possible actions in working memory. That is, during learning in the task it was important for participants to explore the four possible actions (i.e., up, down, left, right.) The prefrontal cortex might be important in mediating the explorative behavior, biasing the system to select actions that have not recently been selected.

The third implication of the fMRI results is that positive reinforcement appears to have more specificity than negative reinforcement. Generally it was found that positive reward-prediction error were more reliable and spatially localized within the striatum, whereas negative reward-prediction error signals were diffuse, more lateralized in the case of right-hand movements, and/or weaker. Models of RL have attempted to use the similarity between phasic dopamine and reward-prediction error as an anchoring point to relate RL algorithms to brain function. However, unlike reward-prediction error in RL models which can take on arbitrary positive and negative values, phasic dopamine levels can only go so far below their baseline levels, and there is no sense in which neurons can exhibit "negative firing levels". Computationally, this suggests that dopamine transmission may not be an effective signal for negative reinforcement. This has led some researchers to propose that other neurotransmitters or neuromodulators (e.g., serotonin: Daw et al., 2002) are primarily tasked with transmitting negative reinforcement. However, although dopamine is apparently incapable of conveying the same amount of variability in negative reinforcement, from a computational perspective this may actually not be problematic – especially given that the primary function of RL is to enable the selection of actions leading to the **maximization** of reward. That is, because the goal of the organism is to maximize positive reinforcement, knowing that an outcome is "bad" (i.e., somewhat negative) is probably sufficient – it may not be necessary to store a precise value indicating how bad two different alternatives are. Interestingly, this suggests that people may not perform very well at selecting the better outcome among two or more negative outcomes, compared to a situation of selecting the best among two or more positive outcomes. Indeed, this

hypothesis is confirmed in a recent study: participants are less accurate, and slower, at selecting the less punished (i.e., correct) object on trials in which both objects were punished ("PunPun" trials) relative to trials on which both objects were rewarded ("RewRew" trials) as well as relative to trials on which one object was rewarded and the other was punished (Blair et al., 2006).

Non-Reward-Related Signaling. As mentioned earlier, another objection has to do with the claim that phasic changes in dopamine reflect reward-prediction error. The objection arises because there are cases in which phasic increases in dopamine, which should signal positive reward-prediction error, occur that appear to be unrelated to reward. These cases, often termed 'novelty' or 'saliency' responses, have been shown to occur very rapidly before a reward prediction can be made, and also have been shown to occur for stimuli that lead to either reward or punishment. The fact that there might be a positive reward-prediction error for a negatively-valued stimulus appears to pose problems for reward-prediction error account of phasic dopamine specifically, and RL in these circuits more generally. However, computational simulations of RL agents can be used to address this concern, and such simulations are reported in Chapter 4.

3.7 CONCLUSION

This study is the first to demonstrate that reward-prediction errors are computed in the striatum for both physical (i.e., motoric) as well as cognitive (i.e., non-motoric) actions.

This suggests that the basal ganglia may perform this function for any action that is represented in regions of neocortex that project to the basal ganglia, notably, even if the action does not involve motor output. Future studies will investigate whether the reinforcement of other non-motor "actions" (e.g., mental rotation, manipulation of items working memory, math, linguistic processing, and/or other cognitive operations) may engage different regions of the basal ganglia in ways similar to the actions examined in this study.

This study also demonstrated anatomically-localized reward-prediction error computations in regions known to participate in specific cortico-basal ganglia loops and in the control of specific actions. This result is important because prior research has focused exclusively on the ventral

striatum in RL, or on a division between ventral and dorsal extents of the striatum (i.e., goal-directed versus habit learning). The present results suggest that rather than a focus on the division between ventral and dorsal striatum, that separate regions of the striatum all perform the same computations for afferent information. Thus, it may be that the striatum is performing the same computation to all neocortical information – so that what differs from one striatal region to another is not the nature of the computation performed there, but rather the object of that computation. For example, in the case of ventral striatum, it is known that the neocortical afferents are primarily limbic in nature, carrying emotional information. One prediction thus is that the reinforcement of emotional states would generate reward-prediction error signals in ventral striatum, much as the reinforcement of hand movements caused reward-prediction errors in the posterior putamen.

Besides showing that the basal ganglia were involved in the RL of physical and cognitive actions, the experimental work presented in this chapter also suggested that the cortico-basal-ganglia loop might be a substrate for reward-prediction error computations. Thus, the results provide a new constraint on a possible mapping between RL and basal ganglia anatomy. Finally, the results also suggested a new mathematical learning rule that is not only more computationally powerful, but also helps explain results from other experimental paradigms such as blocking and downshifting.

The striatal fMRI results were not as statistically significant as they could have been; they did not survive correction for multiple comparison. Part of this may have been because, as mentioned in the background (section 2.2 on page 20), the BOLD response is known to be directly related to dopamine activation (Knutson and Gibbs, 2007), and dopamine is believed to be generally released simultaneously by all dopamine neurons. The flooding of dopamine may have reduced the signal to noise ratio in the striatum during the course of the study by generating an increased average 'background' level of activity.

Although the results suggest a mapping of RL to the basal ganglia, an important question is whether RL is the principal computation of the basal ganglia, or whether the basal ganglia also have other functions that they compute in parallel. Another function might be to compute and signal the presence of salient, but not necessarily reward-related, stimuli (Bunzeck and Düzel, 2006; Horvitz, 2000; Knutson and Cooper, 2006; Zink et al., 2006, 2003). Unless saliency signals could be shown to be compatible with reward-prediction error signals, saliency signals could add complexity to

the computations because saliency appears to interact with reward (e.g., Zink et al., 2004). In the following chapter, a computational study of saliency is carried out in the context of RL simulations to help address this concern.

4.0 USING REINFORCEMENT LEARNING TO UNDERSTAND NOVELTY AND SALIENCY RESPONSES IN THE BASAL GANGLIA

4.1 INTRODUCTION

The basal ganglia appear to be involved in reward-related learning – this fact is supported by the study described in the previous chapter, as well as a variety of prior and current findings. However, a range of findings implicate the basal ganglia in the processing of novel and salient stimuli (Zink et al., 2006, 2003). Although some objects appear to have intrinsic salience (e.g., brightness, edges, sudden temporal onset) or salience brought on by context (e.g., singleton pop-out effects), studies suggest that salience and reward may interact (Zink et al., 2004). A possible explanation for this interaction is that rewarded stimuli may become more perceptually salient through plasticity mechanisms involved in recording rewarding experience. Importantly, there exist serious challenges for the independent study of reward versus salience in real biological systems.

Another aspect of basal ganglia function that appears to reflect the interplay of reward-related learning and saliency is found in the phasic activity of the dopamine neurons. Experimental work by Schultz and colleagues has demonstrated that under certain circumstances, the phasic activity of dopamine-releasing neurons appears to signal differences between predicted and received reward (Schultz, 1998). This fact has been used as the basis for linking Reinforcement Learning (RL) to computational theories of basal ganglia function. However, experimental findings suggest that this 'reward-prediction error' is but one of several signals communicated by that phasic activity: another signal that is also present in phasic dopamine appears to reflect responses to salient but unpredicted non-reward events (Horvitz, 2000; Redgrave and Gurney, 2006).

In an attempt to reconcile reward-related activity with novelty/saliency activity, some modelers

have resorted to making the additional a priori assumption in their computational models of basal ganglia function that novel, unexpected stimuli are rewarding (for a review, see Gillies and Arbuthnott, 2000; Kakade and Dayan, 2002). Although adding this a priori assumption allows these models to capture neural activity quite well, those models consequently cannot shed light on exactly *how* reward-related activity is related to novelty/saliency activity. For example it is conceivable that one type of activity gives rise to the other. Further, if both types of activity are entered into the model as assumptions, it is important to examine the consequences of these combined assumptions on behavior.

The article that is included as the main content of this chapter uses an RL simulation to examine whether novelty and saliency responses might emerge due to basic RL principles. Importantly, the RL simulation makes a minimal number of non-controversial assumptions. Indeed, the simulation is effectively based on a "plain-vanilla" RL agent in a simple one-dimensional environment, with two additional assumptions related to the agent's ability to perceive objects in its environment as well as the need to approach the object in order to extract value from it (i.e., consummatory behavior).

The results of the simulation show that the ability to learn reward predictions within the study constraints is sufficient to generate activity consistent with the apparently non-reward-related phasic dopaminergic response. That is, an agent that is formulated to learn reward predictions will also, as a consequence of limited perceptual ability, generate novelty and saliency responses.

4.2 ARTICLE CONTENT

The next pages contain the full text of Laurent (2008) The emergence of saliency and novelty responses from reinforcement learning principles. *Neural Networks*, 21(10):1493-9. The publisher, Elsevier, grants authors the right to reuse their own articles published by Elsevier, either in part or in full, in a thesis or dissertation. (For information this right, please see page 3 of *Pamphlet #4, Ways to Use Journal Articles Published by Elsevier: A Practical Guide, 3rd Edition*, 2008).

Neural Networks 21 (2008) 1493–1499

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

The emergence of saliency and novelty responses from Reinforcement Learning principles☆

Patryk A. Laurent*

University of Pittsburgh, Centers for Neuroscience and for the Neural Basis of Cognition, 623 LRDC, 3939 O'Hara St., Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 1 October 2007
Received in revised form
15 September 2008
Accepted 18 September 2008

Keywords:
Novelty response
Reinforcement learning
Dopamine
Orienting
Reward-prediction error

ABSTRACT

Recent attempts to map reward-based learning models, like Reinforcement Learning [Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge, MA: MIT Press], to the brain are based on the observation that phasic increases and decreases in the spiking of dopamine-releasing neurons signal differences between predicted and received reward [Gillies, A., & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, *15*(5), 762–770; Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27]. However, this *reward-prediction error* is only one of several signals communicated by that phasic activity; another involves an increase in dopaminergic spiking, reflecting the appearance of salient but unpredicted non-reward stimuli [Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506; Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656; Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975], especially when an organism subsequently orients towards the stimulus [Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27]. To explain these findings, Kakade and Dayan [Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559.] and others have posited that novel, unexpected stimuli are intrinsically rewarding. The simulation reported in this article demonstrates that this assumption is not necessary because the effect it is intended to capture emerges from the reward-prediction learning mechanisms of Reinforcement Learning. Thus, Reinforcement Learning principles can be used to understand not just reward-related activity of the dopaminergic neurons of the basal ganglia, but also some of their apparently non-reward-related activity.

Reinforcement Learning (*RL*) is becoming increasingly important in the development of computational models of reward-based learning in the brain (Gillies & Arbuthnott, 2000). RL is a class of computational algorithms that specifies how an artificial "agent" (e.g., a real or simulated robot) can learn to select actions in order to maximize total expected reward (Sutton & Barto, 1998). In these algorithms, an agent bases its actions on *values* that it learns to associate with various *states* (e.g., the perceptual cues associated with a stimulus). These values can be gradually learned through *temporal-difference learning*, which adjusts state values based on the difference between the agent's existing reward prediction for the state and the actual reward that is subsequently obtained from the environment. This computed difference, termed *reward-prediction error*, has been shown to correlate very well with the phasic activity of dopamine-releasing neurons projecting from the substantia nigra in non-human primates (Schultz, 1998). Furthermore, in humans, the striatum, which is an important target of dopamine, exhibits an fMRI BOLD signal that appears to reflect reward-prediction error during reward-learning tasks (McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Tanaka et al., 2004). This fMRI finding complements the physiology data because striatal BOLD is assumed to reflect, at least in part, afferent synaptic activity (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001) and the dopamine neurons project heavily to the striatum.

Although the aforementioned physiological responses appear to be related to the reward-prediction computations of RL, there is also an increase in dopaminergic phasic activity in response to arousing and/or novel stimuli that is seemingly unrelated to reward (Dommett et al., 2005; Doya, 2002; Horvitz, 2000; Redgrave, Prescott, & Gurney, 1999). A similar phenomenon has been recently observed in humans using fMRI (Bunzeck & Düzel, 2006). There are several reasons why this "novelty" or "saliency" response is said to be unrelated to reward-prediction error: (1) it appears very early,

☆ Contributed article.

* Tel.: +1 412 624 3191; fax: +1 412 624 9149.
E-mail address: patryk@cnbc.cmu.edu.

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.09.004

before the identity of the stimulus has been assessed, so that an accurate reward prediction cannot be generated; (2) it corresponds to an *increase* in neural activity (i.e., it is positive) for both aversive and appetitive stimuli; and (3) it habituates (Redgrave & Gurney, 2006). Indeed, these saliency/novelty responses of the dopamine-releasing neurons are most reliable when the stimuli are unpredicted and result in orienting and/or approach behavior (Schultz, 1998) regardless of the eventual outcome, highlighting the fact that they are qualitatively different from learned reward prediction. The challenge, therefore, has been to explain this apparent paradox (i.e., how novelty affects the reward-prediction error) within the theoretical framework of RL.

Kakade and Dayan (2002) attempted to do exactly this; in their article, they postulate two ways in which novelty responses could be incorporated into RL models of dopaminergic function—both involved the inclusion of new theoretical assumptions. The first assumption, referred to as *novelty bonuses*, involves introducing an additional reward when novel stimuli are present, above and beyond the usual reward received by the agent. This additional reward enters into the computation so that learning is based on the difference between the agent's existing reward prediction and the sum of both the usual reward from the environment and the novelty bonus. Thus, the novelty becomes part of the reward that the agent is attempting to maximize. The second assumption, termed *shaping bonuses*, can be implemented by artificially increasing the values of states associated with novel stimuli. Because the temporal-difference learning rule used in RL is based on the difference in reward-prediction between successive states, the addition of a constant shaping bonus to states concerned with the novel stimuli has no effect on the final behavior of the agent. However, a novelty response still emerges when the agent enters the part of the state space that has been "shaped" (i.e., that is associated with novelty).

Although the addition of each of these assumptions is sufficient to explain many observed effects of novelty, the assumptions also interfere with the progression of learning. As Kakade and Dayan (2002) point out, novelty bonuses can distort the value function (i.e., the values associated with each state by the agent) and affect what is ultimately learned because they are implemented as an additional reward that is intrinsically associated with novel states. The problem is that the agent learns to predict both the primary and novelty components of the reward. Although Kakade and Dayan point out that shaping bonuses do not cause this type of problem because they become incorporated into the reward predictions from preceding states, their addition is still problematic because shaping bonuses introduce biases into the way an agent will explore its state space. Thus, although these additional assumptions may explain how novelty affects the reward-prediction error in RL, they are problematic. Further, the explanations come at the cost of reducing the parsimony of modeling work that attempts to use RL to understand the behavior of real biological organisms.

The simulation reported below was carried out in order to test the hypothesis that a simple RL agent, without any additional assumptions, would develop a reward-prediction error response that is similar to the non-reward-related dopamine responses that are observed in biological organisms. An RL agent was given the task of interacting with two types of object – one positive and the other negative – that appeared at random locations in its environment. In order to maximize its reward, the agent had to learn to approach and "consume" the positive object, and to avoid (i.e., not "consume") the negative object. There were three main predictions for the simulation.

