1-1-2014

Human Reinforcement Learning: Insights from intracranial recordings and stimulation

Ashwin Ramayya
University of Pennsylvania, ashwinramayya@gmail.com

Human Reinforcement Learning: Insights from intracranial recordings and stimulation

Abstract
Reinforcement learning is the process by which individuals alter their decisions to maximize positive outcomes, and minimize negative outcomes. It is a cognitive process that is widely used in our daily lives and is often disrupted during psychiatric disease. Thus, a major goal of neuroscience is to characterize the neural underpinnings of reinforcement learning. Whereas animal studies have utilized invasive physiological methods to characterize several neural mechanisms that underlie

reinforcement learning, human studies have largely relied on non-invasive techniques that have reduced physiological precision. Although ethical limitations preclude the use of invasive physiological methods in healthy human populations, patient populations undergoing certain neurosurgical interventions offer a rare opportunity to directly assay neural activity from the brain during human reinforcement learning. This dissertation presents early findings from this research effort.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Neuroscience

First Advisor
Michael J. Kahana

Keywords
DBS, dopamine, iEEG, microstimulation, reinforcement learning, substantia nigra

Subject Categories
Neuroscience and Neurobiology

HUMAN REINFORCEMENT LEARNING: INSIGHTS FROM INTRACRANIAL RECORDINGS AND STIMULATION

Ashwin G. Ramayya

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Michael J. Kahana
Professor of Psychology

Graduate Group Chairperson

_____

Joshua I. Gold
Professor of Neuroscience

Dissertation Committee:

Joseph Kable, Baird term Assistant Professor of Psychology
Diego Contreras, Professor of Neuroscience
Kareem A. Zaghloul, Investigator, National Institutes of Health

HUMAN REINFORCEMENT LEARNING: INSIGHTS FROM INTRACRANIAL RECORDINGS AND STIMULATION

Acknowledgments

I would like to thank my mentor, Michael J. Kahana, for constantly pushing me to reach my potential. His passion for his students and human memory are both inspiring to me.

I would like to thank my committee of Joe Kable, Josh Gold, Diego Contreras, and Kareem Zaghloul, who have provided me with much guidance and support over these past few years.

Gordon H. Baltuch for mentoring me through the Deep Brain Stimulation studies and who I hope will provide clinical mentorship in the future.

Members of the Kahana lab, including, John F. Burke, Maxwell Merkow, Nicole Long, and Karl Healey for many enjoyable experiences and discussions over these past years.

I would like to thank my parents, Ramesh and Sandhya Ramayya, and my brother, Tarun Ramayya, for their love and constant support. Finally, my fiance, Jehan Bahrainwala, for her undying love, support and encouragement.

ABSTRACT

HUMAN REINFORCEMENT LEARNING: INSIGHTS FROM INTRACRANIAL RECORDINGS AND STIMULATION

Ashwin G. Ramayya

Michael J. Kahana

Reinforcement learning is the process by which individuals alter their decisions to maximize positive outcomes, and minimize negative outcomes. It is a cognitive process that is widely used in our daily lives and is often disrupted during psychiatric disease. Thus, a major goal of neuroscience is to characterize the neural underpinnings of reinforcement learning. Whereas animal studies have utilized invasive physiological methods to characterize several neural mechanisms that underlie reinforcement learning, human studies have largely relied on non-invasive techniques that have reduced physiological precision. Although ethical limitations preclude the use of invasive physiological methods in healthy human populations, patient populations undergoing certain neurosurgical interventions offer a rare opportunity to directly assay neural activity from the brain during human reinforcement learning. This dissertation presents early findings from this research effort.

Contents

List of tables

List of figures

Chapter 1

Introduction

1.1 Introduction

We are often faced with decisions that are associated with vastly distinct outcomes. For example, when a loan officer is presented with an application from a small business, she must decide whether or not to fund the application. A positive outcome would result if the business succeeds and is able to pay the bank the interest on the loan, whereas a negative outcome would result if the business does not succeed and declares bankruptcy. Some applications may be associated with a relatively high probability of a positive outcome and should be funded, whereas other applications may be associated with relatively high probability of a negative outcome and should be rejected. As the loan officer evaluates more applications, she will learn which applications should be funded and which ones should be rejects. This is an example of reinforcement learning (RL), the process by which individuals alter their decisions to maximize positive outcomes and avoid negative outcomes.

RL represents a fundamental cognitive process that is necessary for survival.

Animals must employ RL principles to forage for food in a resource-depleted environment (Stephens, 1986). Humans employ RL to acquire basic skills such as driving a car (Adams, 1987), and even to navigate interpersonal interactions (Klucharev, Hytonen, & Fernandez, 2009). Moreover, several psychiatric disorders, including drug addiction and schizophrenia, may feature pathological RL processes (Maia & Frank, 2011). For these reasons, a major goal for neuroscience is to characterize the neural processes that mediate reinforcement learning. By obtaining physiological control of these neural processes, it may be possible to develop therapies of conditions where there are deficits in RL (Redish, 2013).

Studies in animals have utilized invasive physiological methods to characterize several physiological mechanisms that underlie RL. These studies have demonstrated causal-relations between specific neural processes and learning (Reynolds, Hyland, & Wickens, 2001; Tsai et al., 2009), raising the possibility of obtaining physiological control over human RL. However, there are several challenges in generalizing findings from animal learning to human learning. For example, studies in animal learning typically study learning following primary rewards (e.g., food), whereas human learning often occurs following abstract rewards (e.g., money). Thus, it is important to study the neural processes that underlie human RL. Whereas numerous strides towards this goal have been made by studying human RL using non-invasive techniques (e.g., functional neuroimaging) the level of physiological precision is far from that achieved in animal studies. To obtain a more physiologically precise understanding of human RL, it is necessary to employ invasive methods as used in animal studies. Although ethical limitations preclude the use of these invasive methods in healthy human populations, patient populations undergoing certain neurosurgical interventions offer a rare opportunity to directly assay neural activity from the brain during human RL. This dissertation

presents early findings from this research effort.

1.1.1 Animal studies of Reinforcement Learning: Behavior, Theory, and Neuroscience

When studying how a car works, it is important to first understand the way that it moves before studying the manner in which the engine gives rise to those movements. Similarly, when studying RL, it is important to first understand the behavioral principles that govern RL, before studying the underlying neural processes. Just car's movements can be described based on a set of physical principles, RL behavior can be described in terms of a set of cognitive (or mental) processes. Although RL involves several cognitive processes, the core requirement is the formation of associations related to the selected options and resulting outcomes ("associative learning"). In this section, we discuss seminal animal studies related to associative learning involving reinforcement. We review behavioral results, the computational models that have been proposed to explain those behavioral results, and the neural processes that may implement these computational algorithms.

Pavlovian conditioning: Behavioral studies

The earliest studies in associative learning can be traced back to the work of Ivan Pavlov (1849-1946), a Russian physiologist whose studies on the digestive system earned him a Nobel Prize. Towards the end of his career, Pavlov turned his focus towards studying the formation of associations. From his work on the digestive system, Pavlov learned that dogs would salivate when food was placed in their mouth. However, he also noticed that dogs would begin to salivate following certain cues in the environment that preceded the presentation of food (for example,

the sight of a laboratory worker). Pavlov inferred that this behavior reflected associations that dogs had formed between these cues and the presentation of food over time. Pavlov referred to the food and the resulting salivation as the unconditioned stimulus and response (US, and UR), as this association did not require any training. Whereas he referred to the lab assistant's coat and the resulting salivation as the conditioned stimulus and response (CS, and CR), as this association was acquired over time.

Pavlov's classic experiments involved quantitatively measuring the acquisition of a CR (in terms of drops of salivation) as animals were exposed to multiple pairings of a novel stimulus-food pair (Pavlov, 1927). These early experiments led to several fundamental insights on associative learning, such as the gradual acquisition of the CR over many trials that can be described by a negatively-accelerated learning curve (where the probability of the conditioned response increases more steeply early during learning, and then demonstrates an asymptotic increases after several trials). Another Pavlov demonstrated that CS must be presented prior to the US in order for CR to develop. Even the simultaneous presentation of the CS and US did not result the development of the conditioned response. Together with follow-up experiments by Kamin showing that conditioned responses only emerge following stimuli that provide predictive information about an US (Kamin, 1969), these results suggest that temporal contiguity alone is not sufficient to explain the associations formed during Pavlovian conditioning. Instead, a contingent relation between the CS and US must be perceived by the animal.

Pavlovian conditioning: Computational models

Pavlov's seminal experiments inspired theorists to describe mathematical models that described the manner in which associative learning occurred during Pavlovian

conditioning. The earliest formalization of Pavlovian conditioning was proposed by Bush and Mosteller (Bush & Mosteller, 1951), who proposed that the probability of observing a conditioned response on a trial-by-trial basis could be described by the following iterative equation:

$$P(t+1) = P(t) + \alpha(R(t) - P(t))$$

where P is the probability of observing a conditioned response, t is the trial number, R represents the presence (1) or absence (0) of the unconditioned stimulus following a presentation of the conditioned stimulus and α represents a free parameter that is bound between 0 and 1. The intuition behind this model was the negatively-accelerated learning curve originally described by Pavlov. The equation suggests that the degree to which P changes on a given trial depends on α and the degree of mismatch between R and the the current value of P. Early during training (when P is low), the presentation of an unconditioned stimulus should result in large increases in P, whereas late during learning (when P is high), there should be smaller changes in P. When α is set to 1, the recent trial is heavily weighted, such that the conditioned response develops immediately following the presentation of R, and disappears following the absence of R. On the other hand, when α is set near 0, the recent trial is lightly weighted such that the conditioned response gradually develops after several presentations of R, and gradually disappears after several trials where the conditioned stimulus is presented without the unconditioned stimulus. Thus, α is often referred to as the "learning rate." However, it can also be conceptualized as a forgetting function that describes the decay of past trials ($\alpha = 1$ suggests a steep decay, whereas $\alpha = 0$ suggests a gradual decay; (Glimcher, 2011)).

Whereas Bush and Mosteller's equation provides a description of learning dynamics of associative learning, it did not provide an explanation for several behav-

ioral findings regarding Pavlovian learning. For example, Kamin's blocking effect showed that CR are only formed in association with a CS when the US is not already predicted by other stimuli in the environment. The Rescorla-Wagner model extended Bush and Mosteller's model to allow for interactions between multiple conditioned stimuli under the assumption that the animal generates expectations about upcoming rewards (US) by adding predictive information from the various stimuli in the environment. The Rescorla-Wagner model was successfully able to explain Kamin's Blocking Phenomenon and several additional findings regarding reinforcement learning (Rescorla & Wagner, 1972).

There were two major short-comings of the Rescorla-Wagner model (Niv & Montague, 2009). First, the model treated each trial as a discrete quantity of time, and therefore could not explain changes in reward expectation that may occur within a trial. Second, the model could not explain second-order conditioning, the process by which CR would develop to a stimulus that predicted an upcoming CS (e.g., a tone predicting upcoming reward). Sutton and Barto developed the temporal difference (TD) learning model to overcome these short-comings (Sutton & Barto, 1990). The model introduces several novel features. First, it assumes that the animal maintains expectations about all future rewards (V), not just rewards that are about to occur. Second, it considers each moment within a trial as carrying an independent V. Third, it iteratively updates V at each moment based on the mismatch between currently held predictions $V(t)$ and predictions that follow $V(t+1)$. Using this approach, the TD model can explain behavioral phenomenon such as second-order conditioning and predicts a back propagation of V within a trial as a function of training. The TD learning model extends beyond simple Pavlovian conditioning and can be used to explain a wide variety of complex associative learning phenomenon (Seymour et al., 2004).

In all three models, learning is thought to occur when there is a mismatch between obtained and expected outcomes. The magnitude and direction of this mismatch ("better or worse than expected") is quantified by reward prediction errors (RPEs), that modify reward expectations in the future. Positive RPEs occur when the obtained outcomes are better than expected (unexpected presence of the unconditioned stimulus), whereas negative RPEs occur when the obtained outcomes are worse than expected (unexpected absence of the unconditioned stimulus). Positive RPEs result in an increased expectation of future rewards, whereas negative RPEs result in a decreased prediction of future rewards. An important feature of the TD model is that RPEs are predicted when there is any change in the prediction of future rewards, and thus can occur following neutral stimuli that carry predictive information. In contrast, RPEs predicted by the RW and BM models should only occur following US (rewarding stimuli should result in positive RPEs, whereas aversive stimuli should result in negative RPEs).

Midbrain dopaminergic neurons and reward prediction errors

A major advance in understanding the neural basis of RL was the discovery that dopamine-releasing neurons (DA) within the midbrain demonstrated firing rate changes consistent with reward prediction errors (RPEs). During a Pavlovian conditioning task, DA neurons demonstrated phasic bursts of firing following rewards that were unexpected, and demonstrated pauses in firing when a reward was expected, but omitted (Schultz, Dayan, & Montague, 1997). These firing rate changes can be interpreted as RPEs because increased activity occurs when outcomes are better than expected, whereas pauses in activity occur when outcomes are worse than expected. Moreover, over the course of learning, DA neurons develop phasic bursts of firing following the CS, a pattern specifically predicted by TD learning

models (P. R. Montague, Dayan, & Sejnowski, 1996). The phasic bursts of DA neurons have been shown to correlate with positive RPEs on a trial-by-trial basis, but pauses in firing only showed a weak relation with negative RPEs (Bayer & Glimcher, 2005). These data suggest that phasic bursts DA neurons signal mismatches in predictions of future rewards, and may be suitable to drive RL. Several features of DA neurons make them suitable to encode RPEs and drive learning. First, they project widely throughout the brain (S. N. Haber, Fudge, & McFarland, 2000; S. Haber & Knutson, 2009), suggesting that they have the ability to modulate a variety of neural systems. Second, they are coupled by electrical gap junctions (Vandercasteele, Glowinski, & Venance, 2005) and show a predisposition to demonstrate synchronous bursts in firing rate. Third, dopamine has been shown to facilitate long-term potentiation and induce synaptic plasticity in downstream regions(Otani, Daniel, Roisin, & Crepel, 2003). Together, these properties make DA neurons an ideal candidate to compute a stereotyped RPE representation and project it widely thought the brain (Glimcher, 2011). Recent studies making use of a optical method of neural control (optogenetics) have demonstrated a causal relation between the phasic firing of DA neurons and RPEs. The phasic firing of DA neurons was sufficient to induce a place preference in freely moving mice, suggesting that it was sufficient to induce conditioning (Tsai et al., 2009). More recently, it has been shown that increasing the phasic firing of DA neurons concurrent with reward delivery increases the CR expressed by the animals, consistent with an RPE (Steinberg et al., 2013).