The first prediction was simply that, in order to maximize its reward, the agent would in fact learn to approach and "consume" the positive, rewarding objects while simultaneously learning to avoid the negative, punishing objects. The second prediction was slightly less obvious: that the agent would exhibit an orienting response (i.e., learn to shift its orientation) towards both negative and positive objects. This prediction was made because although the agent could "sense" the appearance of an object and its location, the positive or negative identity of the object (i.e., the cue that the agent would eventually learn to associate with the reward value of the object) could not be determined by the agent until *after* the agent had actually oriented towards the object. Finally, the third (and most important) prediction was related to the simulated dopaminergic phasic response in the model; this prediction was that, when the object appeared, the agent would exhibit a reward-prediction error that was computationally analogous to the phasic dopamine response observed in biological organisms, being positive for both positive *and* negative objects. This response was also predicted to vary as a function of the distance between the agent and the stimulus, which in the context of the simulation was a proxy measure for stimulus "intensity" or salience. As will be demonstrated below, these predictions were confirmed by the simulation results, demonstrating that the apparently non-reward-related dopamine responses can in principle emerge from the basic principles of RL. The theoretical implications of these results for using RL to understand non-reward-related activity in biological organisms will be discussed in the final section of this article.

1. Method

As already mentioned, RL algorithms specify how an agent can use moment-to-moment numerical rewards to learn which actions it should take in order to maximize the total amount of reward that it receives. In most formulations, this learning is achieved by using reward-prediction errors (i.e., the difference between an agent's current reward prediction and the actual reward that is obtained) to update the agent's reward predictions. As the reward predictions are learned, the predictions can also be used by an agent to select its next action. The usual *policy* (defined in Eq. (2)) is for the agent to select the action that is predicted to result in the largest reward. The actual reward that is provided to the agent at any given time is the sum of the immediate reward plus some portion of the value of the state that the agent enters when the action is completed. Thus, if the agent eventually experiences positive rewards after having been in a particular state, the agent will select actions in the future that are likely to result in those rewarded states; conversely, if the agent experiences negative rewards (i.e., punishment) it will avoid actions in the future that lead to those "punished" states.

The specific algorithm that determines the reward predictions that are learned for the various states (i.e., the *value function* V) is called *Value Iteration*[1] and can be formally described as:

[1] Another Reinforcement Learning algorithm, called *Trajectory Sampling* (Sutton & Barto, 1998), is frequently used instead of Value Iteration when the state space becomes so large that it cannot be exhaustively iterated or easily stored in a computer's memory. Rather than iterating over every state in the state space and applying the value function update equation based on the actions that appear to lead to the most reward, Trajectory Sampling works by following paths through the state space. Similar to Value Iteration, the actions leading to the most reward are usually selected from each state, but occasionally a random exploratory action is chosen with some small probability. Thus the algorithm is: *From some starting states, select an action leading to the most reward [e.g.,* reward $+ \gamma V(s')$*] with probability* ε*, or select a random exploratory action with probability* $1 - \varepsilon$*. Apply* $V(s) \leftarrow V(s) + \alpha$*[*reward $+ \gamma V(s') - V(s)$*] during non-exploratory actions from states.*

Besides overcoming the technical limitations of computational time and memory, Trajectory Sampling may be appealing because it may better reflect the manner in which real biological organisms learn: by exploring paths in a state space. On the task described in this paper, Trajectory Sampling yields results that are qualitatively identical to those obtained with Value Iteration. However, for conciseness those

For all possible states s,

$$V(s) \leftarrow V(s) + \alpha \left[\max_{\text{action}\in M} \{\text{reward} + \gamma V(s')\} - V(s) \right] \quad (1)$$

where s corresponds to the current state, $V(s)$ is the current reward prediction for state s that has been learned by the agent, $\max_{\text{action}\in M}\{\}$ is an operator for the maximum value of the bracketed quantity over the set of all actions M available to the agent, $V(s')$ is the agent's current reward prediction for the next state s', α is some learning rate (between 0 and 1), and γ is a discount factor reflecting how future rewards are to be weighted relative to immediate rewards. The initial value function was set so that $V(s)$ was 0 for all states s.

The value function $V(s)$ was implemented as a lookup table, which is formally equivalent to the assumption of perfect memory. Although function approximators such as neural networks have been used with some success to represent value functions (Baird, 1995), a lookup table was used to ensure that the results were not dependent on the types of generalization mechanism that are provided by various function approximators. The agent was trained for 1500 learning iterations over its state space. Because of the unpredictability of the identity of the objects, a value function update parameter of less than one ($\alpha = 0.01$) was used during the learning to allow for averaging of different outcomes. Finally, the discount factor was set to $\gamma = 0.99$ to encourage the agent to seek reward sooner rather than delay its approach behavior until the end of the trial (although changing it from a default value of 1 had no effect on the results reported here). In order to independently determine whether 1500 learning iterations were sufficient for learning to complete, the average amount of change in the learned was monitored and was found to have converged before this number of iterations.

After training, the specific algorithm that governs the agent's behavior (i.e., the *policy* of actions that it takes from each given state) is:

$$\pi(s) = \underset{\text{action}\in M}{\text{argmax}} [\text{reinforcement} + \gamma V(s')] \quad (2)$$

where $\pi(s)$ is the action the agent will select from state s, and the right side of the equation returns the action (e.g., change of orientation, movement, or no action) which maximizes the sum of the reward and the discounted value of the resulting state s'.

In the simulation that is reported below, all of the states that were visited by the agent were encoded as 7-dimensional vectors that represented information about both the external "physical" state of the agent and its internal "knowledge" state. The physical information included both the agent's current position in space and its orientation. The knowledge information included the position of the object (if one was present) and the identity of that object (if it had been determined by the agent). The specific types of information that were represented by the agent are shown in Table 1.

There were a total of 21,120 states in the simulation[2]. However, the states in which there was an unidentified positive and unidentified negative object are, from the perspective of the agent, identical, so there are therefore only 16,280 distinct states. Thus, during each iteration of learning, it was necessary to visit some of those "identical" states twice to allow for the fact that half of the time they might be followed with the discovery of a positive object, and half of the time they might be followed with the discovery of a negative object.[3]

At the beginning of each simulated testing trial, the agent was placed in the center of a simulated linear 11 × 1 unit track with five spaces to the "east" (i.e., to the right) of the agent and five spaces to the "west" (i.e., to the left) of the agent. As Table 1 shows, the agent's state-vector included an element indicating its current location on the track (i.e., an integer from 0 to 10), as well as an element (i.e., a character "*n*", "*s*", "*e*", or "*w*") representing its current orientation (i.e., north, south, east, or west, respectively). The agent's initial orientation was always set to be "north," and no other object was present in the environment (i.e., the value of "OBJECT" in the agent's state-vector was set to equal to "0").

During each time-step of the simulation, the agent could perform one of the following actions: (1) do nothing, and remain in the current location and orientation; (2) orient to the north, south, east or west; or (3) move one space in the environment (east or west). The result of each action took place on the subsequent simulated time-step. All changes in the location and/or orientation of the agent in space occurred through the selection of actions by the agent. However, during every time-step of the simulation, even when a "do nothing" action was selected, time was incremented by 1 until the end of the trial (i.e., time-step 20).

The agent's environment was set up so that half of the time, an object appeared at a random location (but not in the same location as the agent) after ten time steps; 50% of the objects were positive (represented by a "+"; see Table 1) and 50% of the objects were negative (represented by a "−"). The delay before the object appeared was introduced to allow the observation of any behavior the agent may have exhibited before the appearance of the object. If the agent was not oriented towards the object when it appeared, then the element representing the "OBJECT" identity in the agent's state vector was changed from "0" to "?" to reflect the fact that the identity of the object that was now present was currently unknown. However, if the agent was oriented towards the object, then on the subsequent time-step the "OBJECT" element was set to equal to the identity of the object, so that "0" became either "+" or "−" for positive and negative objects, respectively.

If the agent moved to an object's location, then during the next time-step the object vanished. If the object had been positive, then the agent's "CONSUMED" flag was set equal to true and the agent was rewarded (reward = +10); however, if the object

results are not reported here in detail. Value Iteration was selected for the simulation in this paper for two main reasons. First, because Trajectory Sampling involves stochasticity in the selection of trajectories; the large amount of branching that is due to the many possible sequences of actions in this task may result in agents that lack experience with some states unless the exploration–exploitation parameter (i.e., ε-greediness (Sutton & Barto, 1998)) is carefully selected. This lack of experience with particular states can be disruptive of an agent's performance when a lookup table memory structure is used because of the lack of generalization of value to similar (but possibly unvisited) states. Thus, it was preferred to take advantage of the exhaustive exploration of state space that is guaranteed with Value Iteration. Second, the use of Value Iteration obviated the need to specify that additional exploration–exploitation parameter, thereby simplifying the simulation. Note that Trajectory Sampling can ultimately approximate Value Iteration as the number of trajectories approaches infinity (Sutton & Barto, 1998).

[2] The number of 21,120 states can be calculated as follows: 11 possible agent locations × 4 possible agent orientations × (10 time-steps before an object might appear + 10 time-steps where no object appeared + 10 time-steps where the agent had been positively reinforced + 10 time-steps where the object had been negatively reinforced + 11 possible object locations × (10 time-steps with a positive identified object + 10 time-steps with a negative identified object + 10 time-steps with an unidentified positive object + 10 time-steps with an unidentified negative object))].

[3] The existence of these "hidden" states must be considered during training because Value Iteration only looks "one step ahead" from each state in the state space. The fact that states with negative and positive unidentified objects are effectively identical would prevent learning about and averaging the values in the two different subsequent states in which either the positive or negative object becomes identified. A Trajectory Sampling approach on the other hand maintains the hidden state information (i.e., the identity of the unidentified stimulus) throughout the trial and so with that variant of RL the hidden states are not a concern.

Table 1
The dimensions used in the RL simulations and the possible values of those dimensions.

Dimension no.	Description	Possible values
1	Position of agent	Integer (0–10)
2	Orientation of agent	Character ("*n*", "*s*", "*e*", or "*w*")
3	Position of object	Integer (0–10)
4	Identity of object	Character ('0", "?", "+", or "−")
5	"Shocked" by object	Boolean (true or false)
6	"Consumed" object	Boolean (true or false)
7	Time since trial onset	Integer (0–20)

had been negative, then the "SHOCKED" flag was set to true and the agent was punished (reward $= -10$). (Note that the flags were set in this way regardless of whether the agent had or had not identified the object; e.g., the agent could consume an object without ever orienting towards it.) On the subsequent time-step, the "SHOCKED" or "CONSUMED" flag was cleared. The agent was also given a small penalty (reinforcement $= -1$) for each movement or orienting action, and received no reward or punishment (reinforcement $= 0$) if it performed no action.

Both the overt behaviors (i.e., orienting and movement) and a measure of reward-prediction error were quantified for the agent. The overt behavior (i.e., the list of actions selected by the agent) was used as an indication of whether the task had been learned. The measure of reward-prediction error was used to test the hypothesis about the emergence of the non-reward dopaminergic phasic signal. The reward-prediction error, δ, was measured at the time t of the appearance of an object by subtracting the reward prediction at the previous time-step, i.e., $V(s)$ at time step $t-1$, from the reward prediction when the object appeared, i.e., $V(s)$ at time t, yielding the quantity $\delta = V(s_t) - V(s_{t-1})$.

2. Results

Simulated behavior. The overt behavior of the agents was first quantified. The results of this analysis showed that, after training, the agent approached and obtained positive reinforcement from all of the positive objects and never approached any of the negative objects. Together, these results provide behavioral confirmation that the agents learned to perform the task correctly. This conclusion is bolstered by the additional observation that, during the trials when no object appeared, the agent remained motionless. As predicted, the agent oriented to both positive and negative objects.

Simulated reward-prediction error. The central hypothesis of this paper is that the appearance of an unpredictable stimulus will consistently generate a positive reward-prediction error, even if that object happens to be a "negative" object that is always punishing. In support of this hypothesis, the agent exhibited a positive reward-prediction error whenever an (unidentified) object appeared, but not when nothing appeared. Also consistent with the central hypothesis is the fact that the magnitude of the agent's phasic response (δ, measured as described in the Method section) was sensitive to the simulated "intensity" of the stimulus, defined using the distance between the agent and the object (see Fig. 1). A regression analysis indicated that the magnitude of δ was inversely related to the distance from the object, so that closer objects caused a stronger response ($r = -0.999$, $p < 0.001$; $\beta = 0.82$). This negative correlation was caused by the small penalty (reinforcement $= -1$) that was imposed for each movement that the agent was required to make in order to approach the positive object, consume it, and thereby obtain reward.

Given that positive and negative objects appeared in this simulation with equal probability ($p = .25$), the question arises: Why was the agent's reward-prediction error signal positive at the time of the object's appearance? Reasoning along the lines of Kakade and Dayan (2002), one might predict that the signal should reflect the average of all of the learned rewards from such situations, and therefore be equal to zero. The key to understanding this result is to note that not only does RL make an agent less likely to choose actions that result in negative reinforcement, it also makes an agent less likely to enter states which eventually lead to negative reinforcement. This results in a kind of "higher-order" form of learning that is depicted in Fig. 2 and described next.

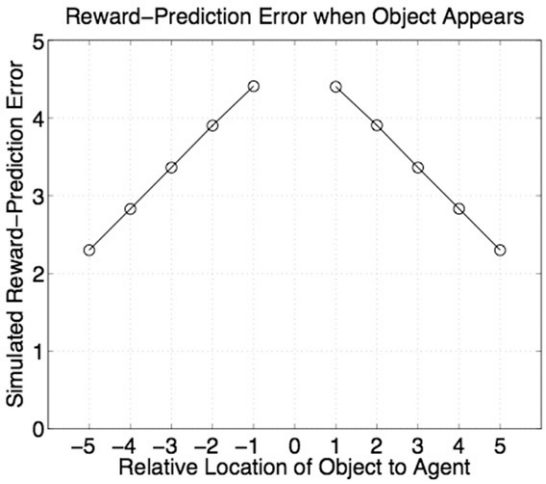

Fig. 1. This figure shows the reward-prediction error (i.e., δ) when the object appeared as a function of the location of the object relative to the location of the agent. The responses are identical for both positive and negative objects. When no object appeared, the response was 0. Note that the size of the response is inversely correlated with distance from the object when it appeared. There is no data for location 0 because the object would be immediately consumed had it appeared there.