Instrumental conditioning: Behavioral studies

The next major advance in the study of RL came with Edward Thorndike (1874-1949) who extended Pavlov's work on associations between stimuli and involun-

tary reflexes (e.g., salivating following food) to the associations involving stimuli and voluntary actions (e.g., pressing a lever) (Thorndike, 1932). His classic experiment involved studying a cat attempting to escape a cage to access a plate of salmon that has been placed just outside the cage. The cat may begin by performing a series of random actions (e.g., scratching the floor) in an attempt to escape the cage but accidentally open the latch of the door and escape. When the cat is replaced into the cage, it would repeat the same sequence of actions in order to escape. But over many trials, the cat would settle on only selecting the actions that were necessary to open the cage (in this case, opening the latch). To explain this pattern of learning, Thorndike proposed the Law of Effect, which stated that rewarding stimuli (the food reward, "reinforcers") strengthened preceding stimulus-action associations, and thus allowed for trial-and-error learning. In the above example, the stimulus-action association of opening the latch when placed in the cage continued to strengthen over every successful trials so as to outcompete associations associated with extraneous actions (e.g., scratching the floor) that may have been reinforced on the first few trials. This form of associative learning involving the reinforcement of voluntary actions in response to a particular stimulus is referred to as operant conditioning, or instrumental learning. The study of instrumental learning was carried forward, in a more rigorous manner, by B.F. Skinner who measured responses that such as lever presses that require less effort and could be more easily measured (*The behavior of organisms: An experimental analysis*, 1938). Skinner's methods allowed for the study of choice (e.g., choosing between two levers that were associated with varying reward rates). The major theoretical contribution of this line of research was the notion that reinforcements become "stamped" into the strength of stimulus-action associations. Thorndike specifically argued against a model where an "images" of past rewards were called into mind when

making subsequent decisions (Thorndike, 1932). Thus, rewarding stimuli modulated stimulus-response associations retrospectively, rather than informing future decisions prospectively. This led to the prediction that associative learning could occur unconsciously (Thorndike, 1932).

Instrumental conditioning: Computational models

The TD learning model also resulted in a formalism that allowed for the modeling of instrumental conditioning (Sutton & Barto, 1990). A challenge involved in modeling instrumental learning is the credit assignment problem, where the agent does not know which of several preceding actions resulted in the obtained reward (Sutton & Barto, 1990). One solution to this problem is proposed by the Q-learning model that builds directly on the TD learning model. Instead of maintaining a reward expectation estimate (V) with each associated moment in the trial, the Q-learning model assumes that each unique stimulus-action pair in the environment is associated with a unique V. Then, on each trial the V is updated based on the incoming feedback using the same learning rule initially proposed by Bush and Mosteller. A simplified version of the model can be written as follows:

$$V_i(t+1) = Q_i(t) + \alpha[R(t) - Q_i(t)] \tag{1.1}$$

where $R(t) = 1$ for correct feedback, $R(t) = 0$ for incorrect feedback and α is the learning rate parameter that adjusts the manner in which previous reinforcements influence current Q values. Large α values (upper bound = 1) heavily weight recent outcomes when estimating Q, whereas small α values (lower bound = 0) incorporate reinforcements from many previous trials.

Moreover, the probability of selecting a particular action when there are mul-

tiple alternatives can be generated by comparing the Q values of the alternatives available during that trial.

$$P_i(t) = \frac{\exp(Q_i(t)/\beta)}{\sum_j \exp(Q_j(t)/\beta)} \tag{1.2}$$

β is a free parameter for inverse gain in the softmax logistic function and can accommodate different relative tendencies to exploit the current action or explore the available alternatives (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). Variants of the Q-learning model allow reward prediction error to be computed by comparing the obtained reward to the maximum valued action, or to the chosen action (Sutton & Barto, 1990).

Neural representations of value

In addition to RPEs, it is also important to characterize the the neural representations of value (V) so as to understand the manner in which those representations are modified over the course of learning. If DA neurons encode RPEs that guide learning by modifying value representations in the brain, then one might expect to identify value representations in regions that receive prominent DA inputs. Anatomical studies have shown that DA neurons send prominent projections to the striatum (S. N. Haber et al., 2000), and indeed, the firing of striatal neurons have been shown to encode the value of chosen actions (Lau & Glimcher, 2008). Based on these data, a basic neural substrate for the Q-learning model emerges—DA neurons encode RPEs following feedback and update value representations in the striatum via dopamine release (P. R. Montague et al., 1996). Directly supporting this view, dopamine release in these regions has been shown to induce synaptic plasticity at cortico-striatal synapses that correlates with instrumental learning (Reynolds et al.,

2001). In addition to the striatum, DA neurons also send projects to several other brain regions, particularly in the prefrontal cortex (S. Haber & Knutson, 2009). As such, neuronal recording studies in monkeys have identified value representations several diverse cortical regions including the orbitofrontal cortex (Padoa-Schiopa & Assad, 2006), dorsolateral prefrontal cortex (Morrison & Salzman, 2009), cingulate gyrus (Wallis & Kennerley, 2011), parietal lobe (Platt & Glimcher, 1999), and amygdala (Paton, Belova, Morrison, & Salzman, 2006). The value representations maintained in the orbitofrontal cortex have been shown to be necessary for DA neurons to encode RPEs (Takahashi et al., 2011), this is consistent with the view that DA neurons must integrate information about reward expectation and incoming feedback in order to generate the RPE signal.

1.1.2 Human studies in reinforcement learning

Before discussing human reinforcement learning, it is important to consider some major differences between studies of human and animal RL. First, animal RL learning is typically studied following primary rewards and punishments (e.g., food rewards) whereas human learning is often motivated by higher-order abstract rewards (e.g., successfully performing the experiment). Second, animal studies require long periods of intense training, whereas much of human learning occurs in novel situations. Third, the issue of whether the stimulus-response associations are unconsciously learned (a key prediction of Thordike's theory), could be directly assessed in these studies.

The earliest studies in human reinforcement learning began soon after Thorndike's work in instrumental conditioning. The major goal of these early studies was to investigate whether the general principles advance by Thordike's Law of Effect could be applied to the manner in which human's formed associations between stimuli

and responses. Greenspoon (1955) showed that the rate of occurrence of verbal responses during spontaneous speech could be modified by providing immediate feedback to the subject (e.g., the experimenter uttering the word "good"). These results were interpreted within the Law of Effect framework, to suggest that the strength stimulus-response associations that resulted in particular verbal phrases could be directly modulated based on feedback. Following studies more precisely showed that the dynamics of associative learning during reinforcement of human behavior was similar to those observed in animals during primary reinforcement (reviewed by Salzinger, 1959). However, later work led by Estes demonstrated that all human associative learning could not be explained by Law of Effect principles, but instead were likely guided by episodic memory and goal-directed decisions (Estes, 1967). Under this framework, when individuals are presented with a stimulus, make a particular response, and obtain feedback, associations are formed between all three events because they occurred close together in time. Then, when faced with the stimulus on a subsequent trial, individuals recall the past outcomes associated with each option, and make a decision by comparing each options' probabilities of providing a positive outcome. In contrast to the Law of Effect framework, where associations are formed based between stimuli and responses based on contingent feedback, within this episodic framework, associations are formed between the stimulus, response and outcome based on contiguity. In the literature on human category learning, these contrasting frameworks have been formalized as decision-bound (Ashby & Maddox, 1993) and exemplar-based models (Estes, 1986), respectively.

It is clear that Law of Effect-type models do not provide the best account of all human associative learning, however, they are able to explain behavior on certain tasks better than their episodic counterparts. For example, Gluck and Bower (1988)

demonstrated that human learning during a probabilistic classification task is better described by a Recorla Wagner learning model than by competing episodic models of categorization (e.g., exemplar models). These results suggest that human associative learning may follow Law of Effect principles in some associative learning tasks (e.g., probabilistic classification), but contiguity-based episodic principles in other associative learning tasks (e.g., list learning). In a landmark study, Knowlton, Mangles, and Squire (1996) showed that patients with Parkinson's disease (who have a dysfunctional dopaminergic system) showed deficits in probabilistic classification, whereas patients with amnesia (who have dysfunctional medial temporal lobe function) have deficits in episodic memory. Thus, humans may possess multiple systems for associative learning that are mediated by distinct neural systems. Although interactions between these systems is a highly significant and active area of research (Redish, 2013), the main goal of this dissertation is to study the neural processes that are related to the Law of Effect (dopamine-dependent) system. We discuss interactions between the multiple learning systems as a future direction (Chapter 5).

Recent studies have provided further support for the role of dopamine in human RL. (M. J. Frank, Seeberger, & O'Reilly, 2004) showed that the administration of DA agonists in patients with Parkinson's disease (hypothesized to enhance DA bursts) can improve their ability to learn from positive outcomes, but decreases their ability to learn from negative outcomes (possibly because they counteract DA pauses) during a two-alternative probabilistic learning task. These results are consistent with the view that DA neurons encode positive RPEs with increases in activity, but encode negative RPEs with decreases in activity. Rutledge, Dean, Caplin, and Glimcher (2010) used computational modeling to more precisely showed that DA agonists resulted in enhanced positive RPEs, but also showed that they resulted in

increased perseveration. Although these pharmacological studies provide important links between dopamine and human RL, there are concerns that DA agonists may improve performance in a non-specific manner. Particularly, DA agonists are known to increase tonic DA levels in the brain, that have been hypothesized to increase motivation and response vigor (Niv, Daw, Joel, & Dayan, 2007). Therefore, the improved performance observed following the administration of DA agonists may (at least in part) be driven by an improvement general arousal, rather than enhanced learning (Shiner et al., 2012). Thus, the role of phasic DA bursts in human RL is currently unknown. We attempt to address this question in Chapters 2 and 3.

In addition to pharmacological manipulations, several studies have examined the neural bases of human RL using functional neuroimaging methods (particularly, functional magnetic resonance imaging; fMRI). Several neuroimaging studies have demonstrated blood oxygen level dependent (BOLD) activity encoding of RPEs in the ventral striatum, regions that receive prominent inputs from DA neurons (McClure, Berns, & Montague, 2003; Berns, McClure, Pagnoni, & Montague, 2001; Rutledge et al., 2010). These changes in BOLD activity are thought to reflect firing rate changes from a large number of neurons, and are thought to emerge as a result of correlated inputs into the region from DA neurons. Consistent with this view, it has been shown that striatal RPE representations are dopamine-dependent and can be modulated by the administration of DA agonists (Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Chowdhury et al., 2013). Neuroimaging studies have also shown that in regions that receive prominent inputs from DA neurons (e.g., ventral striatum, ventromedial prefrontal cortex) encodes expected and obtained value (Bartra, McGuire, & Kable, 2013). There do, however, exist challenges when interpreting changes in BOLD activity in terms of the information encoding by the local neural population. Monkey single-unit studies demonstrate heterogeneous

patterns of firing rate changes within several regions that may not be detected when averaging activity with the region, as is often done in fMRI studies (Wallis & Kennerley, 2011). In an attempt to extract information from distributed neural representations, recent studies have applied multi-voxel-pattern-analyses to fMRI data during RL and have identified distributed representations of several learning-related variables (Kahnt, Heinzle, Park, & Haynes, 2011). (Vickery, Chun, & Lee, 2011) demonstrated that information about outcome valence could be interpreted from almost all cortical and subcortical regions, most of which had not been implicated in valence-encoding based on prior univariate studies. The degree to which these distributed valence signals represent RPEs is not known. We attempt to address this question in Chapter 4.

1.1.3 Studying neural basis of human reinforcement learning in neurosurgical patients

With these behavioral and cognitive principles of human RL in hand, we can return to the question of the underlying neural mechanisms. Generally, the goal is to characterize neural processes that may be related to the various facets of RL. More specifically, we can use the computational models discussed above as a guiding framework to identify the neural processes that implement those cognitive algorithms. Because these cognitive processes occur at a very rapid time scale, and may occur within a very localized neural population, it is important to utilize methods that provide a high spatial and temporal resolution when sampling underlying neural activity. Such signals can be recorded using intracranial electrophysiology where electrodes are positioned within the brain to directly sample activity from neural populations. When the implanted electrodes is small enough (1-2 μm), the

activity of individual neurons may be sampled, whereas with with larger electrodes (1-2 mm), the activity of large neural populations can be sampled. Moreover, electrical stimulation may be applied through these electrodes to modulate the activity of local neural population and study the associated behavioral changes. Whereas such methods are readily available in animal studies, they are too invasive to apply in a healthy human population. Thus, a major obstacle to a mechanistic understanding of human RL is the difficulty of obtaining direct neuronal recordings (Engel, Moll, Fried, & Ojemann, 2005). In this dissertation, we overcome this obstacle by studying neural activity in neurosurgical patients undergoing Deep Brain Stimulation (DBS) surgery for the treatment of Parkinson's Disease (PD) or intracranial electroencephalography monitoring for durg-refractory epilepsy as they perform RL tasks. A handful of studies have investigated the neural basis of RL during DBS surgery. Zaghloul et al. (2009) showed that putative DA neurons in the human substantia nigra (SN) demonstrate neural responses consistent with RPEs. Lega, Kahana, Jaggi, Baltuch, and Zaghloul (2011) and Patel et al. (2012) showed evidence for reward signaling in the ventral striatum, during a later time interval than observed in the SN, suggesting a downstream response. To our knowledge, the neural bases of RL has not previously been investigated using intracranial EEG.

1.2 Overview

In chapter 2, we obtain microelectrode recordings from the substantia nigra of patients undergoing DBS for PD. Previous studies have shown that the SN contains a population of neurons that release dopamine throughout the brain (dopaminergic neurons, DA). Animal studies have shown that DA neurons encode reward prediction errors and may play a critical role in RL (Glimcher, 2011). In this chapter, we

present a study where we assess whether there are functional differences between DA neurons and surrounding neurons in the SN.

In chapter 3, we study the causal relation between these DA neurons and human RL by applying electrical microstimulation as subjects perform the RL task. Microstimulation has been widely used in animal studies to enhance the activity of neural processes near the electrode tip (Histed, Bonin, & Reid, 2009) and assess their causal roles in behavior (Clark, Armstrong, & Moore, 2011). Even though it is routinely used during DBS procedures to improve microelectrode recordings and localization (Lafreniere-Roula, Hutchinson, Lozano, Hodaie, & Dostrovsky, 2009), it has not been applied to study of human cognition. The insights gained from microstimulation experiments would go beyond those gained from studies of patients with neurological lesions (Knowlton et al., 1996), which do not account for compensatory mechanisms, or behavioral studies which apply pharmacological agents (M. J. Frank et al., 2004; Rutledge et al., 2009; Shiner et al., 2012; Chowdhury et al., 2013), which cannot manipulate neural activity during specific time intervals relative to behavioral events. The research described in this chapter lays the groundwork for using microstimulation to alter cognitive processes in a clinical setting.

In Chapter 4, we study feedback signals that are widely distributed throughout the cortex and medial temporal lobe using intracranial electroencephalography and study their functional relevance for learning. We study changes in high frequency activity (HFA, 70-200 Hz), a known indicator of local firing rates (Manning, Jacobs, Fried, & Kahana, 2009). These results build on recent studies that have demonstrated valence representations throughout the cortex and MTL (Vickery et al., 2011).

Chapter 2

Electrophysiological evidence for functionally distinct neural populations in the human substantia nigra

Ashwin G. Ramayya, Kareem A. Zaghloul, Christoph T. Weidemann, Gordon H. Baltuch, and Michael J. Kahana (2014). *Frontiers in Human Neuroscience*, *In Press*

2.1 Abstract

The human substantia nigra (SN) is thought to consist of two functionally distinct neuronal populations—dopaminergic (DA) neurons in the *pars compacta* subregion and GABA-ergic neurons in the *pars reticulata* subregion. However, a functional dissociation between these neuronal populations has not previously been demonstrated in the awake human. Here we obtained microelectrode recordings from

the SN of patients undergoing deep brain stimulation (DBS) surgery for Parkinson's disease as they performed a two-alternative reinforcement learning task. Following positive feedback presentation, we found that putative DA and GABA neurons demonstrated distinct temporal dynamics. DA neurons demonstrated phasic increases in activity (250-500 ms post-feedback) whereas putative GABA neurons demonstrated more delayed and sustained increases in activity (500-1000 ms post-feedback). These results provide the first electrophysiological evidence for a functional dissociation between DA and GABA neurons in the human SN. We discuss possible functions for these neuronal responses based on previous findings in human and animal studies.

2.2 Introduction

Animal studies have shown that the substantia nigra (SN) consists of two functionally distinct neuronal populations—dopaminergic (DA) neurons in the *pars compacta* subregion and GABA-ergic neurons in the *pars reticulata* subregion. DA neurons have been shown to encode reward prediction errors with phasic bursts of firing, that occur when there is a mismatch between obtained and expected outcomes (Schultz et al., 1997; Bayer & Glimcher, 2005). These DA bursts are thought to guide reinforcement learning by adjusting synaptic strength in downstream regions following unexpected outcomes (Reynolds et al., 2001; Tsai et al., 2009). In contrast, GABA neurons are involved in inhibitory regulation of various brain structures including frontal cortical regions (via the thalamus), premotor brainstem nuclei and midbrain DA neurons (Carpenter, Nakano, & Kim, 1976; Hikosaka & Wurtz, 1983; Tepper, Martin, & Anderson, 1995; Henny et al., 2012). Despite these advances in the animal, the functional role of human SN neurons has not been

elucidated.

Patients undergoing deep brain stimulation (DBS) surgery for the treatment of Parkinson's Disease offer a rare opportunity to directly study the functional properties of human SN neurons (Jaggi et al., 2004). Two previous studies in patients undergoing DBS suggest a functional role for the human SN in reinforcement learning. First, it has been shown that a subset of neurons in the SN demonstrate phasic bursts of activity following unexpected rewards, consistent with a reward prediction error (Zaghloul et al., 2009). Second, microstimulation applied in the SN following rewards alters learning by enhancing the reinforcement of preceding actions (Ramayya, Misra, Baltuch, & Kahana, 2014). In both studies, the observed learning-related neural and behavioral patterns were presumed to reflect the function of a healthy subpopulation of DA neurons in the region. Although histochemical studies have shown that DA and GABA neurons co-exist in the human SN (Damier, Hirsch, Agid, & Graybiel, 1999b), a functional dissociation between these SN neural populations has not previously been demonstrated.

In this study, we sought to directly compare the response profiles of DA and GABA neurons recorded from the human SN so as to assess whether these neuron groups represent functionally distinct subpopulations. We obtained recordings from 25 subjects as they performed a two-alternative reinforcement learning task where they selected between stimuli that carried distinct reward probabilities and received positive or negative feedback following each choice. We extracted neuronal spiking activity from each unit and identified putative DA and GABA neurons based on the physiological properties of their recorded waveforms (Ungless & Grace, 2012; Joshua, Adler, Rosin, Vaadia, & Bergman, 2009; Matsumoto & Hikosaka, 2009). If DA and GABA neurons demonstrate distinct task-related responses, it would suggest that they represent functionally distinct neuronal popu-

lations.

2.3 Material & Methods

Electrophysiological recordings We obtained intra-operative microelectrode recordings from 25 Parkinsonian patients undergoing surgery for the implantation of a deep brain stimulator (DBS) in the subthalamic nucleus (STN). Patients who volunteered to take part in the study provided their informed consent during preoperative consultation and received no financial compensation for their participation. Per routine clinical protocol, Parkinson's medications were stopped on the night before surgery (12 h preoperatively); hence subjects engaged in the study while in an OFF state. The study was conducted in accordance with a University of Pennsylvania Institutional Review Board-approved protocol. During surgery, intra-operative microelectrode recordings (obtained from a 1 μm diameter tungsten tip electrode advanced with a power-assisted microdrive) were used to identify the substantia nigra (SN) and the STN as per routine clinical protocol. We obtained microelectrode recordings sampled at 25 kHz using a StimPilot recording system (16 bit analog-to-digital converter) and Spike2 data acquisition software (targeting and recording details are reported elsewhere; (Moyer, Danish, Keating, Finkel, & Baltuch, 2007)). In this study, we present data captured from the SN as subjects performed the reinforcement learning task described below (see "Reinforcement learning task").

Reinforcement learning task Subjects performed a two-alternative probability learning task which has been previously used to study reinforcement learning and value-based decision making (Figure 2.2; (L. M. Frank, Stanley, & Brown, 2004;

M. Frank, Samanta, Moustafa, & Sherman, 2007; Zaghloul et al., 2012)). During the task, three pairs of symbols (denoted here by pairs letters: AB, CD, EF) were presented in random order, and subjects were instructed to choose one of the two stimuli on each trial (Figure 2.2b). Selections were made by pressing buttons on handheld controllers placed in each hand. The three stimulus pairs were characterized by different relative rates of reward (AB, 80% vs 20%; CD, 70% vs 30%; EF, 60% vs 40%). Reward rates associated with each symbol were determined randomly prior to each session and were fixed throughout the experiment. Probabilistic feedback followed each choice. In the event of positive feedback, the screen turned green, and the sound of a cash register was presented. In the event of negative feedback, the selection screen turned red, and an error tone was presented. Each trial consisted of presentation of the stimuli, subjects response, and a 2s display of feedback. Subjects were asked to make selections which maximized their probability of obtaining positive feedback. As in previous reinforcement learning studies in the human SN (Zaghloul et al., 2009; Ramayya et al., 2014), there was no monetary payout and the provided feedback was virtual.

The rationale for including three item pairs with distinct relative reward rates is two-fold. First, we wanted to encourage learning throughout the session. Second, it allowed for the study of subthalamic nucleus neurons during decision conflict in a subsequent experiment. When possible, subjects first performed the task during the preoperative consultation, but in all cases, the task was reviewed with subjects on the morning of surgery. Further instructions were provided prior to beginning the task intra-operatively. During surgery, subjects performed the task on a laptop placed comfortably in front of them while the microelectrode was positioned in the SN. We aligned behavioral data with neural recordings by sending sync-pulses to the neural recording system from the behavioral laptop as participants performed

the task. Some participants were bilaterally implanted with DBS electrodes and performed two intra-operative sessions of the task. The 25 subjects performed 32 sessions in total with a mean (± S.D.) of 123 (±7.1) trials per session. Each session typically lasted ≈ 15 min based on participants' response times.

Extracting neuronal spiking from microelectrode recordings From each microelectrode recording, we extracted neuronal activity using the WaveClus software package (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). We band-pass filtered each voltage recording from 400 to 5000 Hz and manually removed periods of motion artifact. We identified spike events as positive or negative deflections in the voltage trace that crossed a threshold that was manually defined for each recording (≈4 S.D. about the mean amplitude of the filtered signal). The minimum duration between consecutive spike events (censor period) was set to be 1.5 ms. Spike events were subsequently clustered into units based on the first three principal components of the waveform. Noise clusters from motion artifact or power line contamination were manually invalidated. To ensure neuronal isolation, we filtered units based on established measures of isolation quality (IsoI; (Neymotin, Lyton, A.O., & A.A., 2011)). We rejected units if greater than 0.025 of their interspike intervals were refractory period violations ($< 3ms$) or if units were poorly separable from background noise in feature space ($\text{IsoI}_{BG} < 4$). If multiple units on a channel met the aforementioned criteria, but were poorly separated from each other ($\text{IsoI}_{NN} < 4$) they were considered together as a multi-unit, which is appropriate for our analyses because DA and GABA neurons are typically regionally clustered in the SN (Henny et al., 2012). We identified a total of 42 units. 7 units were excluded because of poor separation from background noise and/or refractory period violations. Of the remaining 35 units, two units were poorly separated from

each other and were combined into a multi-unit. Thus, our dataset consisted of 33 single-units ($IsoI_{BG}$ = 6.31 ± 1.13; mean ± SD), and 1 multi-unit ($IsoI_{BG}$ = 7.53, mean $IsoI_{NN}$ = 2.78). These data were identified from 17 of the 25 subjects; 18 sessions yielded one unit, whereas 8 sessions yielded two units.

Identifying putative DA and GABA activity To understand the function of SN DA and GABA neurons, we sought to extract the activity of these neuronal populations from microelectrode recordings. Because *pars compacta* and *pars reticulata* are largely interspersed in the primate SN (Poirier, Giguére, & Marchand, 1983), the location of the microelectrode relative to any anatomical landmarks is typically not used to isolate activity from these neuronal populations (also, see (Menke, Jbabdi, Miller, Matthews, & Zarei, 2010)). Instead, non-human primate electrophysiology studies usually identify putative DA and GABA units based on the properties of extracellular spike waveforms recorded on the microelectrode (Fiorillo, Yun, & Song, 2013). Previous studies which have combined electrophysiological recordings with pharmacological manipulations (Schultz & Romo, 1987) or histochemical techniques (Henny et al., 2012) have shown that DA neurons exhibit slow firing rates and broad waveforms, whereas GABA neurons display fast firing rates and narrow waveforms (Ungless & Grace, 2012). From each unit, we estimated baseline firing rate by computing the mean firing rate over the entire recording session and waveform duration by measuring the peak-to-trough duration (Barto, Singh, & Chentanez, 2004). We identified putative DA units as those which displayed baseline firing rates slower than 15 Hz and waveform durations > 0.8 ms, and GABA units as those which displayed baseline firing rates faster than 15 Hz and waveform durations < 0.8 ms; similar parameters have been used in a prior non-human primate study (Matsumoto & Hikosaka, 2009). For the multi-unit in our

dataset, we considered baseline firing rate to be the average baseline firing rate of the two contributing units to account for the artificial elevation in firing rate that results from combining units.

For each DA and GABA unit, we computed smoothed firing rates during each trial by convolving the spike train with a Gaussian kernel (half-width = 75 ms). To aggregate firing rate responses across units, we computed normalized firing rate responses for each unit. Specifically, we computed a distribution of mean firing rates shown by the unit across all trials (0-1000 ms post-stimulus and -500-2000 ms surrounding response). We z-scored the smoothed firing rate during each trial based on the mean and standard deviation of this distribution. The time intervals used for the normalization process rarely overlapped because subjects demonstrated a mean reaction time of 2047 ms (± 855 ms).

Statistical Methods For all statistical analyses, we aggregated activity within each unit and studied changes in firing rate across units. We studied firing rates from each unit in non-overlapping 250 ms windows (0–750 ms following stimulus presentation, and -500–1500 ms surrounding response trials), that were chosen *a priori* based on prior animal (Schultz et al., 1997; Cohen, Haesler, Vong, Lowell, & Uchida, 2012; Pan, Brown, & Dudman, 2013) and human (Zaghloul et al., 2009) studies of midbrain DA and GABA activity. To assess whether DA and GABA units demonstrated distinct temporal dynamics we performed a 2×2 ANOVA following the three task events (stimulus presentation, responses resulting in positive and negative feedback). We considered time interval and neuron type to be fixed effects. To account for variability that may result from obtaining multiple samples from each population, we included neuron number as random effect nested within the neuron-type fixed effect. We performed post-hoc t-tests to identify specific changes

in neural activity, and corrected for multiple comparisons using a false-discovery rate (FDR) procedure.

2.4 Results

We obtained microelectrode recordings from the substantia nigra (SN) of 25 patients (16 males, mean age = 57.36) undergoing deep brain stimulation surgery for the treatment of Parkinson's disease (PD). As per routine clinical procedure, microelectrodes were advanced into the substantia nigra (SN) in order to identify the inferior border of the subthalamic nucleus, the target for the stimulating electrode (Figure 2.1a; (Jaggi et al., 2004; Zaghloul et al., 2009)). From each SN recording, we extracted neuronal spiking activity and identified putative DA (n = 13, mean rate = 4.56 Hz, mean duration = 0.87 ms) and GABA (n = 10, mean rate = 25.0 Hz, mean duration = 0.62 ms) units based on their baseline firing rates and waveform durations (*Materials and Methods*, Figure 2.1b).

As we obtained recordings, subjects performed a two-alternative probability learning task where they were asked to select between pairs of Japanese characters by pressing buttons on hand-held controllers. Immediately following each response, they probabilistically received positive or negative feedback (Figure 2.2a). Each stimulus carried a distinct probability of reward and each pair always consisted of a high-probability and a low-probability stimulus. During each session, subjects were presented with three item pairs that varied in their relative reward rates (80/20, 70/30, and 60/40). Subjects were instructed to select stimuli that maximized their probability of receiving positive feedback. To index learning on a particular item pair, we measured the tendency that subjects demonstrated towards selecting the high-probability item during the last 10 presentations of that

item pair (Figure 2.2b). We found that subjects reliably demonstrated such a tendency during the 80/20 pair (0.69, $t(30) = 4.64, p < 0.001$). Subjects showed a trend towards such a tendency on the 70/30 pair (0.60, $p = 0.08$), but not on the 60/40 item pair (0.55, $p > 0.2$).