At the beginning of learning (see Fig. 2A), the agent orients to both "+" and "−" objects, approaches them, and is both rewarded and punished by consuming each type of object. If the agent's learned state values were unable to influence the agent's actions (see Fig. 2B), then the agent would continue to approach and consume the objects. The appearance of the cue would then predict an average reward of 0 and there would not be a sudden increase in reward-prediction error. However, the agent in this simulation *does* use learned state values to influence its actions (see Fig. 2C), and although the agent still has to orient to the unknown object to determine its identity, it will no longer consume a negative object if it approached it (as it might if trained with a random exploration algorithm like *trajectory sampling* (see Footnote 1)). Furthermore, because temporal-difference learning allows the negative reward prediction to "propagate" back to preceding states, and because there is a small cost for moving in space, the agent learns to avoid approaching the negative object entirely. Thus, after this information has been learned, the value of the state when the object first appears (indicated as "V" in the first circle in each sequence) is not based on the average of the positive and negative outcome state values, but is instead based on the average of positive and the "neutral" outcome that is attained once the agent learns to avoid the negative objects. This is why the

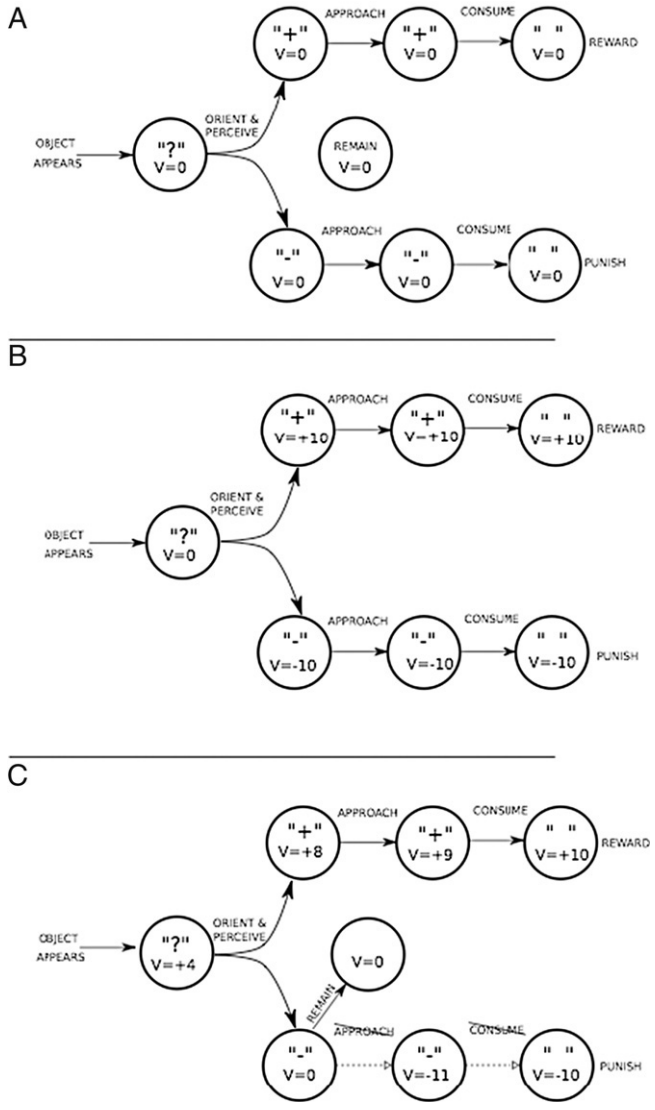

Fig. 2. Illustration showing how an RL agent develops positive reward-prediction error when it is trained with both rewarding and punishing stimuli in its environment and is able to choose whether to approach and consume them. (A) The situation before learning: all states begin with a value of 0, and the agent has not yet learned the rewarding and punishing values of the "+" and "−" stimuli. (B) A temporal-difference learning algorithm is used without allowing those values to affect the actions of the agent: the agent learns reward predictions based on experience but is unable to use the learned values to influence its own behavior. In this case, the reward-prediction error when the object appears will be the average of the positive and negative outcomes (i.e., 0). (C) We show what happened in the present simulation. The agent quickly learns to avoid consuming, or even approaching, the negative object. The result is that when the stimulus appears, the reward-prediction error is based on the average of the positive stimulus and a neutral outcome in which the negative stimulus is avoided and is consistently greater than 0. Note: This figure does not illustrate the fact that in the present simulation, more distant objects require more actions (and therefore more intervening small punishments) in order to approach them. That fact is what causes the decreasing magnitude of the novelty/saliency response for objects that appear more distantly from the agent (e.g., as plotted in Fig. 2).

average of all rewards actually obtained by the trained agent was greater than zero, and explains why the agent's reward prediction (and therefore reward-prediction error when the object suddenly appears) was a net positive. This is illustrated in Fig. 3. In fact, as long as the agent can learn to change its behavior and avoid the negative object, the value of the negative object is ultimately irrelevant to the final behavior of the agent and the magnitude of the novelty/saliency response.

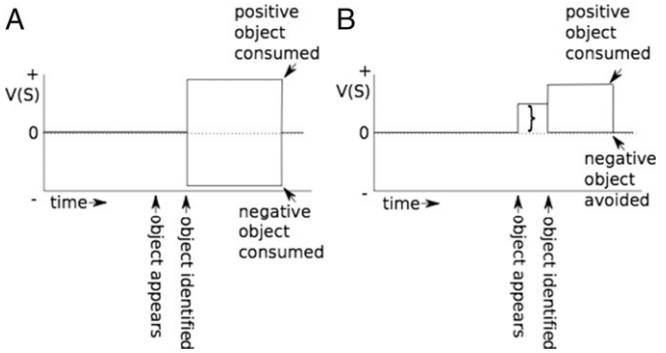

Fig. 3. (A) Demonstrates the changes in reward prediction that would have occurred if RL did not result in higher-order learning (i.e., if the agent could not take measures to avoid the negative outcome), so that the agent was forced to consume all the objects that appeared. When an object appears, the agent does not know yet its identity but generates a net reward prediction of zero because the reward prediction is the average of the positive and negative consequences (i.e., half the time the object has been positive, and half the time it has been negative). (B) Demonstrates what actually occurred: higher-order learning permitted the agent to avoid the negative object, so that when the stimulus appeared, the agent had a greater-than-0 reward prediction because it is the average of the positive outcome and null outcomes. The curly brace spans the difference in reward-prediction values that represents this reward-prediction error.

The simulation results are critically dependant on three assumptions. First, the stimuli had to be "salient" in that the magnitude of the reinforcement predicted by the initial cue was sufficiently large (e.g., +10) relative to the costs of orienting and approaching (e.g., −1). If the magnitude had been relatively small, the agent would not have learned to orient, nor would it have generated the positive reward-prediction error response. Second, a delay prior to recognizing the stimuli was also necessary. (Delay is a proxy for "novelty" under the reasoning that a familiar stimulus would be quickly recognized.) Without a delay, the agent would have simply generated the appropriate positive or negative reward prediction error for the actual perceived object. Finally, the agent's behavior had to be determined by the values that it had learned. If the agent could not use learned values to control its own behavior (i.e., whether to approach the stimuli), then its reward prediction when an object appeared would have equaled 0, the average of the equiprobable positive and negative outcomes.

3. General discussion

The simulation reported in this article demonstrated that a positive reward-prediction error occurs when an unpredictable stimulus, either rewarding or punishing, appears but cannot be immediately identified. Furthermore, the simulation indicated that the size of the reward-prediction error increases with proximity of the stimulus to the agent, which in the context of the simulation is a proxy for stimulus intensity and is thus related to salience. In the theoretical framework of RL, reward predictions are normally understood to reflect the learned value of recognized stimuli, or of the physical and/or cognitive states of an agent (Reichle & Laurent, 2006). However, the reward-prediction error reported here has a qualitatively different interpretation because it is generated before the agent has recognized the object. Together, these results support the hypothesis that RL principles are sufficient to produce a response that is seemingly unrelated to reward, but instead related to the properties of novelty and saliency. This conclusion has several important ramifications for our general understanding of RL and for our interpretation of RL as an account of reward learning in real biological organisms.

First, the reward prediction that is generated by an RL agent when an unidentified stimulus appears is not necessarily a strict average of the obtainable rewards as suggested by Kakade and

Dayan (2002), but can in fact be greater in magnitude than that particular average. Kakade and Dayan would predict that the average reward prediction should be equal to zero because, the trials were rewarded and punished equally often. This surprising result emerged because the agent learned in an "on-policy" manner; that is, the agent learned not only about negative outcomes, but also about its ability to avoid those outcomes. This ability of the reward system to cause an agent to avoid negative outcomes should be carefully considered in translating our understanding of RL to real organisms. This fact is potentially even more important given the apparent asymmetry in the capacity of the dopaminergic phasic response to represent positive reward prediction error better than negative reward prediction error (Niv, Duff, & Dayan, 2005). It may be sufficient to indicate that a particular sequence of events leads to a negative outcome, but that for the purposes of action selection, the magnitude of that outcome is unimportant.

A second ramification of the current simulation is that the novelty response may emerge from an interaction between perceptual processing systems and reward-prediction systems. Specifically, the novelty response may be due to a form of similarity between novel objects and objects that have not yet undergone complete perceptual processing[4]. In this simulation, novelty was implemented by introducing a delay before the object's identity (and consequently its rewarding or punishing nature) became apparent to the agent. This was done under the assumption that novel objects take longer to identify, but this assumption also resulted in the positive and negative objects being perceived similarly when they first appeared (i.e., they were both encoded as "?"). In contrast, Kakade and Dayan (2002) suggest that novelty responses and "generalization" responses are essentially different despite being manifested similarly in the neurophysiology data.

A third ramification of the current simulation results is that they show that the additional assumptions of novelty and shaping bonuses that were proposed by Kakade and Dayan (2002) are not necessary. Instead, novelty-like responses can emerge from realistic perceptual processing limitations and the knowledge of being able to avoid negative outcomes. This is fortunate because, as pointed out by Kakade and Dayan, novelty bonuses distort the value function that is learned by an agent, and shaping bonuses affect the way in which agents explore their state spaces. The inclusion of either of these assumptions thus reduces the parsimony of models based on RL theory. Interestingly, the results presented here also help explain why the biological novelty response might not be disruptive to reward-based learning in real organisms: the novelty response is in fact already predicted by RL. That is, the novelty response reflects behaviors and reward predictions that are inherent in an agent that has already learned something about its environment.

An alternative (but not mutually exclusive) interpretation of the present simulation results is that there is indeed an abstract (perhaps cognitive) reward that agents obtain by orienting towards and identifying objects. In studies of dopaminergic activity, positive phasic responses can occur to unanticipated cues that are known to predict a reward. This simulation, however, demonstrates how these kinds of responses can also occur in response to a cue that could ultimately predict either reward or punishment. The only consistent benefit that is predicted by the cue is the gain in information obtained when the agent determines the identity of the object. Thus, if there is a valid, learned "reward prediction" when the unidentified object appears, it is one that is satisfied after the agent obtains the knowledge about whether to approach or avoid the stimulus. The value of this information is based not on the average of the obtainable outcomes, but is instead based on the knowledge of the effective outcomes – that the agent can either consume the positive reward or avoid the negative reward (see Fig. 2).

Finally, it is important to note that the opportunities to take particular actions (e.g., to orient) may themselves take on rewarding properties through some generalization or learning mechanism not included in this simulation. For example, the very act of orienting and determining "what's out there" could become rewarding to an organism based on the association between that action and the above-demonstrated emergent, always-positive reward-prediction error when new stimuli appear. A similar idea has been recently advanced by Redgrave and Gurney (2006) who hypothesize that an important purpose of the phasic dopamine response is to reinforce actions that occur before unpredicted salient events. The results here are not incompatible with that hypothesis, however it should be noted that Redgrave and Gurney's hypothesis is not directly tested in this simulation because no actions (i.e., exploration) were required of the agent in order for the salient event (the appearance of the object) to occur. However, the simulated phasic signal coincided with the time of the orienting response, suggesting that the two may be strongly related.

In closing, this article has demonstrated that RL principles can be used to explain a type of seemingly non-reward related activity of the dopaminergic neurons. This result emerged from the fact that the temporal-difference learning rule (such as that used by Kakade and Dayan (2002)) was embedded in a simulation in which the agent could select actions that had an effect on the eventual outcome. In the simulation, the agent learned that the outcome of orienting to an object that suddenly appeared could always either be rewarding or neutral because the negative outcome could be avoided. Therefore when the agent had an opportunity to orient, its reward-prediction error was always positive, computationally analogous to the novelty and saliency responses observed in biological organisms.

Acknowledgments

The work described in this article was supported by NIH R01 HD053639 and by NSF Training Grant DGE-9987588. I would like to thank Erik Reichle, Tessa Warren and an anonymous reviewer for helpful comments on an earlier version of this article.

References

Baird, L.C. (1995). Residual algorithms: Reinforcement Learning with function approximation. In A. Priedetis, & S. Russell (Eds.), *Machine learning: Proceedings of the twelfth international conference*.

Bunzeck, N., & Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, *51*(3), 369–379.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., et al. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science*, *307*(5714), 1476–1479.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506.

Gillies, A., & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, *15*(5), 762–770.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656.

[4] One potential objection to the present work is that the orienting response appears to be hard-wired in the mammalian brain, for example, in projections from the superior colliculus (Dommett et al., 2005; Redgrave et al., 1999). In the present simulation, the agents were not hard-wired to orient to objects but instead learned an orienting behavior that permitted the eventual selection of an action (e.g., either approach or avoidance) that maximized reward. Similarly to hard-wired responses, these orienting behaviors occurred very rapidly, before the objects were identified, and were directed towards all objects. The goal of this work was not to make the claim that all such responses are learned, but rather that they can co-exist within the RL framework. Nevertheless, it would be interesting to investigate whether reward-related mechanisms might be involved in setting up connectivity in brainstem areas in order to generate this phasic dopamine response.

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559.
Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157.
McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.
Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, *1*, 6.
O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.
Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975.
Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in Neurosciences*, *22*(4), 146–151.
Reichle, E. D., & Laurent, P. A. (2006). Using Reinforcement Learning to understand the emergence of "intelligent" eye-movement behavior during reading. *Psychological Review*, *113*(2), 390–408.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge, MA: MIT Press.
Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*(8), 887–893.

4.3 SUPPLEMENTAL DISCUSSION

The simulation work demonstrated the theoretical connection between novelty/saliency responses and the reward-prediction error signals that are used in RL algorithms. The results have important implications for the involvement of the basal ganglia in cognition, computation, and behavior.

Cognitively, the simulation results are important for understanding the involvement of the basal ganglia in the cognition of novel and salient stimuli. Stimuli that are novel or salient may generate activation in the basal ganglia through multiple means. One means for generating such signals is in the neocortex itself. Computational work has demonstrated that a network of learning excitatory and inhibitory integrate-and-fire neurons that are similar in distribution to the neocortex will exhibit an increase in activity when novel stimuli are presented (Ho and Rouat, 1998). Thus, neocortical regions involved in processing sensory data may exhibit activation when novel stimuli are presented. The other means for generating such signals is demonstrated by the present experiment simulations: If an object has certain qualities that indicate that it may lead to positive and/or avoidable negative reward, a saliency response will be generated.

Computationally, the results have important implications in the use of RL to interpret the phasic dopamine response.

An additional possible point of contact between saliency and reward-prediction error is in their role in the cognitive act of "selection", which is a generalization of the physical act of orienting towards an external stimulus. This point was indirectly touched upon in the manuscript through the assumption that the agent had to orient to the stimulus that appeared prior to interacting with it (i.e., approaching it or avoiding it). More generally, to decide what action to perform, an organism may need to perform an 'orienting' operation to an unexpected or intense surge of neuronal activity, particularly if doing so is likely to lead to reward (or to avoiding punishment). Part of this "cognitive orienting" may take the form of amplifying the neuronal activity, similarly to what is thought of as an attention shift. The mechanism for achieving this "cognitive orienting" or attention shift may involve dopamine. Dopamine is known to act through D2 receptors on corticostriatal projections to suppress less active inputs (Bamford et al., 2004) while allowing the stronger (i.e., more salient?) inputs to remain active. Thus, the release of dopamine appears to be a mechanism

that mediates the capture and/or direction of attention to salient objects, and therefore comprises a form of "orienting".

These theoretical results suggest that there may be a relationship between saliency, attention, and reinforcement – specifically that the activation of corticostriatal connections and their interaction with dopamine could mediate all three of these phenomena. Thus, an account for a common currency of saliency, attention, and reinforcement leads to three hypotheses for how corticostriatal projections might be recruited: First, corticostriatal projections could be triggered by stimuli that are inherently salient, i.e., by intense (e.g., loud, bright) stimuli. These stimuli would lead to significantly stronger than anticipated activation of cortiostriatal projections. Second, neocortical processing may make the stimuli more salient even if those stimuli are not more physically intense (i.e., local computations in neocortex could increase its response to particular stimuli, e.g., sharp contrasts). Third, the corticostriatal synapses themselves may have been potentiated by previous reward (i.e., RL) associated with the stimulus. So while the stimulus itself is neither intense nor configured to generate a particularly strong neocortical response, it triggers a strong striatal response through potentiated corticostriatal projections. Fourth, some kind of top-down set from prefrontal cortex could act through a combination of (weak) cortico-cortical connections, direct excitation of the striatum, and/or triggering of phasic dopamine release, to result in an increased striatal response to the stimulus.

This theoretical work therefore suggests that further experiments should be aimed at investigating potential interaction between reinforcement, attention, and saliency.

One final point can be made about the value of this kind of simulation through an example demonstrating how the dissociation between reward-prediction error and saliency signals can be studied: A simulation similar to the one described above can be used to elucidate scenarios in which agents are never rewarded, yet still develop a saliency response. Although this is not demonstrated in the present article, it is possible to complete a simulation in which agents only ever receive negative reinforcement, i.e., are never rewarded. In this simulation, objects are either neutral (i.e., never provide reward), or are aversive with a "blast" radius over which the negative reward is distributed after a time delay. In this simulation, the agent will continue to display the results documented above: the agent will rapidly orient towards the unidentified object, and then exhibit

fear-like distancing behavior from the object in the case that it is aversive. Critically, at the time of the appearance of the unidentified object, the error in reward-prediction will still be positive.

5.0 GENERAL DISCUSSION AND CONCLUSION

This dissertation presented experimental, computational and theoretical work focused on investigating the relationship between the basal ganglia, and computations required by a formal theory of learning called RL. The experimental work involved behavioral and neuroimaging studies involving reinforcing human participants for performing hand movements, eye movements and covert attention shifts. The computational work involved implementing numerical simulations of RL agents to make predictions about the brain activity of the human participants as they performed the experimental tasks. Finally, the theoretical work involved performing qualitative computational experiments aimed at interpreting the meaning of the phasic dopamine signal in cases that appeared unrelated to reward. Together these results point to a framework for understanding basal ganglia function in which distributed control of action can be reinforced and shaped by factors relevant to the organism as a whole. Much like an organization that gives bonuses to its employees for good performance, or a conductor that provides feedback to a large orchestra, the brain may release dopamine to encourage all participating brain regions to perform similarly in future similar circumstances.

The discussed behavioral results showed that simple actions such as hand movements and eye movements can be rewarded and learned to the same asymptotic performance levels and at the same learning rates. This suggests a common mechanism of reinforcement, which is consistent with ideas about dopamine and the possible involvement of the nucleus accumbens in reward processing. In the second part, the behavioral studies are adapted into a functional MRI study to test the hypothesis that the basal ganglia is involved in reward-prediction error computations for these different types of action. Results show that in spite of no behavioral differences in learning, signals correlating with reward-prediction error are located in different regions of the striatum depending on the type

of action being reinforced. These findings suggest that these different anatomical regions, sharing similar connectivity as parts of loops with different cortical regions, are implementing the same algorithm.

Most interestingly, the results show that the RL computations that apply to physical actions also apply to cognitive actions like covert shifts of visual attention.

These findings support earlier computational studies in which it is assumed that arbitrary actions, such as eye movements and covert attention shifts, can be reinforced. Those simulations demonstrate that intelligent patterns of eye-movement behavior, such as those seen during reading, can emerge in such cases. This powerful approach is applied to help distinguish between two sophisticated, competing classes of eye-movement control during reading. In the fourth part, another computational study demonstrates how Reinforcement Learning could be used to understand non-reward-related responses in the basal ganglia, such as novelty and saliency responses.

The implications of these findings were discussed with respect to the idea that Reinforcement Learning is a good framework for understanding the function of the basal ganglia. The thesis concludes that the different parts of the striatum are implementing the same computation for different types of action, which are represented in different areas of cortex. Taken together, these findings support the joint neuroanatomical-computational hypothesis the cortico-basal ganglia loops implement a computation which contributes to Reinforcement Learning. This contributes to neuroscience theory suggesting that any neuronal pattern, representing either physical or cognitive states or actions in cerebral cortex, can be positively or negatively reinforced according to this distributed mechanism. Together, the results of this dissertation suggest that RL should be considered in the development of future models of basal ganglia function.

5.1 INNOVATIONS AND CREATIVE APPROACH

The successful completion of this dissertation project required innovation and creativity in several respects, namely, in methods development, experiment design, and data analysis approaches. One innovation that was critical to the success of the fMRI experiment was the development and use of

real-time sound cancellation technology to monitor the participants' oral responses (see Appendix B). Sound-cancellation is necessary in experiments requiring oral reports because MRI scanners emit loud, repetitive noise that results from the movement of gradient magnets that is necessary to scan particular parts of the brain (Huettel et al., 2004). This problem had to be overcome because the experiment design required the experimenter to be able to hear the participants because, as described in the experiment methods, on each trial the participant reported the probe letter they observed and this letter was entered by the experimenter into a computer so that the experiment software could provide the appropriate reinforcement feedback to the participant[1].

Another innovation was the use of real-time eye-tracking technology for feedback during both behavioral and fMRI scanning. Real-time feedback from eye-tracking was important because participants were required to keep their fixation at the center of the screen during hand movement and covert attention shift portions of the experiment, to maximally control which actions were being reinforced. In those conditions, it was of particular interest to reinforce hand movements and covert attention shifts without reinforcing eye movements. The eye-tracker in the fMRI scanner was connected directly to the experiment display computer through a serial link so that it could provide real-time feedback to the participants, warning them and aborting the trial when they moved their eyes from the center. (This task was the first task at the BIRC scanning center to use information about gaze in real time.)

With respect to design, this experiment is innovative because it is the first instance of a paradigm in which a single task was used to study reward-based learning of three different kinds of actions,

[1]Originally, the experiment relied on participants using their hands to indicate the probe letter, however, behavioral pilot work showed that this caused interference in learning hand movements: hand movements were learned slower compared to the other actions, and behavior was fit less well by RL agents. To overcome the problem of interference from the dual use of hand movements, an additional (fourth) action – an oral response – was used in all three conditions (i.e., along with hand movements, eye movements and covert attention shifts). The use of an additional action prior to obtaining reward was a potential risk because the reward could have become associated with the oral response rather than with the particular action. However, RL algorithms that involve learning reward-predictions such as temporal-difference learning reward-prediction posit that learned reward predictions transfer earlier in time to the most reliably predictive state or cue (Schultz, 1998; Sutton and Barto, 1998). Thus, the experiment design assumes that reinforcement will become associated not with the oral response but rather with the stimulus and movement representations. Further, the particular oral responses were randomized on each trial. That is, given a particular cue (e.g., *red boxes*) and action (e.g., *upward hand movement*), the correct oral response was random (e.g., it could have been *A*, *I*, *O*, or *U*.) It was assumed therefore that voice responses would not cause any interference in the hypotheses that were being tested. Indeed, the fact that the results show that reward-prediction errors were associated with action-specific regions of the striatum, rather than all in the same region, reconfirms the assumption that reward prediction transfers to the earliest predictive stimulus.

including both motoric and non-motoric actions. Furthermore, this task combined computational, behavioral, and neuroimaging approaches to generate and test quantitative predictions about both behavior and about neuronal reward-related activity associated with motoric and non-motoric actions. This was possible by implementing RL models and training them to perform the same actions as those performed by the human participants from one trial to the next. Thus, the manner in which the data were analyzed were innovative as well because a single form of computational model was used to make brain and behavioral predictions about both physical and cognitive action.

5.2 FUTURE DIRECTIONS

The general experimental design used in this dissertation lays some groundwork for a number of future experiments that could be undertaken using similar methods.

Using Reward to "Tag" and Observe Cognitive Actions. The paradigm that was used in this dissertation establishes a method for examining the reinforcement of further cognitive actions. Future variants of this task could serve to map cognitive functional zones within the striatum by rewarding participants for performing cognitive actions presumed to involve other neocortical regions, such as mental rotation, manipulations of sentence structures, shifts in working memory, etc. Associating specific reward predictions with particular cognitive operations could serve as a novel method for detecting when those cognitive operations occurred. This method would be particularly relevant for situations in which the neocortical representations of those actions are overlapping or highly diffuse throughout the brain. An example of the former case can be seen in the present data: covert attention shifts and saccadic eye movements share overlapping representations in the *frontal eye fields* (FEF) that have yet to be resolved using fMRI (and are certainly challenging to resolve using electrophysiology; e.g., Murthy et al., 2001; Thompson et al., 2005).

In the present results, the differential localization in the anterior striatum of activations for covert attention shifts and saccadic eye movements suggest that actions with overlapping neocortical representations can be dissociated by examining striatal activation. However, eye movements and covert attention shifts are oculomotor and cognitive "actions", respectively. An interesting

question that should be investigated is whether differences in the location of striatal activity can be used to discriminate between two *cognitive* "actions". For example, participants could be trained to perform a number of linguistic operations, such as checking subject-verb agreement, or recalling an antecedent. Each of these operations could be associated with, for example, low, medium or high rewards. Then, participants could be allowed to read a passage and predictions from a model of real-time language processing could be tested in an event-related fashion. That is, when the model predicts a high reward operation, activity in the striatum could be deconvolved at that time to determine whether a reward-prediction error associated with the cognitive linguistic operation occurred.

Reinforcing "Limbic" Actions in the Ventral Striatum. The current research focused on learning of a limited range of actions that are represented in the connections between neocortex and the basal ganglia. This limited the current investigation to making predictions about activity in dorsal striatum. It is known, however, that there are connections to the basal ganglia from archicortical systems (i.e., hippocampus) as well as the cerebellum. Thus an important future direction of this research method would be to test the approach by driving activity in these other systems. Ventral striatum is predominantly innervated by limbic cortical inputs. An important question to evaluate is whether RL-like learning applies to "limbic actions" such as the expression of particular emotions (i.e., amygdala), the recognition of being in a particular location (e.g., hippocampus), or more abstractly the recognition of the value of an object. Existing electrophysiological work suggests that the hippocampus has strong functional projections to the striatum (e.g., Lansink et al., 2009; Pennartz et al., 2004). Further, it appears that these episodic and reward learning systems interact (Johnson and Redish, 2005). In that simulation, a rat's behavior in a rewarded T-maze task improved after a rest period as though RL had continued in the absence of physical exploration of the environment. These results suggest that hippocampal replay of episodes (i.e., trajectories) to the basal ganglia may have mimicked a form of "trajectory sampling" in mental space without the rat needing to physically explore those trajectories in the environment. In would be particularly interesting to train human participants on trajectories leading to different amounts of reward and monitor their brain activation during a consolidation period, and attempt to correlated BOLD in hippocampus and ventral striatum with later improvements in behavioral performance.

Error Correction of Reward Predictions As discussed in the section 1.5 (page 13), another interesting future direction would be to test the hypothesis that error-correcting capabilities of the cerebellum can be applied to reward-prediction error learning. This is a particularly interesting question because a direct connection between the cerebellum and the basal ganglia has been recently discovered (Hoshi et al., 2005). One prediction to test would be that the cerebellum, and the cerebellar input to the basal ganglia, would be active when some change in perception or context required reward adaptation or adjustment in a range of reward predictions. Other predictions to maximize the involvement of the cerebellum in reward-realted learning could be generated by training Baird's (1995) algorithm on tasks and noting which tasks require larger contributions from the error-correcting gradients, and then creating experimental versions of those tasks for human participants.

Improvements in the Current Experiment Design. In future studies, a number of improvements could be made to the current experiment design. An extra condition in which the mapping between color and action was randomized on every trial would help to control against any kind of associative learning that may have taken place (i.e., between stimulus and cognitive 'direction' of action). This control would have been exactly like the current control condition (CONTROL) except that the mapping between color cue and correct direction would have been randomized on trial to trial (i.e., would not have been consistent.) Another potential improvement might be to randomize the onset of the stimuli and the probe letters to prevent participants from doing multiple and/or anticipatory eye movement and/or attention shifts. We found that participants made anticipatory moves to some of the targets, and this may have resulted in a temporal smearing, and consequently a reduction in power, of the signals of interest. Finally, introduction of jitter between the stimulus onset and reward feedback onset would help to separately deconvolve the reward prediction at the beginning of the trial from the reward-prediction error at the end of a trial.

5.3 CONCLUSION

The work presented in this dissertation was aimed at investigating whether computations performed in the basal ganglia were consistent with predictions made by RL theory. Experimental work was carried out to test specific hypotheses regarding the computation of reward-prediction error in distinct regions of the striatum. The dissertation was motivated by prior anatomical, electrophysiological, and computational neuroscience findings. Prior anatomical studies highlighted questions regarding the pattern of basal connectivity with not just the motor control regions of the prefrontal cortex, but with many other regions including non-motor (i.e., cognitive) regions of neocortex. Prior electrophysiology studies showed that dopamine neurons, which project extensively through the basal ganglia and neocortex, exhibit signals that are consistent with reward-prediction error. Prior computational simulations showed that simple RL principles could give rise to complex adaptive behaviors, and could also explain the emergence of non-reward related signals in phasic dopamine signals. All of these findings suggested that the basal ganglia might implement computations supporting the RL of physical and cognitive actions. However, no experiment had yet been performed in which human participants were reinforced for performing different actions. Thus there was little evidence to support the idea that the basal ganglia might perform a particular computation regardless of the information that was fed into that system.