To compare the functional properties of DA and GABA units, we studied aggregate normalized firing rates from each population aligned to three task-related events—stimulus presentation, responses associated with positive feedback, and responses associated with negative feedback (Figure 2.3). We separately examined neural responses following responses associated with positive and negative feedback because DA units have been shown to demonstrate opposing responses during these trials (Zaghloul et al., 2009). To compare responses from the two groups during these three conditions, we binned firing rates from each unit in non-overlapping 250 ms windows (0–750 ms following stimulus presentation, and -500–1500 ms surrounding response trials) and applied two-factor ANOVAs with neuron-type and time-interval as fixed effects. We included neuron number as random effect nested within the neuron-type fixed effect to account for variability that occurs when obtaining multiple samples from each population (see *Statistical Methods*). Following positive feedback presentation, we observed a significant interaction between neuron-type and time-interval ($F(7, 183) = 6.02$, Mean Squared Error (MSE) $= 0.81, p < 0.001$) suggesting that DA and GABA neurons demonstrated distinct temporal dynamics during these trials. Post-hoc t-tests revealed that DA units demonstrated greater firing rates than GABA units during the 250-500 ms time interval ($t(21) = 2.37, p = 0.028$) whereas GABA units demonstrated greater firing rates than DA units during the 500-750 and 750-1000 ms time intervals ($t(21)$'s$> 2.52, p$'s< 0.029; false-discovery rate (FDR) corrected p's< 0.07). We did not observe significant interactions between neuron-type and time-interval

during stimulus presentation or following negative feedback (p's> 0.16). Thus, we observed distinct responses from DA and GABA neurons following positive feedback presentation, but not following stimulus or negative feedback presentation.

To assess whether differences between DA and GABA firing rates following positive feedback were driven by changes in DA activity, GABA activity or both, we studied changes in each population's firing rates from baseline. We selected the following time intervals of interest based on the results of the previous analysis: 250-500 ms ("early," when DA activity was greater than GABA activity) and 500-1000 ms ("late," when GABA activity was greater than DA activity). For DA units, we observed increased firing rate from baseline during the early time interval ($t(12) = 2.15, p = 0.052$), but did not observe significant changes in firing during the late time interval ($p > 0.2$). For GABA units, we observed the opposite pattern—we did not observe significant changes in firing during the early time interval ($p > 0.2$), but observed significant increases in firing rate during the late time interval ($t(9) = 3.29, p = 0.009$). Thus, the major changes in neural activity following positive feedback presentation included an early increase in DA activity and a late increase in GABA activity. Example DA and GABA units are shown in Figures 2.4 and 2.5, respectively.

2.5 Discussion

We studied neuronal activity in the SN of patients undergoing DBS surgery for the treatment of Parkinson's disease as they performed a two-alternative reinforcement learning task. During each trial of the task, subjects were presented with a pair of stimuli, selected one of the stimuli by pressing buttons on hand-held controllers ("response"), and immediately received positive or negative audio-visual

feedback. We identified putative DA and GABA neurons based on the physiological properties of their extracellular waveforms, and compared the functional properties of the two populations during the task.

DA and GABA neurons in the human SN are functionally distinct. Our main finding was that DA and GABA neurons demonstrated distinct temporal dynamics following responses that resulted in positive feedback. Whereas DA neurons demonstrated phasic bursts in activity (250 – 500 ms post-feedback), GABA neurons demonstrated more delayed and sustained increases in activity (500 – 1000 ms post-feedback). These results provide the first electrophysiological evidence for a functional dissociation between DA and GABA neurons in the human SN. Whereas prior histochemical studies have shown that DA and GABA neurons co-exist in the human SN (Damier et al., 1999b), the only direct evidence for a functional dissociation between these neural populations has come from animal electrophysiology studies (Schultz et al., 1997; DeLong, Crutcher, & Georgopoulos, 1983). Our findings provide a bridge between these studies by demonstrating a functional dissociation between these neural populations in the human SN. As such, our results provide electrophysiological support for neuro-computational theories of human basal ganglia function that ascribe distinct roles to these neural populations during learning and decision-making (Bogacz & Gurney, 2007).

Functional significance of phasic DA bursts. Animal electrophysiology studies have shown that DA neurons demonstrate phasic bursts of activity that correlate with reward prediction errors (Schultz et al., 1997; Bayer & Glimcher, 2005). Enhancement of these DA bursts via electrical microstimulation (Reynolds et al., 2001) or optognetics (Tsai et al., 2009) results in enhanced learning, suggesting a causal

relation between phasic DA bursts and learning. However, several factors limit the generalizability of these studies to human behavior. First, animal learning is typically studied following primary rewards and punishments (e.g., juice and airpuffs) whereas human learning is often motivated by higher-order abstract rewards (e.g., rational and social goals). Second, animals in these studies have typically undergone long periods of intense training, whereas much of human learning occurs in novel situations.

Recent studies in patients undergoing DBS surgery for Parkinson's disease suggest a functional role for phasic DA bursts in human reinforcement learning. (Zaghloul et al., 2009) demonstrated reward prediction error-like responses in a subset of SN neurons that electrophysiologically resemble DA neurons described in animal studies (putative DA neurons). The current study functionally validates the use of these electrophysiological criteria by showing that putative DA neurons demonstrate distinct post-reward responses from other neurons in the region. Consistent with our findings, (Ramayya et al., 2014) found that microstimulation applied near SN neuronal populations that showed post-reward bursts of activity and broad waveforms resulted in altered learning. Generally, our finding that putative DA neurons demonstrated post-reward bursts in activity (Figure 2.4) is consistent with their hypothesized role in providing reinforcement following rewards (Glimcher, 2011).

Similar to the (Zaghloul et al., 2009) study, we observed DA bursts 250-500 ms following feedback, which is later than DA bursts typically observed in animal studies (100-250 ms; (Niv & Montague, 2009)). The more delayed latency might be attributed to the presentation of abstract audio-visual rewards, rather than primary rewards, each of which might engage DA neurons through distinct processes (prefrontal vs. brainstem mechanisms, respectively (Glimcher, 2011)). Unlike

the (Zaghloul et al., 2009) study, however, we did not observe clear evidence that post-reward DA bursts represented a reward prediction error (although, see *Supplemental Data*). This may be because subjects demonstrated limited learning during the task. Additionally, whereas (Zaghloul et al., 2009) observed DA pauses during the 150-300 ms post-feedback interval, we did not observe reliable decreases in activity across DA neurons (although, see Figure 2.4b). This discrepancy may be explained by the fact that the negative feedback condition in the (Zaghloul et al., 2009) study was associated with an absence of reward, whereas in our study, it was associated with the presentation of a salient negative stimulus. Previous animal studies have shown that pauses in DA activity are less frequently observed following the presentation of aversive, salient stimuli (Matsumoto & Hikosaka, 2009).

Functional significance of GABA activity. In contrast to DA neurons, GABA neurons demonstrated delayed, and sustained increases in activity following positive feedback. These patterns are consistent with findings from animal studies that have have shown sustained changes in midbrain GABA activity following visual stimulus and reward presentation (Handel & Glimcher, 2000; Sato & Hikosaka, 2002; Joshua et al., 2009; Cohen et al., 2012). We speculate that these post-feedback GABA responses are related to a reciprocal interaction with DA neurons. Previous work has shown that GABA neurons demonstrate increased firing rates when exposed to dopamine (Waszczak & Walters, 1983), suggesting that DA neurons may exert excitatory control of GABA firing. Conversely, SN GABA neurons exhibit inhibitory projections onto midbrain DA neurons, and may exert inhibitory control over DA neurons (Tepper et al., 1995; Lobb, Wilson, & Paladini, 2011; Henny et al., 2012; Pan et al., 2013). Then, following a phasic DA burst, GABA neurons

might display an increase in firing rate that might act to regulate DA firing and suppress subsequent DA phasic bursts. GABA responses might be more prominent following positive compared to negative feedback if the majority of DA neurons that provide inputs to GABA neurons demonstrate preferential increases phasic activity following positive feedback compared to negative feedback. Although the majority of SN GABA neurons reside in the *pars reticulata*, a subset of GABA neurons are also known to exist in the *pars compacta* region (Ungless & Grace, 2012; Nair-Roberts et al., 2008).

Some GABA neurons also demonstrated robust pauses in activity soon after feedback was presented (see Figure 2.5). Pauses in GABA-ergic activity typically suggest a release of inhibition on downstream structures, and have been classically observed during movement and saccade generation (DeLong et al., 1983; Hikosaka & Wurtz, 1983). These pauses in activity are thought to decrease inhibition on ("disinhibit") downstream motor structures (e.g., superior colliculus; (Carpenter et al., 1976)), and allow for the execution of a movement. Thus, the observed GABA pauses may be related to some movement expressed by subjects immediately following the presentation of salient sensory stimuli during the feedback condition (possibly orienting saccades; (Hikosaka & Wurtz, 1983)). However, we are unable to test this hypothesis because we did not monitor eye movements during the study. Alternatively, the observed pauses in GABA activity may be related to decreased inhibition on DA neurons that would facilitate post-feedback DA bursting (Luscher & Ungless, 2006; Lobb et al., 2011).

Limitations We note several limitations to our study. First, we are unable to provide direct histochemical evidence that these electrophysiologically-identified neural subgroups reflect distinct neuronal populations. However, there is a large

body of evidence from animal studies suggesting that these electrophysiological criteria may be used to identify distinct midbrain neuronal populations (Ungless & Grace, 2012). As such, several animal studies rely on electrophysiological criteria alone to identify functional subpopulations within the midbrain (Matsumoto & Hikosaka, 2009; Fiorillo et al., 2013). Second, the population we studied in this experiment–patients undergoing DBS for Parkinson's disease–is known to have degeneration of neurons in SN. Ideally, one would like to study the function of SN neurons in healthy human subjects, but at present such recordings may not be ethically obtained in any other human population. Converging evidence from histochemical (Damier, Hirsch, Agid, & Graybiel, 1999a) and electrophysiological studies (Zaghloul et al., 2009; Ramayya et al., 2014) in patients with Parkinson's disease and in animals (Hollerman & Grace, 1990; Zigmond, Abercrombie, Berger, Grace, & Stricker, 1990; Wang et al., 2010) indicate that a significant population of viable DA neurons remain in the parkinsonian SN. We suggest that the observed DA and GABA responses reflect activity from the subpopulation of healthy neurons that remain in the SN.

Supplemental Data

Comparing DA and GABA responses following positive and negative feedback
To shed light on the functional properties of DA and GABA neurons, we compared their firing rates following positive and negative feedback obtained during the early and late time intervals, respectively. For DA neurons, we did not observe significant differences in activity following the two feedback conditions ($p > 0.14$). Thus, although individual DA neurons demonstrated differential activity following positive and negative feedback (Figure 2.4), we did not observe reliable differences

across the population of DA neurons, which may be due to a lack of power. For GABA neurons, we observed a trend towards greater firing rates following positive compared to negative feedback during the late time interval ($t(9) = 2.24, p = 0.052$). If GABA responses reflect a reciprocal interaction with DA neurons (see *Discussion*), more prominent tonic GABA responses following positive feedback might suggest that excitatory DA inputs onto these neurons are stronger following positive compared to negative feedback.

Relating post-reward DA bursts to reward prediction error Theories of learning posit that decisions are altered based on a reward prediction error, or the mismatch between obtained and expected rewards (Rescorla & Wagner, 1972). Previous studies have shown that DA neurons encode a reward prediction error because they selectively show post-reward bursts in activity when rewards are unexpected (Schultz et al., 1997; Zaghloul et al., 2009). Because subjects demonstrated poor learning during the task (Figure 2.2), the vast majority of rewards obtained during the task were unlikely to be predicted based on past experience, and would be classified as "unexpected." Thus, we were limited in our ability to evaluate whether post-reward DA bursts represented a reward prediction error.

Our behavioral analyses suggested that subjects demonstrated evidence of learning on the 80/20 pair, but not the 70/30, or the 60/40 pair (see *Results*, Figure 2.2). Thus, rewards obtained during the last 10 trials of the 80/20 pair would be better predicted by subjects than those obtained during the first 10 trials. To assess whether post-reward DA bursts reflected a reward prediction error, we compared DA activity during the 250-500 ms post-feedback interval following rewards obtained during the first 10 trials of the 80/20 pair ("unexpected"), and those obtained during the last 10 trials of the 80/20 pair ("expected"). We observed greater phasic

DA activity during the unexpected reward condition compared to the expected condition ($t(22) = 2.49, p = 0.02$), which is consistent with a reward prediction error. We did not observe significant differences between phasic DA activity obtained during early and late reward trials associated with the other item pairs (p's> 0.4), or following negative feedback or stimulus presentation (p's> 0.15). Also, we did not observe reliable differences in tonic GABA activity (500-1000 ms) during early and late trials, following positive feedback, negative feedback, or stimulus presentation associated with the 80/20 condition (p's> 0.29).

References

Adams, J. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin*, *101*(1).

Addison, P. S. (2002). *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance.* Bristol: Institute of Physics Publishing.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current opnion in neurobiology*, *12*(2), 169-177.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*, 6.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal ofMathematical Psychology*, *37*, 372-400.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Barto, A., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd international conference on development and learning.*

Bartra, O., McGuire, J., & Kable, J. (2013). The valuation system: A coordinate-based meta-analysis of bold fmri experiments examining the neural correlates

of subjective value. *NeuroImage*, *76*, 412-427.

Bayer, H., & Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Bayer, H., & Glimcher, P. (2007). Statistics of midbrain dopaminergic neuron spike trains in the awake primate. *Journal of Neurophysiology*, *98*(3), 1428-1439.

The behavior of organisms: An experimental analysis. (1938). Chicago.

Behrens, T., Woolrich, M. W., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214-1221.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Series B*, *57*, 289-300.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*(8), 2793-2798.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*, 442–477. doi: 10.1162/neco.2007.19.2.442

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*.

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *NeuroImage*, *85 Pt. 2*, 834–843.

Burke, J. F., Zaghloul, K. A., Jacobs, J., Williams, R. B., Sperling, M. R., Sharan, A. D., & Kahana, M. J. (2013). Synchronous and asynchronous theta and gamma activity during episodic memory formation. *Journal of Neuroscience*,

33(1), 292-304.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413.

Buzsáki, G. (2006). *Rhythms of the brain*. New York: Oxford University Press.