Thus this dissertation raised the question: Are the basal ganglia involved in the RL of both physical and cognitive actions? The results of this work suggest the conclusion that yes, the basal ganglia are a system which is capable of performing a particular set of operations on action representations in neocortex, whether those representations are for physical or cognitive actions. The fMRI results suggest that the basal ganglia are able to compute and store, likely at the synapses between corticostriatal axons and medium spiny neuron dendrites, reward predictions associated with neocortical action representations. Furthermore, the behavioral results suggest that the BOLD fMRI response in the striatum does not simply reflect the action of dopamine there, but rather interaction of dopamine with corticostriatal afferent information. Indeed, a possible conclusion from this work is that the fMRI BOLD response is indicative of metabolic processes involved in plasticity in the striatum. A result that supports this conclusion is that both positive- and negative-reward

prediction errors, which should both result in plastic changes in the form of increases or decreases in the reward predictions hypothesized to be stored in corticostriatal synapses, both resulted in positively-correlated BOLD responses in the striatum.

The experiment design and the results shown here may lead to a number of future investigations of the RL of physical, cognitive, and potentially other forms of actions (e.g., including limbic actions of emotional expression). The methodological innovations, such as the real-time cancellation of fMRI noise to allow voice responses, could also be exploited to investigate novel paradigms in which participants can use a non-manual response during experiments. Finally, by demonstrating that separate actions can be reinforced separately in the brain, this work contributes theoretical insights to the field of RL research, which involves not only neuroscience, but also cognitive psychology, computer science, and economics.

6.0 BIBLIOGRAPHY

Albin, R. L., Young, A. B., and Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in neurosciences*, 12(10):366–75.

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience*, 9:357–81.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, 111(4):1036–60.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., and O'Reilly, R. C. (2007). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nature neuroscience*, 10(1):126–31.

Attwell, D. and Iadecola, C. (2002). The neural basis of functional brain imaging signals. *Trends in neurosciences*, 25(12):621–5.

Baird, L. C. (1999). Reinforcement learning through gradient descent. Technical Report CMU-CS-99-132 (PhD Thesis), Carnegie Mellon University.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *The Behavioral and brain sciences*, 20(4):723–42; discussion 743–67.

Balleine, B. W., Delgado, M. R., and Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *Neuroscience*, 27(31):8161–5.

Bamford, N. S., Zhang, H., Schmitz, Y., Wu, N.-P., Cepeda, C., Levine, M. S., Schmauss, C., Zakharenko, S. S., Zablow, L., and Sulzer, D. (2004). Heterosynaptic dopamine neurotransmission selects sets of corticostriatal terminals. *Neuron*, 42(4):653–63.

Bar-Gad, I., Havazelet-Heimer, G., Goldberg, J. A., Ruppin, E., and Bergman, H. (2000). Reinforcement-driven dimensionality reduction–a model for information processing in the basal ganglia. *Journal of basic and clinical physiology and pharmacology*, 11(4):305–20.

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in neurobiology*, 71(6):439–73.

Blair, K., Marsh, A. A., Morton, J., Vythilingam, M., Jones, M., Mondillo, K., Pine, D. C., Drevets, W. C., and Blair, J. R. (2006). Choosing the lesser of two evils, the better of two goods: specifying the roles of ventromedial prefrontal cortex and dorsal anterior cingulate in object choice. *Neuroscience*, 26(44):11379–86.

Bolam, J. P., Hanley, J. J., Booth, P. A., and Bevan, M. D. (2000). Synaptic organisation of the basal ganglia. *Journal of anatomy*, 196 (Pt 4):527–42.

Botvinick, M. M., Niv, Y., and Barto, A. C. (2008). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*.

Brown, M. R. G., Goltz, H. C., Vilis, T., Ford, K. A., and Everling, S. (2006). Inhibition and generation of saccades: rapid event-related fmri of prosaccades, antisaccades, and nogo trials. *NeuroImage*, 33(2):644–59.

Brown, M. R. G., Vilis, T., and Everling, S. (2007). Frontoparietal activation with preparation for antisaccades. *Journal of neurophysiology*, 98(3):1751–62.

Bunzeck, N. and Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/vta. *Neuron*, 51(3):369–79.

Calzavara, R., Mailly, P., and Haber, S. N. (2007). Relationship between the corticostriatal terminals from areas 9 and 46, and those from area 8a, dorsal and rostral premotor cortex and area

24c: an anatomical substrate for cognition to action. *The European journal of neuroscience*, 26(7):2005–24.

Carr, D. B. and Sesack, S. R. (2000). Projections from the rat prefrontal cortex to the ventral tegmental area: target specificity in the synaptic associations with mesoaccumbens and mesocortical neurons. *Neuroscience*, 20(10):3864–73.

Cavada, C. and Goldman-Rakic, P. S. (1991). Topographic segregation of corticostriatal projections from posterior parietal subdivisions in the macaque monkey. *Neuroscience*, 42(3):683–96.

Chklovskii, D. B. (2004). Exact solution for the optimal neuronal layout problem. *Neural computation*, 16(10):2067–78.

Clatworthy, P. L., Lewis, S. J. G., Brichard, L., Hong, Y. T., Izquierdo, D., Clark, L., Cools, R., Aigbirhio, F. I., Baron, J.-C., Fryer, T. D., and Robbins, T. W. (2009). Dopamine release in dissociable striatal subregions predicts the different effects of oral methylphenidate on reversal learning and spatial working memory. *Neuroscience*, 29(15):4690–6.

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., and D'Esposito, M. (2009). Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration. *Neuroscience*, 29(5):1538–43.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., and Petersen, S. E. (1991). Selective and divided attention during visual discriminations of shape, color, and speed: functional anatomy by positron emission tomography. *Neuroscience*, 11(8):2383–402.

Corrado, G. and Doya, K. (2007). Understanding neural coding through the model-based analysis of decision making. *Neuroscience*, 27(31):8178–80.

Crutcher, M. D. and DeLong, M. R. (1984). Single cell studies of the primate putamen. ii. relations to direction of movement and pattern of muscular activity. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 53(2):244–58.

Daw, N. D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and

dopamine. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6):603–16.

Delgado, M. R., Locke, H. M., Stenger, V. A., and Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. *Cognitive, affective & behavioral neuroscience*, 3(1):27–38.

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., and Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of neurophysiology*, 84(6):3072–7.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks : the official journal of the International Neural Network Society*, 12(7-8):961–974.

Doya, K. (2002). Metalearning and neuromodulation. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6):495–506.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–91.

Feekes, J. A. and Cassell, M. D. (2006). The vascular supply of the functional compartments of the human striatum. *Brain : a journal of neurology*, 129(Pt 8):2189–201.

Földiák, P. (1990). Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 64(2):165–70.

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *Journal of neurophysiology*, 63(4):814–31.

Gerardin, E., Lehéricy, S., Pochon, J.-B., Tézenas du Montcel, S., Mangin, J.-F., Poupon, F., Agid, Y., Le Bihan, D., and Marsault, C. (2003). Foot, hand, face and eye representation in the human striatum. *Cerebral cortex (New York, N.Y. : 1991)*, 13(2):162–9.

Gillies, A. and Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement disorders : official journal of the Movement Disorder Society*, 15(5):762–70.

Good, M. and Honey, R. C. (1997). Dissociable effects of selective lesions to hippocampal subsystems on exploratory behavior, contextual learning, and spatial learning. *Behavioral neuroscience*, 111(3):487–93.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. i. a new functional anatomy. *Biological cybernetics*, 84(6):401–10.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. ii. analysis and simulation of behaviour. *Biological cybernetics*, 84(6):411–23.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Neuroscience*, 28(22):5623–30.

Harris, A. and Aguirre, G. K. (2008). The representation of parts and wholes in face-selective cortex. *Journal of cognitive neuroscience*, 20(5):863–78.

Haruno, M. and Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of neurophysiology*, 95(2):948–59.

Hazy, T. E., Frank, M. J., and O'Reilly, R. C. (2006). Banishing the homunculus: making working memory work. *Neuroscience*, 139(1):105–18.

Hazy, T. E., Frank, M. J., and O'reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1485):1601–13.

Ho, T. and Rouat, J. (1998). Novelty detection based on relaxation time of a network ofintegrate-and-fire neurons. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 2.

Hoffman, J. E. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & psychophysics*, 57(6):787–95.

Holland, P. C. (1984). Unblocking in pavlovian appetitive conditioning. *Journal of experimental psychology. Animal behavior processes*, 10(4):476–97.

Holroyd, C. B. and Coles, M. G. H. (2008). Dorsal anterior cingulate cortex integrates reinforcement history to guide voluntary behavior. *Cortex*, 44(5):548–59.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96(4):651–6.

Hoshi, E., Tremblay, L., Féger, J., Carras, P. L., and Strick, P. L. (2005). The cerebellum communicates with the basal ganglia. *Nature neuroscience*, 8(11):1491–3.

Houk, J., Adams, J., and Barto, A. (1995). *A model of how the basal ganglia generate and use neural signals that predict reinforcement*, pages 249–270. The MIT press.

Huettel, S., Song, A., and McCarthy, G. (2004). *Functional magnetic resonance imaging*.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral cortex (New York, N.Y. : 1991)*, 17(10):2443–52.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6):535–47.

Johnson, A. and Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural networks : the official journal of the International Neural Network Society*, 18(9):1163–71.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6):549–59.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., and Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *NeuroImage*, 46(3):786–802.

Knutson, B. and Cooper, J. C. (2006). The lure of the unknown. *Neuron*, 51(3):280–2.

Knutson, B. and Gibbs, S. E. B. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology*, 191(3):813–22.

Kowler, E., Anderson, E., Dosher, B., and Blaser, E. (1995). The role of attention in the programming of saccades. *Vision research*, 35(13):1897–916.

Krimer, L. S., Muly, 3rd, E. C., Williams, G. V., and Goldman-Rakic, P. S. (1998). Dopaminergic regulation of cerebral cortical microcirculation. *Nature neuroscience*, 1(4):286–9.

Landau, S. M., Lal, R., O'Neil, J. P., Baker, S., and Jagust, W. J. (2009). Striatal dopamine and working memory. *Cerebral cortex (New York, N.Y. : 1991)*, 19(2):445–54.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., and Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS biology*, 7(8):e1000173.

Laurent, P. A. (2008). The emergence of saliency and novelty responses from reinforcement learning principles. *Neural networks : the official journal of the International Neural Network Society*, 21(10):1493–9.

Laurent, P. A. and Reichle, E. D. (submitted). Serial or parallel? using reinforcement learning to examine attention allocation during reading.

Leblois, A., Boraud, T., Meissner, W., Bergman, H., and Hansel, D. (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *Neuroscience*, 26(13):3567–83.

Lehéricy, S., Ducros, M., Krainik, A., Francois, C., Van de Moortele, P.-F., Ugurbil, K., and Kim, D.-S. (2004a). 3-d diffusion tensor axonal tracking shows distinct sma and pre-sma projections to the human striatum. *Cerebral cortex (New York, N.Y. : 1991)*, 14(12):1302–9.

Lehéricy, S., Ducros, M., Van de Moortele, P.-F., Francois, C., Thivard, L., Poupon, C., Swindale, N., Ugurbil, K., and Kim, D.-S. (2004b). Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. *Annals of neurology*, 55(4):522–9.

Leichnetz, G. R. (2001). Connections of the medial posterior parietal cortex (area 7m) in the monkey. *The Anatomical record*, 263(2):215–36.

Lévesque, M. and Parent, A. (2005). The striatofugal fiber system in primates: a reevaluation of its organization based on single-axon tracing studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11888–93.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–7.

McClure, S. M., Berns, G. S., and Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339–46.

Middleton, F. A. and Strick, P. L. (1996). The temporal lobe is a target of output from the basal ganglia. *Proceedings of the National Academy of Sciences of the United States of America*, 93(16):8683–7.

Middleton, F. A. and Strick, P. L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain research. Brain research reviews*, 31(2-3):236–50.

Minsky, M. (1986). The Society of Mind Simon and Schuster.

Müller, J. R., Philiastides, M. G., and Newsome, W. T. (2005). Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):524–9.

Murthy, A., Thompson, K. G., and Schall, J. D. (2001). Dynamic dissociation of visual selection from saccade programming in frontal eye field. *Journal of neurophysiology*, 86(5):2634–7.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, N.Y.)*, 304(5669):452–4.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–37.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–73.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science (New York, N.Y.)*, 314(5796):91–4.

O'Reilly, R. C. and Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2):283–328.

Pagnoni, G., Zink, C. F., Montague, P. R., and Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature neuroscience*, 5(2):97–8.

Pasupathy, A. and Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433(7028):873–6.

Pennartz, C. M. A., Lee, E., Verheul, J., Lipa, P., Barnes, C. A., and McNaughton, B. L. (2004). The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *Neuroscience*, 24(29):6446–56.

Perry, R. J. and Zeki, S. (2000). The neurology of saccades and covert shifts in spatial attention: an event-related fmri study. *Brain : a journal of neurology*, 123 (Pt 11):2273–88.

Posner, M. I. (1980). Orienting of attention. *The Quarterly journal of experimental psychology*, 32(1):3–25.

Posner, M. I., Snyder, C. R., and Davidson, B. J. (1980). Attention and the detection of signals. *Journal of experimental psychology*, 109(2):160–74.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., and Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and aging*, 21(3):448–65.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature reviews. Neuroscience*, 7(12):967–75.

Redgrave, P., Prescott, T. J., and Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in neurosciences*, 22(4):146–51.

Reichle, E. D. and Laurent, P. A. (2006). Using reinforcement learning to understand the emergence of "intelligent" eye-movement behavior during reading. *Psychological review*, 113(2):390–408.

Rescorla, R. and Wagner, A. (1972). *A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*, pages 64–99. In: A.H. Black and W.F. Prokasy, Editors, Classical Conditioning. II: Current Research and Theory, Appleton-Century-Crofts, New York.

Reynolds, J. N. J. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6):507–21.

Rizzolatti, G., Riggio, L., Dascola, I., and Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A):31–40.

Robbins, T. and Everitt, B. (1992). Functions of dopamine in the dorsal and ventral striatum. In *Seminars in Neuroscience*, volume 4, pages 119–127. Elsevier.

Rolls, E. (2000). The orbitofrontal cortex and reward. *Cerebral cortex*, 10(3):284.

Salamone, J. D. (1994). The involvement of nucleus accumbens dopamine in appetitive and aversive motivation. *Behavioural brain research*, 61(2):117–33.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306):1593–9.

Shipp, S. (2004). The brain circuitry of attention. *Trends in cognitive sciences*, 8(5):223–30.