Buzsaki, G., Anastassiou, C., & Koch, C. (2012). The origin of extracellular fields and currents - eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, *13*, 407-419.

Carpenter, M., Nakano, K., & Kim, R. (1976). Nigrothalamic projections in the monkey demonstrated by autoradiographic technics. *Journal of Comparative Neurology*, *165*(4).

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Duzel, E., & Dolan, R. (2013). Dopamine restores reward prediction errors in old age. *Nature Neuroscience*, *16*(5), 648-653.

Clark, K., Armstrong, K., & Moore, T. (2011). Probing neural circuitry and function with electrical microstimulation. *Proceedings of the Royal Society B: Biological Sciences*.

Cohen, J., Haesler, S., Vong, L., Lowell, B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the Ventral Tegmental Area. *Nature*, *482*, 85-88.

Cools, R., Barker, R., Sahakian, B., & Robbins, T. (2001). Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. *Cerebral Cortex*, *11*, 1136–1143.

Dale, A. M., Fischl, B., & Sereno, M. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179-194.

Damier, P., Hirsch, E., Agid, Y., & Graybiel, A. M. (1999a). The substantia nigra of the human brain ii. patterns of loss of dopamine-containing neurons in

Parkinson's disease. *Brain*, *122*, 1437-1448.

Damier, P., Hirsch, E., Agid, Y., & Graybiel, A. M. (1999b). The substantia nigra of the human brain i. nigrosomes and the nigral matrix, a compartmental organization based on calbindin d28k immunohistochemistry. *Brain*, *122*(8), 1421-1436.

Daw, N., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4-6), 603-616.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876-879.

DeLong, M., Crutcher, M., & Georgopoulos, A. P. (1983). Relations between movement and single cell discharge in the substantia nigra of the behaving monkey. *Journal of Neuroscience*, *3*(8), 1599-1606.

Desikan, R., Segonne, B., Fischl, B., Quinn, B., Dickerson, B., Blacker, D., . . . Killiany, N. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968-80.

Doll, B., Shohamy, D., & Daw, N. (2014). Multiple memory systems as substrates for multiple decision systems. *Neurobiology of learning and memory*.

Duncan, S., & Barrett, L. (2007). The role of the amygdala in visual awareness. *Trends in cognitive sciences*, *11*(5), 190-192.

Engel, A. K., Moll, C. K. E., Fried, I., & Ojemann, G. A. (2005). Invasive recordings from the human brain–clinical insights and beyond. *Nature Reviews Neuroscience*, *6*, 35–47.

Estes, W. K. (1967). *Reinforcement in human learning*. Defense Technical Information Center.

Estes, W. K. (1986, Oct). Array models for category learning. *Cognitive Psychology*,

18(4), 500-549.

Fiorillo, C., Yun, S., & Song, M. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *Journal of Neuroscience*, *33*(11), 4693-709.

Fischl, B., Sereno, M., Tootell, R., & Dale, A. M. (1999). High-resolution inter-subject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*, 272-284.

Foerde, K., Race, E., Verfaellie, M., & Shohamy, D. (2013). A role for the medial temporal lobe in feedback-driven learning: Evidence from amnesia. *Journal of Neuroscience*, *33*(13), 5698-5704.

Frank, L. M., Stanley, G., & Brown, E. (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience*, *24*(35), 7681–7689.

Frank, M., Samanta, J., Moustafa, A., & Sherman, S. (2007). Hold your horses: Impulsivity , deep brain stimulation, and medication in parkinsonism. *Science*, *318*, 1309–1312.

Frank, M., & Surmeier, D. (2009). Do substantia nigra dopaminergic neurons differentiate between reward and punishment? *Journal of Molecular Cell Biology*, *1*, 15-16.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*, 1940–1943.

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, *33*(20), 8590 - 8595.

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, USA*, *108*(3), 15647-15654.

Gluck, M., & Bower, G. (1988). Evaluating an adaptive network model of human learning. *Journal Of Memory And Language*, *27*(2), 166-195.

Grattan, L., Rutledge, R., & Glimcher, P. (2011). Increased dopamine concentrations increase the perceived value of an action. In *Program No. 732.12. Society for Neuroscience Meeting Planner.* San Diego, CA.

Haber, S., & Knutson, B. (2009). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*(1), 4–26.

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, *20*(6), 2369-2382.

Handel, A., & Glimcher, P. (2000). Contextual modulation of substantia nigra pars reticulata neurons. *Journal of Neurophysiology*, *83*(5), 3042-3048.

Henny, P., Brown, M., Northrop, A., Faunes, M., Ungless, M., Magill, P., & Bolam, J. (2012). Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons. *Nature Neuroscience*, *15*(4), 613-619.

Hikosaka, O., & Wurtz, R. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. i. relation of visual and auditory responses to saccades. *Journal of Neurophyiology*, *49*(5), 1230-1253.

Histed, M., Bonin, V., & Reid, C. (2009). Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. *Neuron*, *63*, 508-522.

Hollerman, J., & Grace, A. (1990). The effects of dopamine-depleting brain lesions on the electrophysiological activity of rat Substantia Nigra dopamine neurons. *Brain Research*, *533*, 203-212.

Hunga, Y., Smith, M., Bayle, D., Mills, T., Cheyne, & Taylor, M. J. (2010). Unattended emotional faces elicit early lateralized amygdala-frontal and fusiform

activations. *NeuroImage*, *50*(2), 727-733.

Jaggi, J., Umemura, A., Hurtig, H., Siderowf, A., Colcher, A., Stern, M., & Baltuch, G. (2004). Bilateral subthalamic stimulation of the subthalamic nucleus in Parkinson's disease: surgical efficacy and prediction of outcome. *Stereotactc & Functional Neurosurgery*, *82*, 104–14.

Joshua, M., Adler, A., Rosin, B., Vaadia, E., & Bergman, H. (2009). Encoding of probabilistic rewarding and aversive events by pallidal and nigral neurons. *Journal of Neurophysiology*, *101*, 758-772.

Kable, J., & Glimcher, P. (2009). The neurobiology of decision: Consensus and controversy. *Neuron*, *63*, 733-745.

Kahnt, T., Heinzle, J., Park, S., & Haynes, J. (2011). Decoding the formation of reward predictions across learning. *Journal of Neuroscience*, *31*(41), 14624-14630.

Kamin, L. (1969). Selective association and conditioning. In N. Mackintosh & W. Honig (Eds.), *Fundamental issues in instrumental learning* (pp. 42–64). Halifax, Canada: Dalhousie University Press.

Klucharev, V., Hytonen, R. M. S. A., K., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140-151.

Knowlton, B., Mangles, J., & Squire, L. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402.

Lafreniere-Roula, M., Hutchinson, W., Lozano, A., Hodaie, M., & Dostrovsky, J. (2009). Microstimulation-induced inhibition as a tool to aid targeting the ventral border of the subthalamic nucleus. *Journal of Neurosurgery*, *111*(4), 724-728.

Lau, B., & Glimcher, P. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451-463.

Lega, B. C., Kahana, M. J., Jaggi, J. L., Baltuch, G. H., & Zaghloul, K. A. (2011). Neuronal and oscillatory activity during reward processing in the human ventral striatum. *NeuroReport*, *22*(16), 795-800.

Lobb, C., Wilson, C., & Paladini, C. (2011). High-frequency, short-latency disinhibition bursting of midbrain dopaminergic neurons. *Journal of Neurophsyiology*, *105*, 2501-2511.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157.

Luscher, C., & Ungless, M. (2006). The mechanistic classification of addictive drugs. *PLoS Medicine*, *3*(11).

Ma, S., Rinne, J., Collan, Y., Roytta, M., & Rinne, U. (1996). A quantitative morphometrical study of neuron degeneration in the substantia nigra in Parkinson's disease. *Journal of the neurological sciences*, *140*(1-2), 40–45.

Maia, T., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154-162.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003, Jul). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, *19*(3), 1233–1239.

Manning, J. R., Jacobs, J., Fried, I., & Kahana, M. J. (2009). Broadband shifts in LFP power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience*, *29*(43), 13613 - 13620.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(11), 837-841.

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.

Menke, R., Jbabdi, S., Miller, K., Matthews, P., & Zarei, M. (2010). Connectivity-based segmentation of the Substantia Nigra in human and its implications in Parkinson's disease. *Neuroimage*, *52*, 1175-1180.

Montague, P., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417-448.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Morita, K., Morishima, M., Sakai, K., & Kawaguchi, Y. (2012). Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends in Neurosciences*, *35*(8), 457-467.

Morris, J., Ohman, A., & Dolan, R. (1999). A subcortical pathway to the right amygdala for mediating "unseen" fear. *Proceedings of the National Academy of Science USA*, *96*(4), 1680-1685.

Morrison, S., & Salzman, D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *Journal of Neuroscience*, *29*(37), 11471-11483.

Moyer, J., Danish, S., Keating, G., Finkel, L., & Baltuch, G. (2007). Implementation of dual simultaneous microelectrode recording systems during deep brain stimulation surgery for Parkinson's disease: Technical note. *Operative Neurosurgery Supplement I*, *60*, E177–78.

Nair-Roberts, R., Chatelain-Badie, S., Benson, E., White-Cooper, H., Bolam, J., & Ungless, M. (2008). Stereological estimates of dopaminergic, GABA-ergic, and glutamatergic neurons in the Ventral Tegmental Area, Substantia Nigra

and Retrorubal Field in the rat. *Journal of Neuroscience*, *152*(4), 1024-1031.

Nassar, M., Rumsey, K., Wilson, R., Parikh, K., Heasly, B., & Gold, J. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*, 1040-1046.

Neymotin, S., Lyton, W., A.O., O., & A.A., F. (2011). Measuring the quality of neuronal identification in ensemble recordings. *Journal of Neuroscience*, *31*(45), 16398-16409.

Niv, Y., Daw, N., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507-520.

Niv, Y., & Montague, P. (2009). Theoretical and empirical studies of learning. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (chap. 22). London: Academic Press.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452-454.

Otani, S., Daniel, H., Roisin, M., & Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cerebral Cortex*, *13*(11), 1251-1256.

Otto, R., Gershman, S., Markman, A., & Daw, N. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751-761.

Padoa-Schiopa, C., & Assad, J. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*, 223-226.

Pan, W. X., Brown, J., & Dudman, J. (2013). Neural signals of extinction in the inhibitory microcircuit of the ventral midbrain. *Nature Neuroscience*, *16*(1), 71-78.

Patel, S., Sheth, S., Gale, J. T., Greenberg, B., Dougherty, D., & Eskandar, E. N. (2012). Single-neuron responses in the human nucleus accumbens during a financial decision-making task. *Journal of Neuroscience*, *32*(21), 7311-5.

Paton, J., Belova, M., Morrison, S., & Salzman, C. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865-870.

Pavlov, I. (1927). *Conditioned reflexes*. New York, NY, US: Oxford University Press.

Pearce, J., & Hall, G. (1980). A model for pavlovian conditioning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–555.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. (2006). Dopamine-dependent prediction errors underpin reward-seeking behavior in humans. *Nature*, *442*, 1042-1045.

Platt, M., & Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233-238.

Poirier, L., Giguére, M., & Marchand, R. (1983). Comparative morphology of the substantia nigra and ventral tegmental area in the monkey, cat and rat. *Brain Research Bulletin*, *11*, 371-397.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(23), 1102–1107.

Ramayya, A. G., Misra, A., Baltuch, G. H., & Kahana, M. J. (2014). Microstimulation of the human substantia nigra following feedback alters reinforcement learning. *Journal of Neuroscience*, *34*(20), 6887–6895.

Ray, S., & Maunsell, J. (2011). Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biology*, *9*(4), e1000610.

Redish, A. D. (2013). The mind within the brain: How we make decisions and how those decisions go wrong. *Oxford University Press*.

Rescorla, R., & Wagner, A. (1972). A theory of pavolvian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Reynolds, J., Hyland, B., & Wickens, J. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67–70.

Roesch, M., Esber, G., Li, J., Daw, N., & Schoenbaum, G. (2012). Surprise! neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190-1200.

Rutledge, R., Dean, M., Caplin, A., & Glimcher, P. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*, *30*(40), 13525-13536.

Rutledge, R., Lazzaro, S., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *Journal of Neuroscience*, *29*(48), 15104-15114.

Sato, M., & Hikosaka, O. (2002). Role of primate substantia nigra pars reticulata in reward-oriented saccadic eye movement. *Journal of Neuroscience*, *22*(6), 2363-2373.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Schultz, W., & Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology*, *57*, 201-217.

Seymour, B., O'Doherty, J., Dayan, P., Koltzenburg, M., Jones, A., Dolan, R. J., ... Frackowiak, R. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664-667.

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, D. P., K.P., & Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: evidence from parkinson's disease. *Brain*, *135*, 1871-1883.

Steinberg, E., Keiflin, R., Boivin, J., Witten, I., Deisseroth, K., & Janak, P. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966-973.

Stephens, D. (1986). *Foraging theory*. Princeton Univ. Pr.

Sugrue, L., Corrado, G., & Newsome, W. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*, 363-375.

Sutton, R., & Barto, A. (1990). Time-derivative models of pavolovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Takahashi, Y., Roesch, M., Wilson, R., Toreson, K., O'Donnell, P., Niv, Y., & Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature Neuroscience*, *14*(12), 1590-1597.

Tepper, J., Martin, L., & Anderson, D. (1995). GABA-A receptor-mediated inhibition of rat Substantia Nigra dopaminergic neurons by pars reticulata projection neurons. *Journal of Neuroscience*, *15*(4), 3092-3103.

Thorndike, E. L. (1932). *The fundamentals of learning*. New York: Bureau of Publications, Teachers College.

Tsai, H., Zhang, F., Adamatidis, A., Stuber, S., Garret, Bonci, A., Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for

behavioral conditioning. *Science*, *324*(5930), 1080-1084.

Ungless, M., & Grace, A. (2012). Are you or aren't you? Challenges associated with physiologically identifying dopamine neurons. *Trends in Neurosciences*, *35*, 422-30.

Vandercasteele, M., Glowinski, J., & Venance, L. (2005). Electrical synapses between dopaminergic neurons of the substantia nigra pars compacta. *Journal of Neuroscience*, *25*(2), 291-298.

Vickery, T., Chun, M., & Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron*, *72*(1), 166-177.

Wallis, J. D., & Kennerley, S. (2011). Contrasting roles of reward signals in the orbitofrontal and anterior cingulate cortex. *Annals of the New York Academy of Sciences*, *1239*, 33-42.

Wang, Y., Zhang, Q., Ali, U., Gui, Z., Hui, Y., Chen, L., & Wang, T. (2010). Changes in firing rate and pattern of GABA-ergic neurons in subregions of the Substantia Nigra pars reticulata in rat models of Parkinson's Disease. *Brain Research*, *1324*, 54-63.

Waszczak, B., & Walters, J. (1983). Dopamine modulation of the effects of gamma-aminobutyric acid on Substantia Nigra pars reticulata neurons. *Science*, *220*(218-221).

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human Substantia Nigra neurons encode unexpected financial rewards. *Science*, *323*, 1496–1499.

Zaghloul, K. A., Lega, B. C., Weidemann, C. T., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal activity in the human Subthalamic Nucleus encodes decision conflict during action selection. *Journal of Neuroscience*, *32*(7), 2453–2460.

Zigmond, M., Abercrombie, E., Berger, T. W., Grace, A., & Stricker, E. (1990). Compensations after lesions of central dopaminergic neurons: some clinical and basic implications. *Trends in Neurosciences*, *13*, 290-296.

Figures

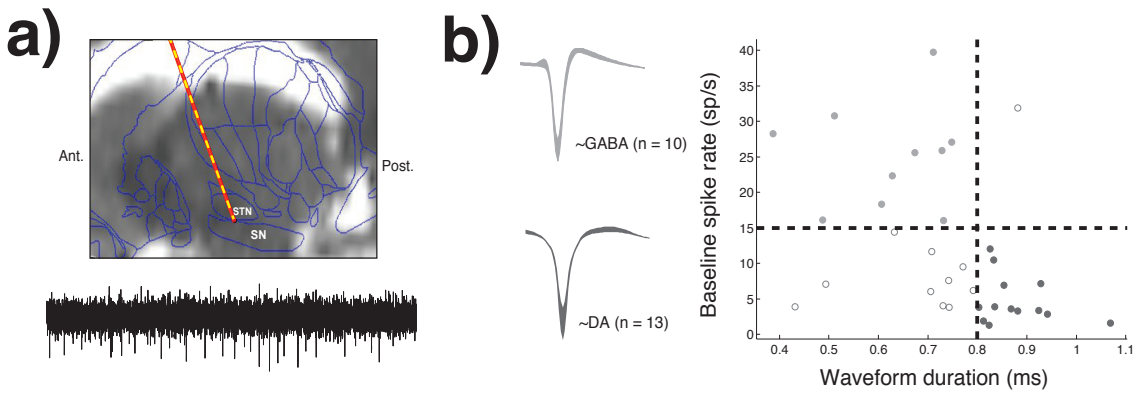

Figure 2.1: **A.** During deep brain stimulation (DBS) surgery, a microelectrode is advanced into the substantia nigra (SN) to identify the inferior border of the subthalamic nucleus (STN). Top–an example pre-operative MRI scan (sagittal view) overlaid with a standard brain atlas and estimated microelectrode position is shown. This figure is adapted from (Zaghloul et al., 2009). Bottom–an example 500 ms band-pass (400–3000 Hz) filtered signal filtered voltage trace is shown. We extracted neuronal spiking activity from each microelectrode recording by identifying spikes in the filtered signal that demonstrated sufficient separation from background noise (*Materials and Methods*). **B.** We identified putative DA (n = 13, dark grey) and GABA (n = 10, light grey) units based on their baseline firing rate and waveform durations. Left: mean waveforms from DA and GABA fast-spiking units. Width represents standard error of mean (S.E.M). Units that did not fall in either category are marked with open circles.

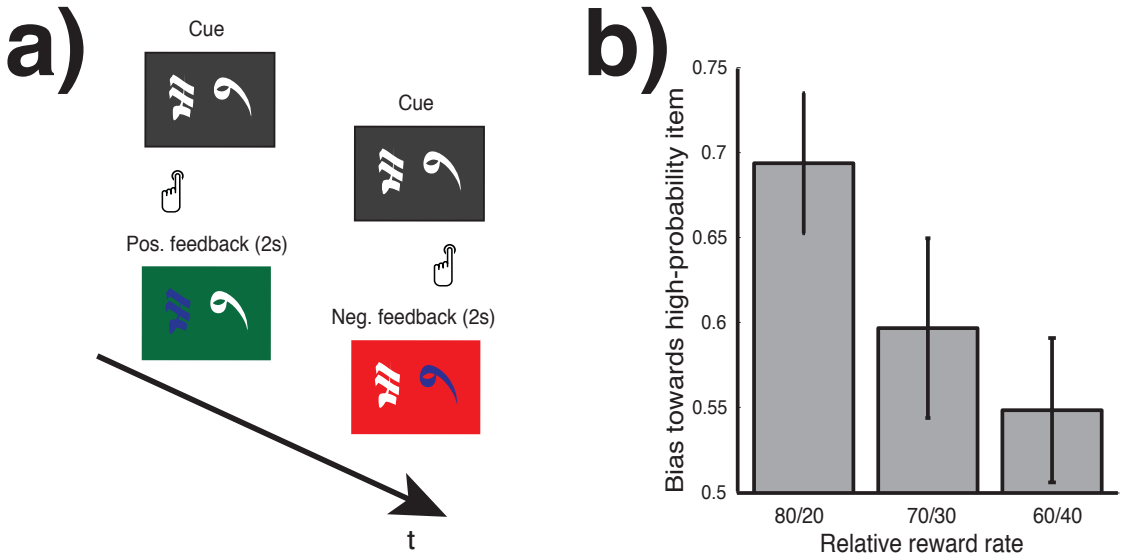

Figure 2.2: A. Reinforcement learning tasks. During surgery, subjects performed a two-alternative reinforcement learning task where they were asked to select between pairs of Japanese characters by pressing buttons on hand-held controllers. Immediately following each response, positive feedback (green screen, sound of cash register) or negative feedback (red screen, error tone) was probabilistically provided. An example positive and negative feedback trial are illustrated. Subjects were informed that each stimulus carried a distinct probability of reward and that their goal was to maximize positive feedback during the session. **B. Behavioral performance.** During each session, subjects were presented with three stimulus pairs that varied in their relative reward rates (80/20, 70/30, and 60/40). To index subjects' learning during the task, we measured their bias towards selecting the high probability item during the final 10 trials of a given pair. Subjects reliably demonstrated a bias on the 80/20 pair (0.69) and a modest bias on the 70/30 pair (0.6). We did not observe a bias on the 60/40 pair (0.55). Error bars represent S.E.M across subjects. See main text for statistics.

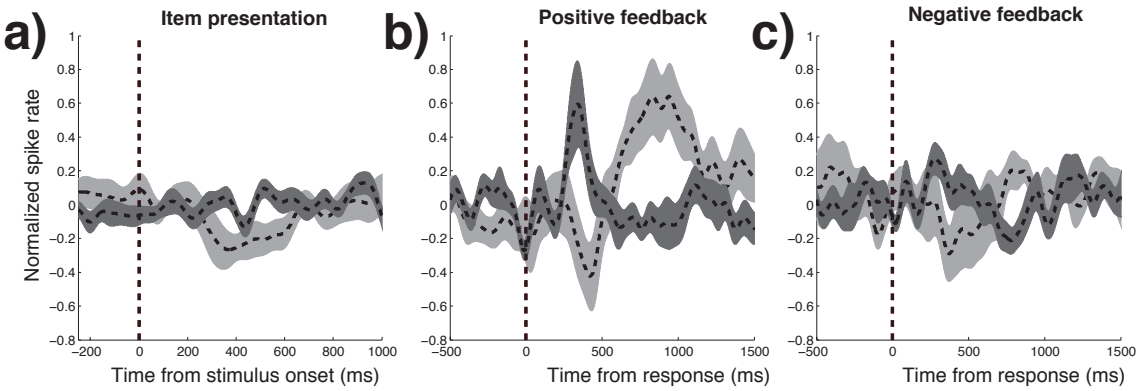

Figure 2.3: We studied aggregate normalized firing rates from DA (n = 13, dark grey) and GABA (n = 10, light grey) units in relation to three task events—stimulus presentation (**A.**) subject responses that resulted in positive feedback (**B.**), and negative feedback (**C.**). We observed distinct responses from DA and GABA units following responses associated with positive feedback, but not following stimulus presentation or responses associated with negative feedback. Firing rate responses were smoothed using a Gaussian-kernel (half-width = 75 ms). Width of each response represents S.E.M across units.

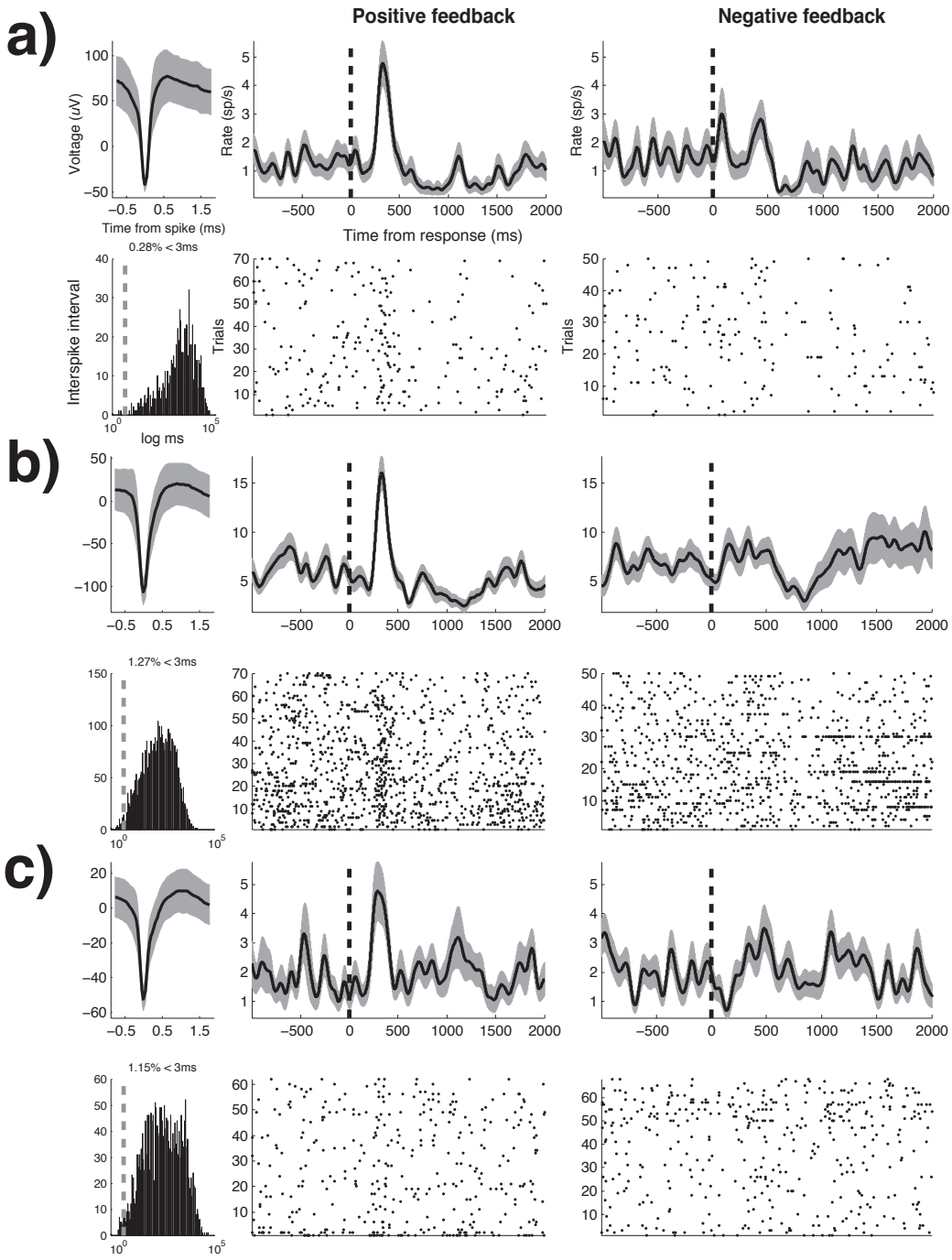

Figure 2.4: Three representative DA units are shown. For each unit, average waveform (top left), inter-spike intervals on a logarithmic time scale (bottom left, vertical line indicates 3 ms), smoothed rate (half-width = 75 ms) and raster following responses associated with positive (middle) and negative feedback (right), respectively (vertical line indicates response). Width of smooth rate represents standard error of mean (S.E.M). Baseline firing rates, waveform durations for the three units are as follows. **A.** 1.31 Hz, 0.82 ms **B.** 6.91 Hz, 0.85 ms **C.** 3.35 Hz, 0.92 ms.