Stern, C. and Passingham, R. (1996). The nucleus accumbens in monkeys (Macaca fascicularis): II. Emotion and motivation. *Behavioural Brain Research*, 75(1-2):179–193.

Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 121(3):350–4.

Sutton, R. S. and Barto, A. G. (1998). Reinforcement learning: an introduction. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 9(5):1054.

Sweeney, J. A., Mintun, M. A., Kwee, S., Wiseman, M. B., Brown, D. L., Rosenberg, D. R., and Carl, J. R. (1996). Positron emission tomography study of voluntary saccadic eye movements and spatial working memory. *Journal of neurophysiology*, 75(1):454–68.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature neuroscience*, 7(8):887–93.

Tepper, J. M. and Bolam, J. P. (2004). Functional diversity and specificity of neostriatal interneurons. *Current opinion in neurobiology*, 14(6):685–92.

Tepper, J. M., Koós, T., and Wilson, C. J. (2004). Gabaergic microcircuits in the neostriatum. *Trends in neurosciences*, 27(11):662–9.

Tesauro, G. (1994). Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219.

Thinus-Blanc, C., Save, E., Poucet, B., and Buhot, M. C. (1991). The effects of reversible inactivations of the hippocampus on exploratory activity and spatial memory. *Hippocampus*, 1(4):365–71.

Thompson, K. G., Biscoe, K. L., and Sato, T. R. (2005). Neuronal basis of covert spatial attention in the frontal eye field. *Neuroscience*, 25(41):9479–87.

Tricomi, E., Balleine, B. W., and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *The European journal of neuroscience*, 29(11):2225–32.

Volkow, N. D., Wang, G.-J., Fowler, J. S., Logan, J., Jayne, M., Franceschi, D., Wong, C., Gatley, S. J., Gifford, A. N., Ding, Y.-S., and Pappas, N. (2002). "nonhedonic" food motivation in humans involves dopamine in the dorsal striatum and methylphenidate amplifies this effect. *Synapse (New York, N.Y.)*, 44(3):175–80.

Volz, K. G., Schubotz, R. I., and von Cramon, D. Y. (2005). Variants of uncertainty in decision-making and their neural correlates. *Brain research bulletin*, 67(5):403–12.

Vonsattel, J. P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D., and Richardson, Jr, E. P. (1985). Neuropathological classification of huntington's disease. *Journal of neuropathology and experimental neurology*, 44(6):559–77.

Webster, M. J., Bachevalier, J., and Ungerleider, L. G. (1993). Subcortical connections of inferior temporal areas te and teo in macaque monkeys. *The Journal of comparative neurology*, 335(1):73–91.

Wickens, J. (1993). *A theory of the striatum*. Elsevier Science Inc. New York, NY, USA.

Wickens, J. R., Budd, C. S., Hyland, B. I., and Arbuthnott, G. W. (2007). Striatal contributions to reward and decision making: making sense of regional variations in a reiterated processing matrix. *Annals of the New York Academy of Sciences*, 1104:192–212.

Wittmann, B. C., Bunzeck, N., Dolan, R. J., and Düzel, E. (2007). Anticipation of novelty recruits reward system and hippocampus while promoting recollection. *NeuroImage*, 38(1):194–202.

Yeterian, E. H. and Pandya, D. N. (1995). Corticostriatal connections of extrastriate visual areas in rhesus monkeys. *The Journal of comparative neurology*, 352(3):436–57.

Yeung, N., Botvinick, M. M., and Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4):931–59.

Zink, C. F., Pagnoni, G., Chappelow, J., Martin-Skurski, M., and Berns, G. S. (2006). Human striatal activation reflects degree of stimulus saliency. *NeuroImage*, 29(3):977–83.

Zink, C. F., Pagnoni, G., Martin, M. E., Dhamala, M., and Berns, G. S. (2003). Human striatal response to salient nonrewarding stimuli. *Neuroscience*, 23(22):8092–7.

Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., Chappelow, J. C., and Berns, G. S. (2004). Human striatal responses to monetary reward depend on saliency. *Neuron*, 42(3):509–17.

APPENDIX A

FMRI ANALYSIS

This appendix chapter details the fMRI analysis stream that was used to obtain the results discussed in this dissertation. The principal tool in the fMRI analysis was the free, open source NIMH-sponsored AFNI software package[1]. The other tool that was used in this work was University of Pennsylvania's ANTS Version 1.0 package, which was used to automatically transform the anatomy of individual participant brains to a common reference brain (i.e., TT_N27) in Talairach space[2].

Prior to statistical analysis, the fMRI data requires reconstruction and various normalization procedures known an *pre-processing*. The pre-processing steps can be divided into participant-specific and participant-general parts. Participant-specific steps include parameters that varied on a particular scan. For example, some participants may have moved between the acquisition of the structural scan and the acquisition of the functional scans, or those scans may have been obtained at differing obliquities: in those cases, some nudging and/or rotation was required prior to structural-to-functional alignment. In contrast, in several participants, there was significant movement between the acquisition of the anatomical and functional scans, and so additional nudging and/or rotation steps were required.

After the participant-specific steps, the bulk of the preprocessing took place using participant-general steps. This included aligning each participant's functional scans to their anatomical scans (correcting for movement from one functional scan to the next); the normalization of the brain ac-

[1] AFNI is available at http://afni.nimh.nih.gov/
[2] ANTS is available at http://www.picsl.upenn.edu/ANTS/

tivation intensity values throughout each run, etc. They also included transforming the anatomical image to the group anatomical image in Talairach space, which was taken in this study to be Colin's (i.e., TT_N27) brain.

Note that in the design of this fMRI analysis stream, it was preferred to minimize the amount of processing applied to the functional images. The reasons for this are two fold: (1) applying transformations or computation to the entire set of functional images would be computationally intensive compared to simply processing the structural image, and (2) most transformations degrade the data in some way, and so the goal was to minimize this degradation.

A.1 FMRI PRE-PROCESSING

A.1.1 PARTICIPANT-SPECIFIC STEPS

Depending on logistics of each participant's scans, the following two steps differed slightly from participant to participant. Copies of the scan sheets provided by the scanning center for each participant are attached for future reference.

The first step in the preprocessing was reconstructing the images from the scanner to convert them into AFNI format. The numbers preceding the /*.dcm referred to the run number (i.e., the nth time the scanner was started for a continuous run):

```
# Anatomical image
to3d -orient ASR -assume_dicom_mosaic -anat -prefix mprage 3/*.dcm

# Resting state correlation data
to3d -assume_dicom_mosaic -time:zt 38 256 2s alt+z -prefix resting 5/*.dcm

# Experimental data (alternating CONTROL and LEARN runs)
to3d -assume_dicom_mosaic -time:zt 38 216 2s alt+z -prefix control1 6/*.dcm
to3d -assume_dicom_mosaic -time:zt 38 432 2s alt+z -prefix learn1 7/*.dcm
to3d -assume_dicom_mosaic  -time:zt 38 216 2s alt+z -prefix control2 9/*.dcm
to3d -assume_dicom_mosaic -time:zt 38 432 2s alt+z -prefix learn2 10/*.dcm
to3d -assume_dicom_mosaic -time:zt 38 216 2s alt+z -prefix control3 11/*.dcm
to3d -assume_dicom_mosaic -time:zt 38 432 2s alt+z -prefix learn3 12/*.dcm
```

AFNI includes an automated coregistration tool called **align_epi_anat.py** that was used to align the structural scan to the functional scans. However, If the structural and functions were significantly out of spatial alignment, first a rough hand alignment of the functional and anatomical datasets was done using the **nudge** and **rotate** tools as demonstrated below. Note that the number of

type of nudges differed from participant to participant until the images were approximately aligned. Also note that **align_epi_anat.py**, which performs the automated alignment of the structural image to the functionals, has options such as "-big_move" and "-huge_move" which may overcome the need for the coarse manual alignment.

```
# Manually nudging and rotating for a rough initial alignment.
3dSkullStrip -input mprage+orig -prefix mprage_3dskullstrip
3dAnatNudge -x 5 -y 20 -z 5 -anat mprage_3dskullstrip+orig -epi resting+orig -prefix mprage_nudge
3dAnatNudge -x 5 -y 20 -z 5 -anat mprage_nudge+orig -epi resting+orig -prefix mprage_nudge2
3dAnatNudge -x 5 -y 20 -z 5 -anat mprage_nudge2+orig -epi resting+orig -prefix mprage_nudge3

3drotate -prefix mprage_nudge3_rot13R -rotate 13R 0 0 mprage_nudge3+orig

3dcopy mprage_nudge3_rot13R+orig anat_mprage
```

A.1.2 PARTICIPANT-GENERAL STEPS

Auto-Talairaching. AFNI provides an automatic Talairaching algorithm which brings participants' brains into a common space. Although the results in this dissertation are based on the use of a separate transformation (ANTS; see below), the automated Talairaching is mentioned here because is generally sufficiently good for studies of neocortical activations. The syntax for a pre-stripped brain (e.g., *anat_mprage_al+orig* from above) is as follows:

```
# Warp a structural named anat_mprage_al+orig to Colin (TT_N27) yielding anat_mprage_al+orig__ref_TT_N27
@auto_tlrc -no_ss -base TT_N27+tlrc -input anat_mprage_al+orig -suffix __ref_TT_N27
```

Time-slicing correction. As in most scanning protocols, the functional data for each participant was captured in an interlaced fashion: that is, each functional image contained a mixture of acquired voxels from two distinct time intervals. For example, half of the slices in the functional image captured at t=10s were captured at t = 6s-8s, and the remainder were captured at t=8s-10s. This interlacing of data is largely irrelevant for block-design experiments because the intensity values are summed or averaged across the entire block. However, the present experiment is an event-related design experiment. Therefore, to maximize our power to detect event-related effects at particular times, the images were de-interlaced. This is termed *slice-timing correcton*.

```
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix resting_tshift resting+orig
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix control1_tshift control1+orig
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix control2_tshift control2+orig
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix control3_tshift control3+orig
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix learn1_tshift learn1+orig
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix learn2_tshift learn2+orig
3dTshift -Fourier -tpattern alt+z -ignore 5 -prefix learn3_tshift learn3+orig
```

Motion correction. All of the functional images from the experiment were concatenated sequentially and then AFNI's motion correction algorithm was applied to the data to bring all images into register with the 5th image (i.e., image 4.). This operation also outputted an estimate of movement in three different dimensions.

```
# Concatenate all runs of the scan.
3dTcat -prefix allruns resting_tshift+orig control1_tshift+orig \\
    learn1_tshift+orig control2_tshift+orig  learn2_tshift+orig  \\
    learn3_tshift+orig control3_tshift+orig

# Perform motion correction, aligning all images (every 2 seconds) to the fifth image
3dvolreg -base allruns+orig'[4]' -Fourier -zpad 4 -dfile rollpitchyawSILRPA.txt \\
   -prefix allruns_volreg allruns+orig
```

Align structural to functionals. Rather than aligning each participant's functional images to their anatomical image, it was preferred to align the anatomical image to the functional images. This was less computationally intensive, and helped to minimize the number of transformations (and therefore distortions) that occurred to the functional data in the course of the analysis.

```
# Align the MPRAGE to the functionals (for anatomical localization,
# and eventual Talairaching of statistical results.)

align_epi_anat.py -volreg off -deoblique off -anat_has_skull no \\
    -anat anat_mprage+orig -epi allruns_volreg+orig -epi_base 5
```

Intensity Normalization. To adjust for changes in scanner signal intensity from one run to the next, the images within each run were normalized to a common intensity value.

```
3dAutomask -clfrac .25 -eclip -prefix allruns_volreg_automasked allruns_volreg+orig

3dTstat -mean -prefix resting_mean allruns_volreg+orig'[0..255]'
3dTstat -mean -prefix control1_mean allruns_volreg+orig'[256..471]'
3dTstat -mean -prefix learn1_mean allruns_volreg+orig'[472..903]'
3dTstat -mean -prefix control2_mean allruns_volreg+orig'[904..1119]'
3dTstat -mean -prefix learn2_mean allruns_volreg+orig'[1120..1551]'
3dTstat -mean -prefix control3_mean allruns_volreg+orig'[1552..1767]'
3dTstat -mean -prefix learn3_mean allruns_volreg+orig'[1768..2199]'

3dcalc -fscale -a allruns_volreg+orig'[0..255]' -b resting_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix resting_normalized
3dcalc -fscale -a allruns_volreg+orig'[256..471]' -b control1_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix control1_normalized
3dcalc -fscale -a allruns_volreg+orig'[472..903]' -b learn1_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix learn1_normalized
3dcalc -fscale -a allruns_volreg+orig'[904..1119]' -b control2_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix control2_normalized
3dcalc -fscale -a allruns_volreg+orig'[1120..1551]' -b learn2_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix learn2_normalized
3dcalc -fscale -a allruns_volreg+orig'[1552..1767]' -b control3_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix control3_normalized
3dcalc -fscale -a allruns_volreg+orig'[1768..2199]' -b learn3_mean+orig \\
     -c allruns_volreg_automasked+orig -exp 'c*(a/b*100)' -prefix learn3_normalized

```

```
# Re-assemble all the runs into a single file for analyses.
3dTcat -prefix allruns_normalized resting_normalized+orig control1_normalized+orig \\
     learn1_normalized+orig control2_normalized+orig  learn2_normalized+orig  \\
     control3_normalized+orig  learn3_normalized+orig
```

Separating Experiments. The preceding steps complete the pre-processing. Finally, because resting state and the experiment will be treated separately, they are divided into two separate files.

```
3dTcat -prefix restingstate allruns_normalized+orig'[0..255]'
3dTcat -prefix mmrl allruns_normalized+orig'[256..2199]'

3dTcat -prefix restingstate_smoothed allruns_normalized_smoothed+orig'[0..255]'
3dTcat -prefix mmrl_smoothed allruns_normalized_smoothed+orig'[256..2199]'
```

A.2 SPATIAL TRANSFORMATIONS

Performing spatial functional neuroimaging analyses at the group level requires some type of transformation to bring the different brains of the participants into a common space. Often, this common space is in a standardized coordinate frame (i.e., Talairach space) so that the coordinates of regions of interest can be reported and compared.

There are a number of transformation methods that can be performed, each with their advantages and disadvantages. For example, the bulk of fMRI studies use a method like that provided by the AFNI package, which is an affine transformation. An **affine transformation** is a series of linear transformations which include a combination of translating (i.e., shifting in the x-,y-, or z-plane), rotating, scaling and shearing (i.e., titling in the x-, y-, or z-plane). In studies focusing on neocortical BOLD activation, which relatively large smoothing kernels (e.g., 8mm or more), the affine transformation may be adequate. However, when an affine transformation is not adequate, there are a number of alternative methods that can be used. One such method is **nonlinear warping**. When examining the regions that are the focus of this study, such as the striatum, the inadequacy of an affine transform becomes especially apparent. The shape of the striatum, especially in its most dorsal portions, varies considerably from individual to individual, a fact possibly related to its juxtaposition to the ventricles. According to a recent comparative review of 14 non-linear warping algorithms, ANTS (Advanced Normalization Tools, University of Pennsylvania) was superior

especially with respect to basal ganglia normalization (Klein et al., 2009). To provide a qualitative assessment of performance on the present data set, Figure A1 (page 162) compares the alignment of AFNI's automated affine transformation tool (i.e., auto_tlrc) to the alignment of ANTS' Syn warping algorithm. In a number of respects, the results using ANTS were clearly superior.