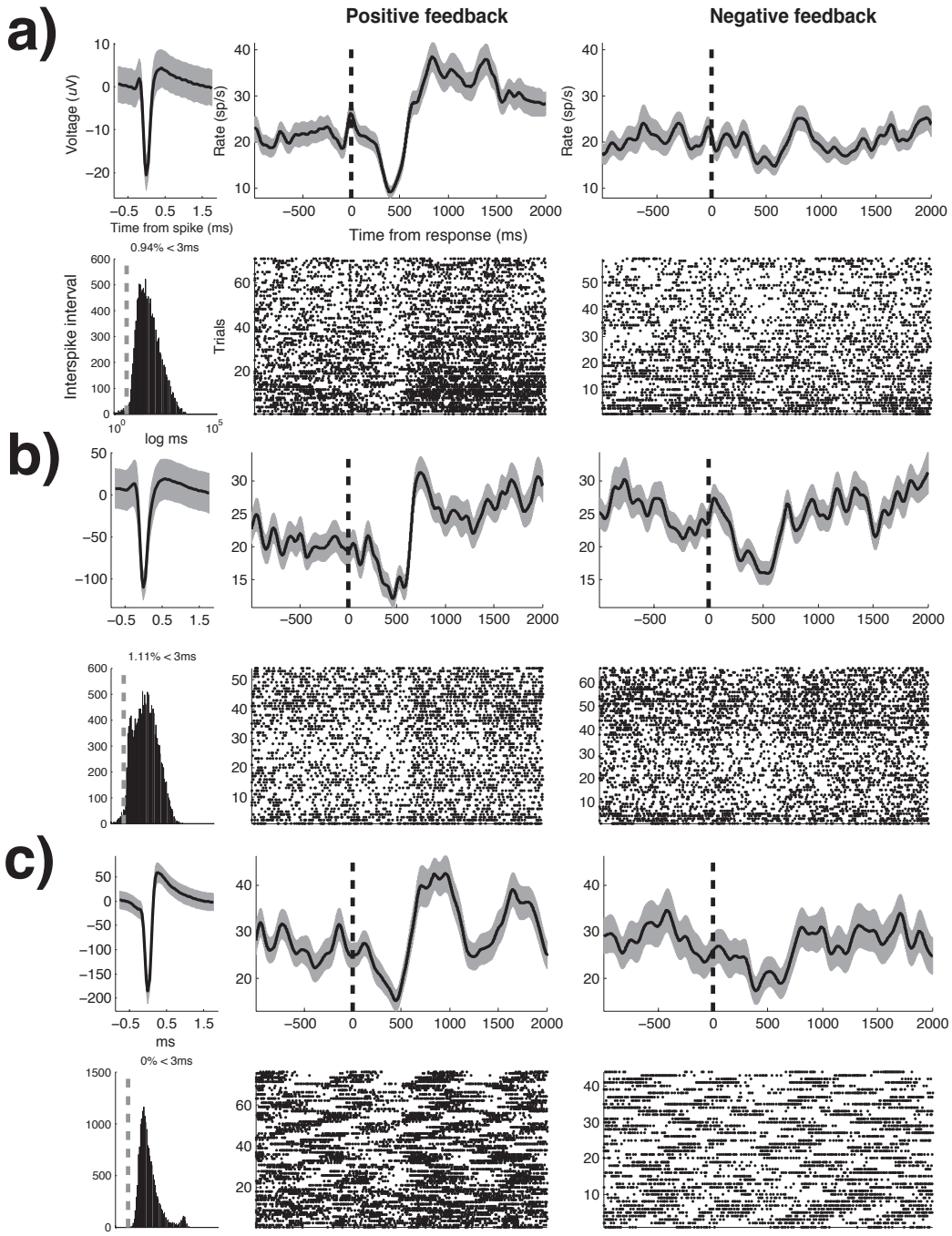

Figure 2.5: Three representative GABA units are shown. Same conventions as in Figure 2.4. Baseline firing rates and waveform durations are as follows. **A.** 25.6 Hz, 0.67 ms **B.** 27.0 Hz, 0.75 ms **C.** 28.3 Hz, 0.39 ms.

Chapter 3

Microstimulation of the human substantia nigra alters reinforcement learning

Ashwin G. Ramayya, Amrit C. Misra, Gordon H. Baltuch, and Michael J. Kahana (2014). *The Journal Neuroscience*, 34 (20), 6887-6895

3.1 Abstract

Animal studies have shown that substantia nigra dopaminergic (DA) neurons strengthen action-reward associations during reinforcement learning, but their role during human learning is not known. Here we applied microstimulation in the substantia nigra (SN) of 11 patients undergoing deep brain stimulation (DBS) surgery for the treatment of Parkinson's disease (PD) as they performed a two-alternative probability learning task, where rewards were contingent on stimuli, rather than actions. Subjects demonstrated decreased learning following reward trials that were

accompanied by phasic SN microstimulation, compared to reward trials without stimulation. Subjects who showed large decreases in learning also showed an increased bias towards repeating actions following stimulation trials; thus, stimulation may have decreased learning by strengthening action-reward associations, rather than stimulus-reward associations. Our findings build on previous studies that implicate SN DA neurons in preferentially strengthening action-reward associations during reinforcement learning.

3.2 Introduction

Contemporary theories of reinforcement learning posit that decisions are modified based on a reward prediction error (RPE), the difference between the experienced and predicted reward (Sutton and Barto, 1990). A positive RPE (outcome better than expected) strengthens associations between the reward and preceding events (e.g., stimuli, actions) such that a rewarded decision is more likely to be repeated. Animal electrophysiology studies have shown that dopaminergic (DA) neurons in the ventral tegmental area and substantia nigra (SN) display phasic bursts of activity following unexpected rewards (Schultz et al., 1997; Bayer & Glimcher, 2005), leading to the hypothesis that they encode positive RPEs (Glimcher, 2011). Because SN DA neurons predominantly send projections to dorsal striatal regions that mediate action selection (S. N. Haber et al., 2000; Lau & Glimcher, 2008), they have been hypothesized to preferentially strengthen action-reward associations during reinforcement learning (P. R. Montague et al., 1996). Supporting this hypothesis, a previous rodent study has shown that SN microstimulation reinforces actions and strengthens cortico-striatal synapses in a dopamine-dependent manner (Reynolds et al., 2001).

In humans, much of the evidence linking DA activity to reinforcement learning has come from studies in patients with Parkinson's disease (PD), who have significant degeneration of SN DA neurons (Ma, Rinne, Collan, Roytta, & Rinne, 1996) and show specific deficits on reward-based learning tasks compared to age-matched controls (Knowlton et al., 1996). Administration of DA agonists in these patients improves reinforcement learning performance (M. J. Frank et al., 2004; Rutledge et al., 2009), suggesting that DA plays an important role in human reinforcement learning. However, both PD and DA agonists manipulate tonic DA levels throughout the brain in addition to phasic DA responses. Because altered tonic DA levels may influence performance on learning tasks through non-specic changes in motivation (Niv et al., 2007), these studies do not specifically implicate the phasic activity of DA neurons in human reinforcement learning (Shiner et al., 2012).

To study the role of phasic DA activity during human reinforcement learning, we applied microstimulation in the SN of patients undergoing deep brain stimulation (DBS) surgery for the treatment of PD. Microstimulation has been shown to enhance neural responses near the electrode tip (Histed et al., 2009) and is widely used in animal electrophysiology studies to map causal relations between particular neural populations and behavior (Clark et al., 2011). Although microstimulation is often applied during DBS to aid in clinical targeting (Lafreniere-Roula et al., 2009), it has not been previously applied in association with a cognitive task. Here we applied microstimulation during the 500-ms following a subset of feedback trials as subjects performed a reinforcement learning task, where rewards were contingent on stimuli, rather than actions (putative DA neurons in the human SN have been shown to display RPE-like responses during this post-feedback time interval; (Zaghloul et al., 2009)). If phasic SN responses preferentially strengthen

action-reward associations during reinforcement learning, stimulation following reward trials should induce a bias to repeating actions, rather than stimuli, and disrupt learning during the task.

3.3 Materials and Methods

Subjects: Eleven patients undergoing deep brain stimulation (DBS) surgery for the treatment of Parkinson's Disease volunteered to take part in this study (8 male, 3 female, age = 63 ± 7, mean ± S.D). Subjects provided their informed consent during pre-operative consultation and received no financial compensation for their participation. Per routine clinical protocol, Parkinson's medications were stopped on the night before surgery (12 h preoperatively); hence subjects engaged in the study while in an OFF state. The study was conducted in accordance with a University of Pennsylvania Institutional Review Board-approved protocol.

Intra-operative methods: During surgery, intra-operative microelectrode recordings (obtained from a 1 μm diameter tungsten tip electrode advanced with a power-assisted microdrive) were used to identify the substantia nigra (SN) and the subthalamic nucleus (STN) as per routine clinical protocol ((Jaggi et al., 2004); Figure 1a). Electrical microstimulation is routinely applied through the microelectrode to aid in clinical mapping of SN and STN neurons, and was approved for use in this study by the University of Pennsylvania IRB. Once the microelectrode was positioned in the SN, we administered a two-alternative probability learning task through a laptop computer placed in front of the subject. Subjects viewed the computer screen through prism glasses placed over the stereotactic frame and expressed choices by pressing buttons on handheld controllers placed in each hand.

Reinforcement learning task: Subjects performed a two-alternative probabil-

ity learning task with feedback, which has been widely applied to the study of reinforcement learning (Figure 1b; (Sugrue, Corrado, & Newsome, 2005)). Subjects chose between pairs of items and probabilistically received positive or negative feedback following each choice. One item in each pair was associated with a high probability of reward (e.g., 0.8), whereas the other item was associated with a low probability of reward (e.g., 0.2). Subjects were informed that each stimulus in a presented pair was associated with a distinct reward rate and that their goal was to maximize rewards over the entire session. In order to achieve this goal, subjects needed to learn the underlying reward probabilities associated with each stimulus by trial and error and adjust their choices accordingly. Each trial consisted of the presentation of stimuli, subject choice, and feedback presentation. In the event of positive feedback ("wins"), the screen turned green and the sound of a cash register was presented. In the event of negative feedback ("losses"), the screen turned red and an error tone was presented. The item pairs consisted of colored images of simple objects that were matched based on normative data (e.g., semantic similarity, naming agreement, familiarity, and complexity; Rossion and Pourtois, 2004). The same pairs of stimuli were used across subjects, however, the assignment of reward probabilities to each stimulus in the pair was randomly assigned for each subject. The arrangement of the items on the screen, and thus the button associated with each item (left and right) was randomized from trial to trial.

Each session consisted of 150 trials (15 minutes of testing time) and was subdivided into three stages (50 trials each, Figure 1c). Each stage consisted of two novel pairs of stimuli (two sets of stimuli) that resulted in two independent learning conditions per stage. Such a design was used so that we could study the effects of stimulation on learning while controlling for various extraneous factors that might inuence performance. To ensure a fair comparison between the two item pairs

within each stage, the relative reward rates for each pair were set to 0.8 vs. 0.2. If the subject selected the high-probability item on at least 80% of trials on Stage 1, the relative reward rates for both pairs in subsequent stages were set to 0.7 vs. 0.3 to encourage learning during the remainder of the session, otherwise, they remained the same. Furthermore, the item pairs were presented in alternating trains of 3 to 6 trials. This method of item presentation allowed subjects to learn reward probabilities associated with a single item pair for multiple sequential trials, while ensuring that the two pairs within a stage were associated with similar levels of motivation, or arousal, which likely vary slowly throughout the session.

During Stage 1, we did not provide stimulation in association with either pair, but during the subsequent stages, we applied microstimulation following a subset of feedback trials (see *Stimulation parameters*). During Stage 2, one of the pairs was associated with SN microstimulation during *positive feedback* following a high reward-probability choice ($STIM^+$), whereas the other pair did not receive stimulation ($SHAM^+$). During Stage 3, one pair received SN microstimulation during *negative feedback* following an low reward-probability choice ($STIM^-$), whereas the other pair did not receive stimulation ($SHAM^-$). During Stage 2, we sought to assess the effect of stimulation on learning from wins by comparing performance on the $STIM^+$ and $SHAM^+$ pairs, whereas during Stage 3, we sought to assess the effect of stimulation on learning from losses by comparing performance on the $STIM^-$ and $SHAM^-$ pairs.

Because the goal of the study was to assess whether there were stimulation-related changes in learning across the various item pairs, it was crucial to minimize within-subject, across-pair variability in choice behavior. To reduce such variability, we ensured that reward probabilities of the items did not drastically fluctuate of the course of each stage by employing deterministic reward schedules (e.g.,

for a reward probability of 0.8, we ensured that 4 out of every 5 selections of that stimulus result in positive feedback). These deterministic reward schedules were not true binomial processes and may allow for distinct learning strategies than reward schedules typically used in probability learning tasks. However, by reducing within-subject variability in choice behavior, these schedules allowed us to more effectively detect stimulation-related changes in learning and take full advantage of the rare clinical opportunity offered by this patient population. When possible, subjects first performed the task during preoperative consultation, but in all cases, the task was reviewed with subjects on the morning of surgery. Further instructions were provided prior to beginning the task intra-operatively. Subject #3 did not perform Stage 1 due to a technical difficulty during the experiment, but completed Stages 2 and 3 of the task (Table 2). The design also included a fourth stage consisting of a $STIM^+$ and a $STIM^-$ pair to allow for a direct comparison between the two conditions, however, because only a subset of subjects (n = 6) completed this stage due to fatigue, these data were not analyzed for this study.

Stimulation parameters: Stimulation was provided through the microelectrode immediately following feedback presentation during the learning task using an FHC Pulsar 6b microstimulator using the following parameters: bi-phasic, cathode phase-lead pulses at 90 Hz, lasting 500 ms at an amplitude of 150 μAmps and a pulse width of 500 μs. Similar stimulation parameters have induced learning in the rodent SN (Reynolds et al., 2001) and the non-human primate VTA (Grattan, Rutledge, & Glimcher, 2011). An LED on the front chasse of the stimulator indicated the onset of stimulation, however, this was not visible to the patient as they performed the task. There was no sound associated with stimulation. Thus, stimulation trials were not signaled to subjects in any manner. None of the subjects reported a perceptual change following the application of microstimulation.

Reinforcement learning model simulations: To better understand subjects' behavior during the task, we simulated the performance of various reinforcement learning models (see below) on a two-alternative probability learning task with inconsistent stimulus-response mapping. Each simulated session consisted of 25 trials (similar to one item pair in our task) and consisted of a single item pair with reward probabilities of 0.8 and 0.2. Each item was randomly assigned to an action from trial to trial.

Q-learning model: This standard reinforcement learning model maintains independent estimates of reward expectation (Q) values for each option i at each time t (Sutton and Barto, 1990). A choice is probabilistically generated on each trial by comparing the Q values of available options on that trial using the following logistic function: $P_i(t) = \frac{\exp(Q_i(t)/\beta)}{\sum_j \exp(Q_j(t)/\beta)}$. β is a free parameter for inverse gain in the softmax logistic function (which accommodates noise in the choice process or different relative tendencies for exploration vs. exploitation; (Daw et al., 2006)). Once an item is selected by the model, feedback is received, and Q values are updated using the following learning rule: $Q_i(t+1) = Q_i(t) + \alpha[R(t) - Q_i(t)]$, where $R(t) = 1$ for correct feedback, $R(t) = 0$ for incorrect feedback and α is the learning rate parameter that adjusts the manner in which previous reinforcements influence current Q values. Large α values (upper bound = 1) heavily weight recent outcomes when estimating Q, whereas small α values (lower bound = 0) more evenly weight reinforcements from previous trials. To simulate the behavioral changes associated with decreasing learning rates, we studied the performance of 34 $Q-$model agents that varied in their α values (0.01 to 1, with a step size of 0.03; (M. Frank et al., 2007)), while fixing the β parameter at 0.2. Similarly, to simulate behavioral changes associated with increasing noise in the choice policy, we studied the performance of 34 agents that varied in their β values (0.01 to 1, with a step size of 0.03), while fixing the α

parameter at 0.2. Q values associated with each item were initialized to 0.5. We simulated the performance of these agents on 1000 randomly generated sessions.