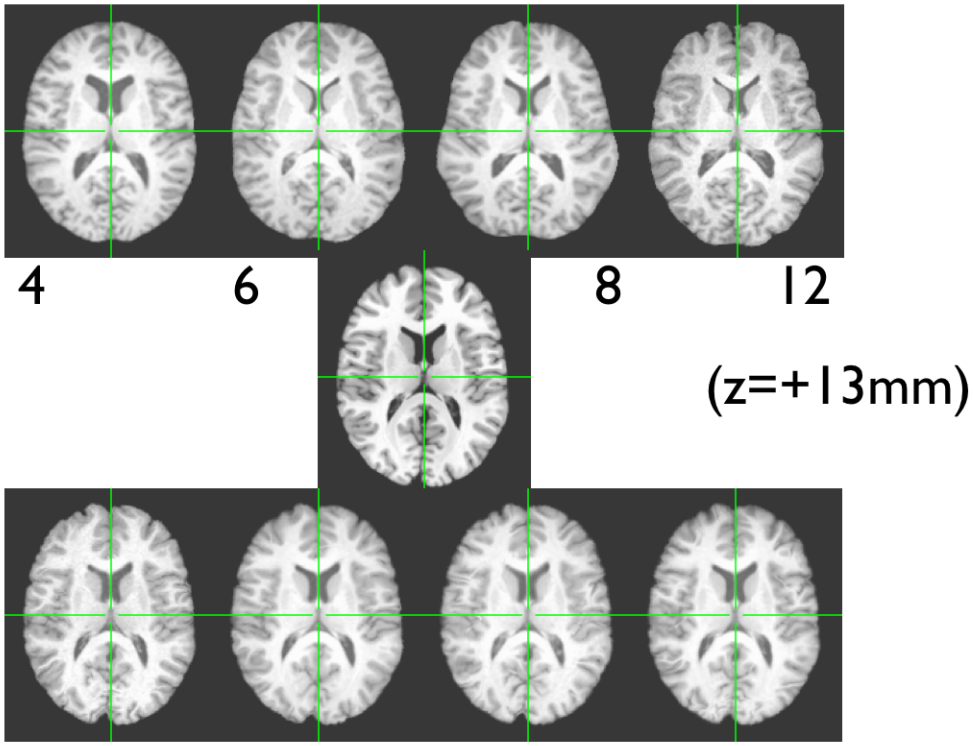

Figure A1: **Comparison of results from AFNI's affine transformation (top) to the ANTS non-linear warping algorithm, default settings (bottom).** The top row shows the same axial slice from four experiment participants. The image in the center is the slice at the same level from the target brain, TT-N27 (Colin). The bottom row is the result of the automated warping algorithm provided by ANTS. The improvement in the coregistration is very apparent in this particular slice, especially for the caudate nucleus, as well as the grey matter in the frontal and occipital cortices.

Using the anatomical images that have been nudged, rotated, and aligned (volume registered in AFNI) to the functional data (i.e., *anat_mprage_al+orig*), we apply the following script to each participant. For example, for participant f004,

```
cd WARP_ANTS
```

```
  ln -s ../f004/anat_mprage_al+orig.BRIK ./f004_needsResampleToRAI+orig.BRIK
  ln -s ../f004/anat_mprage_al+orig.HEAD ./f004_needsResampleToRAI+orig.HEAD
  3dresample -orient RAI -prefix f004 -inset f004_needsResampleToRAI+orig

  3dAFNItoNIFTI f004+orig
  ants.sh 3 TT_N27.nii f004.nii ants_f004 > ants_f004.log
cd ..
```

This command uses the default "ants.sh" script, which uses as its parameters: the transformation algorithm itself is SyN[0.25], the MaxIterations is 30x90x20, the number of MultiResolution levels is 3, Regularization is Gauss[2,0] and the metric is "PR[".

The result of this transformation is a number of files that contain the transformed anatomical image, as well as matrix representations of the affine and nonlinear transformations that were performed.

```
ants_f004Affine.txt
ants_f004InverseWarpxvec.nii
ants_f004InverseWarpyvec.nii
ants_f004InverseWarpzvec.nii
ants_f004Warpxvec.nii
ants_f004Warpyvec.nii
ants_f004Warpzvec.nii
ants_f004deformed.nii
```

The results of this warping are: (1) an anatomical image of the participant that has been warped to the target brain (i.e., ants_f004deformed.nii) and (2) A series of transformation matrices that can be applied to functional data to bring them into the group space (i.e., the remainder of the files in the above listing).

A.2.1 FORWARD TRANSFORMATION INTO GROUP SPACE

Using ANTS. After deconvolution analyses were performed at the single subject level (using the single subject regressors produced by the reinforcement learning model), group-level analyses (i.e., a 3dttest) requires that each participant's statistical map be translated to the group space using the resulting matrices from the nonlinear anatomical warping described in the preceding section.

Applying this transformation to the output from 3dDeconvolve (i.e., DeconvOutput.nii) for a particular participant (e.g., f004) can be performed by using this command:

```
3dAFNItoNIFTI Deconv+orig
WarpImageMultiTransform 3 Deconv.nii DeconvWarped_to_TTN27.nii -R TT_N27.nii --ANTS-prefix ants_f004
```

Using AFNI.

A similar transformation can be performed using AFNI's simpler affine transformation given that a participant anatomy has been Talairached (e.g., using @auto_tlrc).

```
@auto_tlrc -apar anat_mprage_al__ref_TT_N27+tlrc -suffix _to_TTN27 -input Deconv+orig
```

A.2.2 INVERSE TRANSFORMATION TO INDIVIDUAL SPACES

A number of fMRI hypotheses require performing tests on original data, rather than statistical maps arising from deconvolution or GLM analyses. For example, a researcher may be interested in extracting time-course information from particular voxels in each subject's brain specified by a group-level region of interest or mask. Another example is to perform functional connectivity analysis, which requires examining temporal correlations in the original signal. Performing these kinds of analyses requires either transforming all of the raw functional data to the group space (a computationally expensive operation), or, the solution preferred here, transforming the group-level maps or regions of interest back into the original subject spaces.

Both ANTS and AFNI provide transformations that are invertible. This section shows example syntax that demonstrates how to perform these inverse transformations. (Note that inverse transformations are not required to obtain a group level time-courses if impulse-response functions are saved for each voxel, as they are in deconvolution approaches. Those impulse-response functions can be forward transformed into the group space and then extracted or averaged over group-level masks.)

Using ANTS. To invert a map from the group space into the space of a single subject, one could perform the following command.

```
WarpImageMultiTransform 3 mask.nii mask_inverted_result.nii -R f004.nii \
    --ANTS-prefix-invert ants_f004 --ANTS-invert
```

Using AFNI. If mask was obtained on group data obtained by applying @auto_tlrc, that transformation can be inverted by using the 3dfractionize command, e.g.,

```
3dfractionize -template anat_mprage+orig -input groupmask+tlrc -warp anat_mprage_at_TT_N27+tlrc -clip 0.2
```

APPENDIX B

REAL-TIME FMRI SCANNER NOISE CANCELLATION

This chapter details the sound cancellation technique that was developed for this fMRI study so that oral reports of the participants could be recorded as they performed the task. Importantly, the oral reports needed to be heard and entered into the computer online, i.e., while the experiment was running, so that participants could obtain reward feedback.

This sound cancellation approach used here was selected because the algorithm is simple (i.e., a dozen lines of code), and it was not very computationally-intensive.

B.1 OVERVIEW

One of the requirements of the experimental task was that the experimenter could hear participants' oral reports of the probe letters they observed during the task. This was required because the letters had to be entered into the experiment computer so that it could provide real-time reward feedback on every trial. However, the sound from the fMRI scanner (i.e., "scanner noise") is loud and typically prevents experimenters from hearing participants speak while the scanner is running. Functional MRI scans (echo planar imaging) use a repetitive scanning procedure; in this study in particular, the repeated scanning occurred every two seconds (repetition time, TR = 2). The scanner noise generated on each TR is nearly identical to the sound generated on the previous TR. Using this fact, it is possible to cancel out scanner noise in real-time by recording the previous TR, phase-inverting

it, and then superimposing it onto the current sound. These operations effectively subtract out or "cancel" the scanner noise for the experimenter and allow the experimenter to hear.

B.2 METHOD

A simpler approach might have been to record the scanner noise from a single TR and use this as a template to cancel out all of the TRs during an experiment. However, this approach is not successful because there are differences in the sound signal received by the microphone due to moment-to-moment variations in the acoustic environment within the scanner. For example, slight body postural movements or hand movements that occur during the hand-movement condition of the experiment cause differences in how the sound waves are reflected from the participant and the scanner cavity into the microphone. For this reason, it is insufficient to simply capture a template of the sound from one TR and use it to cancel out the scanner sound from all subsequent TRs.

A more sophisticated procedure involves updating the template only when the previous TR has changes significantly. However this is more computationally intensive than the approach of simply consistently using he previous TR to cancel out the current TR, because it requires comparison operations between the template and each TR.

Thus, the simplest technological solution to the "scanner noise" problem is to continually (i.e., on every TR) update the template used to cancel out the sound in the subsequent TR in real-time as the experimenter monitors the participant's reports. The only undesired consequence of this approach is that a additional, but phase-inverted, version of the participant's response is repeated on the TR after the response was made. That is, each oral response made by the participant is repeated by the sound cancellation system on the subsequent TR.

Real-time processing of sound is easily achieved using the program called ChucK developed at Princeton (http://chuck.cs.princeton.edu/). An example script follows below. Sound from the microphone is fed into the computer's microphone input, and the experimenter wears headphones. Sound cancellation begins after the first TR has been read into the system as illustrated in Figure B1, page 171.

```
// ---------------------------------------------------------------
// invert_delay_sound.ck -- realtime scanner noise cancellation
//   for fMRI paradigms.
// Written by: Patryk Laurent, University of Pittsburgh, 2009.
// ---------------------------------------------------------------

2000.042::ms => dur delayLength;    // Based on duration of TR, fine tuned for computer.

// Set up input-output copy through a virtual y-plug, and a path for a delayed input.
adc.left => Gain half1 => Gain yPlug => dac;
adc.left => Gain half2 => Gain phaseInverter => DelayL delay => yPlug;
0.5      => half1.gain;
0.5      => half2.gain;
1.0      => yPlug.gain;

// Set up inverted, delayed copy.
-1          => phaseInverter.gain;
delayLength => delay.max => delay.delay;

while(true)
{
    100::ms => now;
}
```

We noted that the precise timing period of the fMRI scanner noise varied from day to day. Presumably this is because of the facts that (1) the noise is generated by the movement of physical parts, and (2) there are slight variations in the scanner environment from day-to-day (e.g., temperature and humidity). Thus, it was noted that a TR of 2 seconds was not cancelled out exactly by a delay loop of 2000ms. (In the code example above, 2000.042 ms was the optimal delay for canceling scanner noise on the majority of the scans.) Also, it is possible that the resolution or calibration of the sound-canceling computer's internal clock was an issue.

It is therefore ideal to test the sound-canceling software prior to data collection. Fortunately, the fMRI experiment included a period of *resting-state* data collection prior to runs in which oral responses needed to be collected. During this initial resting-state data collection period, were able to measure the effectiveness of the scanner noise cancelation and optimize the delay. Critically, participants did not speak during this period, and this allowed us to quantify the amount of scanner noise remaining after the cancelation operation for a variety of delay loop durations. We then chose the duration that minimized the remaining residual noise, reported numerically by the search code below:

```
// ---------------------------------------------------------------
// To enable searching for optimal delay loop length, insert
// this code at the beginning of the ChucK script.
// ---------------------------------------------------------------
spork ~ keepIncrementingDelayAndPrintingRMS();
0.001::ms    => dur searchStepSize;     // If you enable search, step size.

// .
// .
// .
//  main sound cancellation code goes here
// .
// .
// .

// ======================== SEARCH CODE START ====================
// (if you don't need to automatically search for the best delay,
//  you can ignore this code.)
// Call using: spork ~ keepIncrementingDelayAndPrintingRMS();
// (This line is below, right above the main program loop.)

// ---------------------------------------------------------------
// Here we set up the analyzer so we can measure how much sound
// is not being cancelled out.  If it's high, we should search
// for a better delay.
// ---------------------------------------------------------------
dac => FFT fft =^ RMS rms => blackhole;
1024 => fft.size;
Windowing.hann(1024) => fft.window;
// ---------------------------------------------------------------
// This function is started up by the "spork" command below, and
// increments the delay while measuring RMS (total sound energy).
// We want the lowest RMS possible.
// ---------------------------------------------------------------
fun void keepIncrementingDelayAndPrintingRMS()
{
        while (true)
        {
                delayLength + searchStepSize => delayLength;
                delayLength => delay.max => delay.delay;
                <<< "  " >>>;
                <<< delayLength / 1::ms >>>;
                2::second => now;

                0 => float sum;
                for (0 => int x; x < 4; x++)
                {
                        rms.upchuck() @=> UAnaBlob blob;
                        sum + blob.fval(0) => sum;
                        // <<< blob.fval(0) >>>;
                        fft.size()::samp => now;
                        500::ms => now;
                }
                <<< "average" >>>;
                <<< sum / 4.0 >>>;
        }
}


// ======================== SEARCH CODE END ====================
```

B.3 LIMITATIONS

One potential limitation of this technique in fMRI scanning studies is that it is likely limited to vowels only, for the reasons specified above. I found that using vowels did not create extreme amounts of head movement. Further testing is required to determine whether the use of consonants in syllables would cause too much movement or artifact. If more than 4 or 5 responses are required, experimenters may wish to consider using a larger set of vowel sequences (e.g., "a-y-o", "e-y-a", etc.)

An important limitation is that this technique cannot be used to cancel out scanner noise for an experiment participant *within* the scanner. This is because the microphone that feeds into the sound cancellation algorithm is located in a position that is different from the location of each of the participant's headphones, and it therefore is canceling out sound that is different from the sound arriving at the participant's left and right ears. In other words, this technique only works for the experimenter because the sound being cancelled out and the desired sound are both being detected by the same microphone. However, preliminary experimentation suggests that if the microphone is placed within the ear of the participant and appropriate delay is set, the sound can be effectively canceled.

B.4 FUTURE DIRECTIONS

A future application of this technique is to combine it with automatic transcription of the finite number of oral responses. This would help overcome the current limitation that the experimenter has to attentively listen to the speech reports during all portions of the experiment during which feedback has to be given back to the participant (that is, during the LEARN runs.) On every trial, the experimenter must enter the letter on the keyboard that corresponds to what the participant reports they saw. The recognition of vowels based on format analysis can be easily automated by a computer, and could be done in real time using a neural network. Thus a future application of this technique could allow the oral responses of participants to be interpreted directly by the experiment

display computer, and therefore reduce the need for the experimenter to attend to each trial.

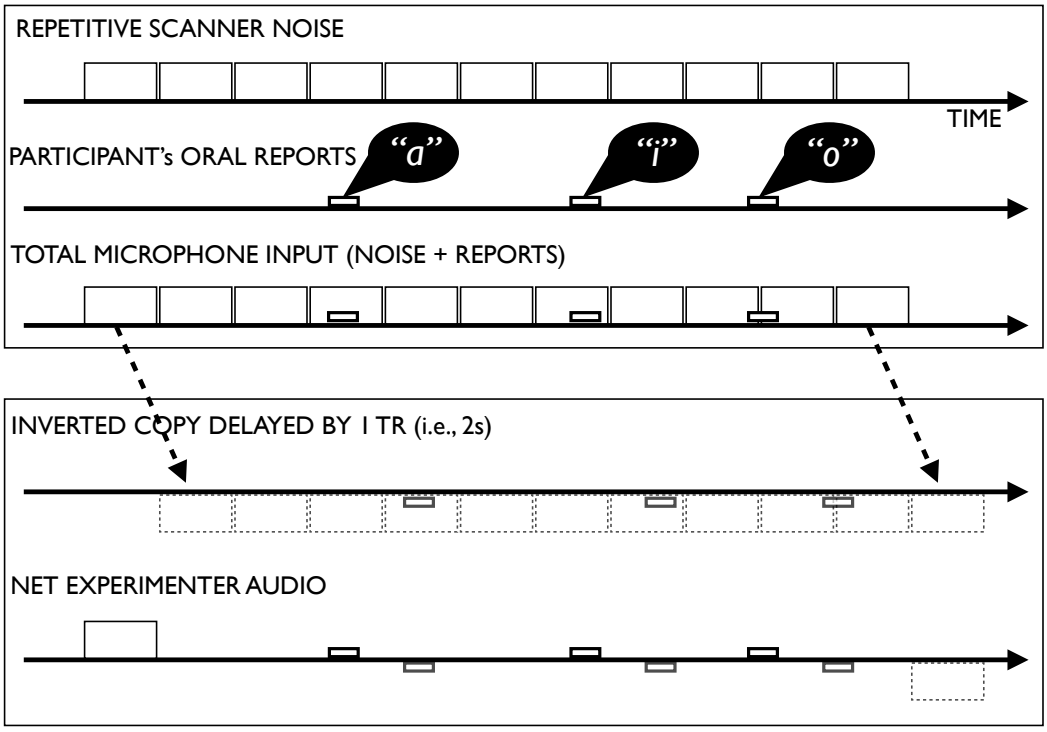

Figure B1: **Echo Cancellation Technique for Oral Reports of Participants during an fMRI scan**. During an EPI (echo-planar imaging) scan, the scanner noise repeats with a period of 2 seconds. Superimposed on this repetitive scanner noise are the much quieter oral responses of the experiment participants. The microphone thus detects the combination of the scanner noise and the oral responses. Echo cancellation involves digitally manipulating the real-time audio captured by the microphone. A delay loop is created whereby the previous 2 seconds (the period of the TR scanner) of sound are phase-inverted, and then added to the live audio being detected by the microphone. The result is that the scanner noise is cancelled out, except for at the beginning of the scan (i.e., the first TR is ineffectively canceled out by the silence that precedes the scan.) A consequence of this cancelation technique is that the participants' oral reports are heard twice: the first time is in real time when they vocalize their response, and the second time is the phase-inverted delayed copy of the vocalization. This is illustrated in the bottom-most panel, where a copy of the oral report occurs 2 seconds after this first report. Depending on the experiment design, this could be an advantage because it provides the experimenter with a second opportunity to confirm what the participant reported.

APPENDIX C

DECONVOLUTION ANALYSIS ILLUSTRATION

This appendix chapter contains a brief illustrative example of deconvolution analyses, which is an important technique used in this dissertation. Deconvolution analysis is a generalization of multiple regression analysis that is used to determine the extent to which patterns in the data (i.e., voxel activation over time) can be explained by a set of regressors arranged in a temporal sequence. Unlike standard multiple regression approaches in which a single linear regression coefficient is solved for each time series, deconvolution analysis creates multiple time-shifted linear regression coefficients, one for each time step of the expected response.

In such conventional analyses, a hemodynamic response function is selected *a priori*, and under the assumptions of a general linear model (i.e., linear additivity of bold responses), multiple regression is performed to determine the extent to which each event can explain changes in the observed BOLD response. This approach works quite well and has been used extensively in the neuroimaging literature.

Deconvolution analysis has a number of advantages over conventional multiple regression when analyzing BOLD fMRI data. Because a separate coefficient is computed for each time-step of the response, deconvolution analysis does not require an *a priori* specification of a hemodynamic response function shape for each event in the time series. This approach is particularly appealing because fewer assumptions go into the analysis. Obviating the need to specify the hemodynamic response function shape is especially relevant in the present study because striatal BOLD responses may vary depending on their cause and location. Deconvolution analysis also has the advantage

that the signal recovered for a particular event is only that proportion of the signal which could not be explained by all the other regressors. That is, all the other time series are given a chance to explain the variance in the signal before the reported time series.

The first step of this demonstration is to generate events of interest. Because the purpose of this demonstration is to illustrate the usefulness of this analysis for the experiment described in the dissertation, the events will resemble the events during the experiment. First there will be a sequence of 'control' movement trials without reinforcement, all with magnitude 1. Then, there will be a similar sequence of movement events, each of which is followed – with a fixed delay of 1 time-step – by a reward event of varying magnitude. Although there is jitter between trials, there is no jitter within a trial. Code for generating this kind of event sequence is provided here:

```
import java.util.*;
import java.io.*;

public class GenerateStims
{
        public static void main(String args[]) throws Exception
        {
                Random rnd = new Random(12345);
                PrintWriter m = new PrintWriter(new BufferedWriter(new FileWriter("movement.txt")));
                PrintWriter r = new PrintWriter(new BufferedWriter(new FileWriter("reward.txt")));

                int trialsBeforeRewards = new Integer(args[0]);
                int trialsWithRewards = new Integer(args[1]);
                System.out.println("trialsBeforeRewards = " + trialsBeforeRewards);
                System.out.println("trialsWithRewards = " + trialsWithRewards);
                double maxJitter = 3.0d;

                for (int i = 0; i < trialsBeforeRewards; i++)
                {
                        m.println(1); r.println(0);
                        m.println(0); r.println(0);
                        m.println(0); r.println(0);
                        int jitter = (int) (maxJitter * rnd.nextDouble());
                        for (int j = 0; j < jitter; j++) { m.println(0); r.println(0); }
                }
                for (int i = 0; i < trialsWithRewards; i++)
                {
                        double reward = 0.25 * (int) (5.0d * rnd.nextDouble());
                        m.println(1); r.println(0);
                        m.println(0); r.println(0);
                        m.println(0); r.println(reward);
                        int jitter = (int) (maxJitter * rnd.nextDouble());
                        for (int j = 0; j < jitter; j++) { m.println(0); r.println(0); }
                }
                m.close(); r.close();
        }
}
```

Running this code outputs event numbers into the 'movement' and 'reward' regressor files. Events from one such run are plotted in Panel (1) of Figure C1. In that figure, the 'movement' signals are

the full-height blue bars throughout the simulated experiment data. The 'reward' signals are the variable height red bars interspersed with the blue bars during the latter half of the simulated experiment data. The next step in this demonstration is to generate a hemodynamic response function that will be use to simulate the continuous BOLD response in a voxel arising from the events of interest. (Note that one could also generate two separate BOLD responses of different shape for each event type. Note that, in general, differences in shapes of the two responses would make them more easy to deconvolve.)

```
waver -WAV -delaytime 0 -risetime 2 -falltime 9  -undershoot 0.2 -restoretime 1 -peak 100 -dt 1 > waver.1D
```

The hemodynamic response function is plotted in panel 2 of Figure C1. Now, the movement and reward events can be convolved with this hemodynamic response function to yield a simulated BOLD response for an imaginary voxel:

```
3dConvolve -polort 3  \
 -base_file base.txt \
 -nfirst 0 \
 -nlast `cat movement.txt | wc -l | sub1` \
 -num_stimts 2 \
 -input1D \
 -stim_file 1 movement.txt \
 -stim_minlag 1 2 \
 -stim_maxlag 1 14 \
 -stim_file 2 reward.txt \
 -stim_minlag 2 2 \
 -stim_maxlag 2 14 \
 -iresp 1 waver.1D \
 -iresp 2 waver.1D \
 -output convolved
```

The simulated hemodynamic response containing the superposition of movement and reward activities is plotted in Panel 3 of Figure C1.

```
3dDeconvolve -polort 3 \
 -num_stimts 2 \
 -input1D convolved.1D \
 -stim_file 1 movement.txt \
 -stim_minlag 1 0 \
 -stim_maxlag 1 20 \
 -stim_label 1 'movement' \
 -stim_file 2 reward.txt \
 -stim_minlag 2 0 \
 -stim_maxlag 2 20 \
 -stim_label 2 'reward' \
 -iresp 1 movement.irf \
 -iresp 2 reward.irf \
 -gltsym 'SYM: +movement' -glt_label 1 'movglt' \
 -gltsym 'SYM: +reward' -glt_label 2 'rewglt' \
 -bucket deconvolved
```

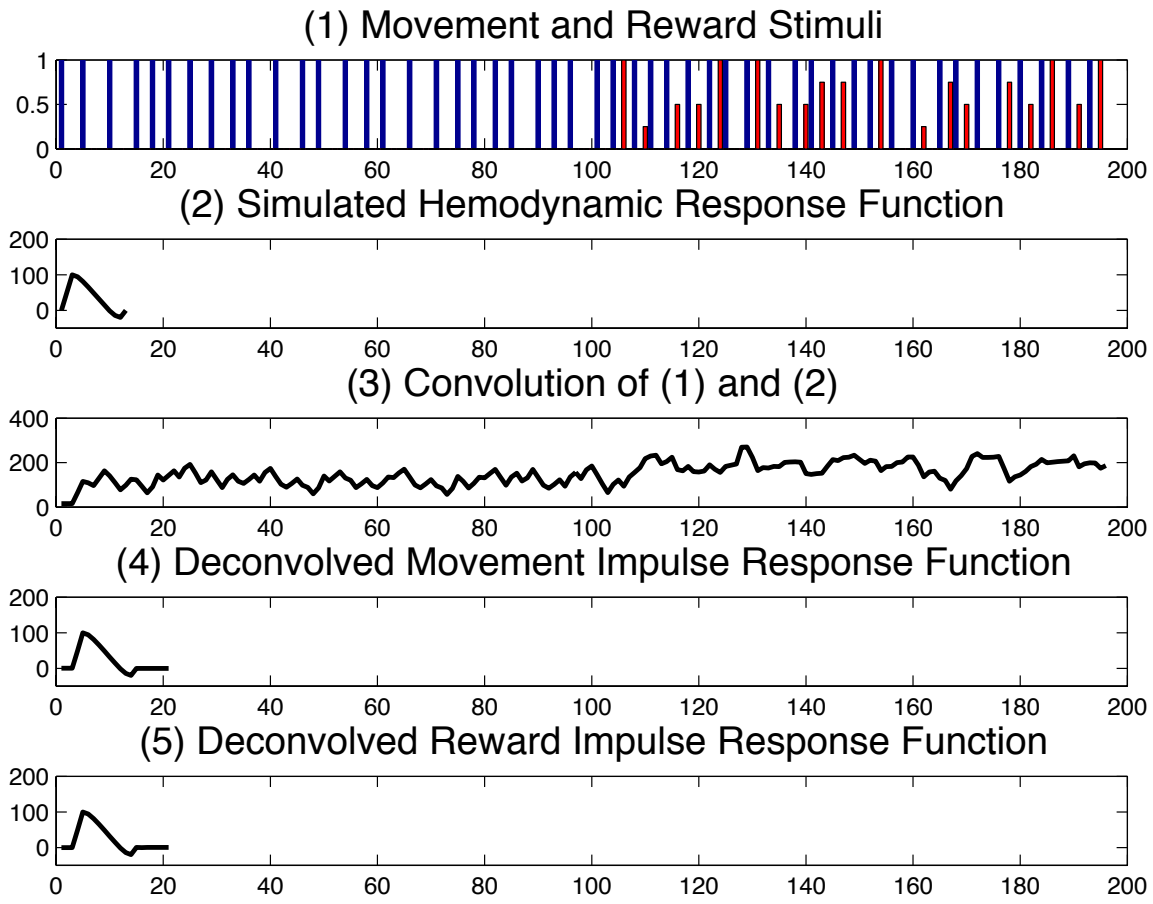

Figure C1: **Illustration of Deconvolution Analysis**. **(1)** Sequence of Events. The blue bars represent binary movement events. The red bars represent continuously-valued scalar reward. The rewards are provided at a fixed delay after movements. **(2)** A simulated hemodynamic response function that will be convolved with the events shown in (1) to generate a simulated BOLD time-course for an imaginary voxel. **(3)** The results of convolving the events with the hemodynamic response function. Note that the original shape of the hemodynamic response function is not discernable in this signal. **(4-5)** The recovered hemodynamic response function shapes for the movement and reward events, thanks to deconvolution analysis.

Despite its advantages, there are several potential disadvantages of using deconvolution analysis. One potential disadvantage is that achieving significant might require more statistical power

because it divides the data points among multiple coefficients. However, the fact that there is a large quantity of data provided for each time point makes this not a major concern. Further, the increase in power through a better fit of the true underlying data may offset the possible loss in power through the need to fit additional coefficients. Another potential disadvantage of using deconvolution is that the coefficients in a particular time window may be either positive and negative. This may cause a problem because group-level statistics are usually based on the sum of the coefficients within a time window surrounding the event of interest, and negative coefficients can cancel out positive coefficients. Thus, attention must be paid to the time window over which summation is made. Generally, summation should be limited to the time period over which the hemodynamic response is expected to occur. Alternatively, one way to overcome this difficulty could be to summate absolute values of the coefficients rather than raw values.

Deconvolution analysis therefore is a useful tool and important tool for the analysis of BOLD fMRI data. It is particularly attractive in that it makes few assumptions about the shape of the underlying hemodynamic response and thus is well-suited for studies in subcortical regions. However, its use requires care and exploration because of the issues highlighted above.