Hybrid Action-Stimulus (AQ) learning model: To extend the Q-learning model to a task with inconsistent stimulus-response mapping, we developed a hybrid action-stimulus (AQ) learning model. Similar to the standard $Q-$model, the hybrid-AQ model tracks reward expectations associated with each stimulus using a recency-weighted exponential decay function that is controlled by the learning rate α (ranging from 0 to 1). However, in addition, the hybrid-AQ model also tracks the reward expectations associated with each available action (A). To limit addition of free parameters, the α associated with the action values is assumed to be the same for tracking stimulus and action values. A weighting parameter (W_A, ranging from 0 to 1) determines the aggregate reward expectation associated with a particular action/stimulus combination (AQ) in the following manner. $AQ_{i,j}(t) = W_A(A_i(t)) + (1 - W_A)(Q_j(t))$, where i indexes a particular stimulus, j indexes a particular action, and t represents a particular trial. Similar to the Q model, the hybrid-AQ model computes the probability of selecting from each action/stimulus combination using the following softmax-logistic function: $P_{i,j}(t) = \frac{\exp(AQ_{i,j}(t)/\beta)}{\sum_j \exp(AQ_{i*,j*}(t)/\beta)}$, where $AQ_{i*,j*}$ represents all other available action-stimulus combinations, and β is a free parameter for inverse gain in the softmax logistic function. In summary, the hybrid-AQ model has three free parameters—the learning rate (α), noise in the choice policy (β) and an action-value weighting parameter (W_A). To simulate the behavioral changes that would be observed following strengthened reward-action associations, we simulated the behavior of 34 hybrid-AQ models at various levels of the W_A parameter (0.01 to 1, with a step size of 0.03), while fixing α and β at 0.2.

Fitting reinforcement learning models to subjects' behavioral data: To directly study the relation between stimulation-related behavioral changes and the parameters of the reinforcement learning models, we fit the two-parameter Q-learning model and the three parameter hybrid-AQ model to each subject's behavioral data. We fit each model separately to subjects' choices on each item pair so as to compare changes in the parameter values across stimulation conditions. To identify the set of best-fitting parameters for a given pair, we performed a grid-search through each model's parameter space (0.01 to 1, with a step size of 0.03) and selected the set of parameters that resulted in the most positive log-likelihood estimate (LLE) of the model's predictions of the subject's choices ($i*$). $LLE = log(\prod_t P_{i*}, t)$. To assess the goodness-of-fit of each model fit across the dataset, we computed a LLE of each model's predictions of all subject choices during each item pair. To assess whether model predictions were better than chance, we computed a pseudo-R^2 statistic (r-LLE)/r, where r represents the LLE of purely random choices ($P = 0.5$ for all choices; (Daw et al., 2006)). To allow for a fair comparison between the two and three parameter model fits, we penalized each model for complexity by using the Akaike Information Criterion (AIC; (Akaike, 1974)). Because we were computing goodness-of-fit on the group level, we considered the $Q-$model to have 22 parameters (2 parameters for each subject), and the hybrid-AQ model to have 33 parameters (3 parameters for each subject).

Extracting spiking activity from microelectrode recordings: We obtained microelectrode recordings as subjects performed Stage 1 prior to applying microstimulation during the experiment. Because these recordings were of a relatively short duration ($\approx$ 5 min.) and only associated with 50 trials, their main purpose was to aid in interpretation of the stimulation results, rather than to characterize the

functional properties of human SN neuronal activity (Zaghloul et al., 2009). To assess whether stimulation-related behavioral changes were related to the properties of the neuronal population near the electrode tip, we extracted multi-unit activity from each microelectrode recording using the WaveClus software package (Quiroga et al., 2005). We band-pass filtered each voltage recording from 400 to 5000 Hz and manually removed periods of motion artifact. We identified spike events as negative deflections in the voltage trace that crossed a threshold that was manually defined for each recording (≈3.5 S.D about the mean amplitude of the filtered signal). The minimum duration between consecutive spike events (censor period) was set to be 1.5 ms. Spike events were subsequently clustered into units based on the first three Principal Components of the waveform. Noise clusters from motion artifact or power line contamination were manually invalidated. We considered spikes from all remaining clusters together as a multi-unit. From each multi-unit, we extracted two features that are characteristic of DA activity — the mean waveform duration and the phasic post-reward response (Zaghloul et al., 2009; Ungless & Grace, 2012). We quantified the waveform duration as the mean peak-to-trough duration for all spikes and the phasic post-reward response as the difference between the average spike rate during 0-500 ms post-reward interval, and that during the -250-0 and 500-750 ms intervals. We did not consider responses following negative outcomes because DA neurons are not homogenous in their responses following negative outcomes (Matsumoto and Hikosaka, 2009). We obtained multi-unit activity from 9 of the 11 subjects. We were unable to obtain recordings from one subject (#3) and could not distinguish spiking activity from noise contamination in another subject (#11).

3.4 Results

We applied microstimulation in the substantia nigra (SN) of eleven patients undergoing deep brain stimulation (DBS) surgery for the treatment of Parkinson's disease (PD; Figure 1a). Subjects performed a two–alternative probability learning task where they selected between pairs of items (images of common objects) and probabilistically received abstract rewards ("wins") or punishments ("losses") following each choice (Figure 1b). Subjects were instructed that one item in each pair carried a higher reward probability that the other item in the pair, and that their goal was to maximize the number of rewards they obtained during the session. We indexed learning on a given item pair by calculating the probability that subjects selected the high-probability item on trials associated with that pair. Because items were randomly assigned to an action (left or right button) on each trial, subjects were required to encode stimulus-reward associations, rather action-reward associations in order to perform well during the task. The task was divided into multiple stages (50 trials each) with each stage consisting of two item pairs matched in their relative reward rates (see *Materials and Methods*, Figure 1c). During Stage 1, we did not provide stimulation in association with either item pair (SHAM) so that subjects could become acclimated to the learning task. Across the 50 trials of Stage 1, subjects selected the high-probability item on 63% of trials, which trended towards being greater than chance (50%, $t(9) = 2.07, p = 0.068$). In each of the next two stages, one item pair was associated with microstimulation (STIM), whereas the other was not (SHAM). By comparing learning on the STIM and SHAM pair within each stage, we sought to assess the effects of SN microstimulation on learning.

During Stage 2, we assessed the effect of stimulation on reward learning by applying stimulation following positive outcomes associated with the high reward-

probability item on one of the pairs ($STIM^+$). We found that subjects were less likely to select the high-probability item on the $STIM^+$ pair compared to the SHAM pair during this stage ($t(10) = 2.56$, $p = 0.029$, Figure 2, Table 1). This difference in performance could be attributed to a stimulation-related decrease in learning; subjects demonstrated learning on the SHAM pair (accuracy = 67%, $t(10) = 3.05$, $p = 0.012$) but did not perform better than chance on the $STIM^+$ pair (accuracy = 48%, $p > 0.5$). To directly study the behavioral changes following stimulation trials during this stage, we compared subjects' tendencies to repeat their selection of the high-reward probability item following rewards ("win-stay") on the $STIM^+$ and the SHAM pair. We found that subjects reliably demonstrated decreased win-stay following reward trials accompanied by stimulation compared to reward trials without stimulation ($t(10) = 2.71$, $p = 0.022$). Thus, subjects demonstrated decreased learning following reward trials that were accompanied by phasic SN microstimulation compared to reward trials without stimulation. During Stage 3, we applied stimulation following negative feedback associated with the low-reward probability item on one item pair ($STIM^-$) to study the effect of SN stimulation on learning from negative outcomes. We did not observe differences in learning between the $STIM^-$ pair and the SHAM pair within the same stage, either in terms of overall accuracy (Figure 2) or their probability repeating an item choice following stimulation trials ($p's > 0.3$).

Our main finding is that SN microstimulation following rewards during Stage 2 disrupted learning of stimulus-reward associations. Because SN DA neurons have been hypothesized to preferentially strengthen action-reward associations (P. R. Montague et al., 1996; S. N. Haber et al., 2000; M. Frank & Surmeier, 2009) the observed decrease in learning might have occurred because stimulation induced a bias towards repeating actions rather than stimuli following high-probability reward trials. Such a bias would result in decreased performance because the map-

ping between stimuli and actions (left vs. right button) was randomized from trial to trial during the task; repeating the same action following the selection of a high reward-probability item would result in the selection of the low reward-probability item on approximately half the trials. If this is the case, subjects should show an increased bias towards repeating the same button following high-probability reward trials ("win-same button") on the STIM$^+$ pair compared to the SHAM pair. We did not observe a reliable stimulation-related increase in win-same button across subjects ($p > 0.4$), however, we observed a positive correlation between stimulation-related decreases in accuracy and increases in win-same button ($r = 0.77, p = 0.006$, Figure 3a). Thus, subjects who showed the greatest stimulation-related decreases in learning also showed an increased bias towards repeating actions following stimulation trials.

The positive correlation between stimulation-related decreases in accuracy and increases in win-same button suggests that stimulation may have disrupted learning by strengthening action-reward associations during the task. However, one might wonder whether this positive correlation might simply occur in association with decreased learning during our task. To assess whether this was the case, we simulated the performance of a standard two-parameter reinforcement learning model (Q-model; Sutton and Barto, 1990) performing a two-alternative probability learning task with inconsistent stimulus-response mapping (*Materials and Methods*, Figure 3b,c). Briefly, the model estimates the expected reward associated with each stimulus based on a recency-weighted average of recent outcomes (forgetting function), and probabilistically makes a selection by comparing the expected reward associated with the available options. The model has two free-parameters: a learning rate (α) that controls the rate of decay of the forgetting function, and noise in the choice policy (β). We found that both decreases in α and increases in β were

associated with decreases in accuracy and win-stay, but no accompanying change in win-same button. Thus, the positive correlation between decreased accuracy and increased win-same button cannot be explained by parametric changes in the standard two-parameter Q-model, and is not a necessary result of the task design.

To assess whether the observed stimulation-related behavioral changes could be explained by strengthened action-reward associations, we developed a hybrid action-stimulus (AQ) learning algorithm that independently tracks reward expectations associated with each available action in addition to those associated with each available stimulus (*Materials and Methods*). The model selects between available options by comparing the aggregate reward expectancies associated with the available action/stimulus combinations (e.g., house and left button press vs. candle and right button press). A weighting parameter (W_A) controls the strength of action value representations relative to stimulus value representations (higher W_A values result in strengthened action-reward associations). In total, the model has three free parameters—α (the learning rate), β (noise in the choice policy), and W_A (strength of action-reward associations). We studied the behavior of the hybrid-AQ model at various levels of W_A to simulate the behavioral changes that would be observed following strengthened action-reward associations (*Materials and Methods*, Figure 4a). We found that increasing levels of W_A were associated with decreased accuracy, decreased win-stay, and an increased win-same button. Thus, increasing the strength of action-reward associations in the hybrid-AQ model is able to explain the major stimulation-related behavioral changes, including the positive correlation between decreases in accuracy and increases in win-same button. Consistent with the behavior predicted by these model simulations, the 5 subjects who showed stimulation-related increases in win-same button showed a mean (± S.E.M.) win-same button of 0.77 (± 0.11) during the STIM$^+$ condition, and

0.48 (± 0.11) during the SHAM condition.

To directly investigate whether stimulation-related behavioral changes were related to strengthened action-reward associations, we fit the two-parameter Q-model and the three-parameter hybrid-AQ model to each subjects' choice behavior during the STIM$^+$ and SHAM conditions (*Materials and Methods*). For each subject, we identified the parameter sets that provided the best fit to subjects' choices during each pair using a grid-search across each model's parameter space. We assessed whether the three-parameter hybrid-AQ model provided a better explanation of subjects' choice behavior than the two-parameter Q-learning model using Akaike Information Criterion (AIC), a goodness-of-fit measure that applies a penalty for model complexity (Akaike, 1994). We found that the hybrid-AQ model provided a better fit to subjects' choice behavior during the STIM$^+$ condition, whereas the $Q-$model provided a better fit to subjects' choice behavior during the SHAM condition (Table 2). Then, using the parameter estimates obtained from the hybrid-AQ model, we assessed whether stimulation-related decreases in accuracy during Stage 2 were best explained by changes in α, β, or W_A by applying the following linear regression model: $R = \beta_0 + \beta_A A + \beta_B B + \beta_W W$, where R was a vector containing the decrease in accuracy for each subject. A, B and W were vectors containing changes in α, β, and W_A, respectively. We found that simulation-related decreases in accuracy demonstrated a significant, positive relation with increases in W_A ($\beta_W = 0.22, t(10) = 2.48, p = 0.017$), but not with changes in α or β (p's> 0.3). These results provide further support for the hypothesis that stimulation-related decreases in accuracy were related to strengthened action-reward associations.

Strengthened action-reward associations following feedback trials should result in improved accuracy during congruent trials (where the the rewarded item is associated with the same action as the previous trial), but decreased accuracy during

incongruent trials (where the the rewarded item is no longer associated with the same action as the previous trial). Our finding that increases in win-same button were correlated with decreases in accuracy suggests that strengthened action-reward associations may have preferentially occurred during incongruent trials. To assess whether this was the case, we studied raw probabilities of win-same button during the SHAM and STIM$^+$ pairs in subjects who showed a stimulation-related increase in win-same button, but separately for congruent and incongruent trials (n = 5, Figure 5a). During incongruent trials, these subjects showed a mean win-same button of 0.75 (± 0.19) during the STIM$^+$ condition, but a win-same button of 0.24 (± 0.15) during the SHAM condition. However, during congruent trials, these subjects showed a mean win-same button of 0.67 (± 0.21) and 0.87 (± 0.08) during the STIM$^+$ and SHAM conditions, respectively. To relate these behavioral patterns to the earlier model-based analyses, we examined the predicted win-same button probabilities of the various model simulations during congruent and incongruent trials. We found that the predictions of the Q-learning model were inconsistent with the observed behavior as both decreases in α and increases in β were associated with symmetric changes in win-same button (decreases during congruent trials and increases during incongruent trials to chance level; Figure 5 b,c). In contrast, increases in W_A of the hybrid-AQ model were associated with asymmetric changes in win-same button (increases in win-same button during incongruent trials to above chance levels, and modest decreases in win same-button during congruent trials; Figure 5d), consistent with the observed stimulation-related behavioral changes. One might have predicted that strengthened action-reward associations should result in increased win-same button following congruent trials as well as incongruent trials. However, because each action is associated with a reward probability of 0.5, this would only occur in the setting of very high α values.

These results suggest that stimulation may have strengthened action-reward associations during the task, possibly by enhancing phasic DA activity in the SN (Reynolds et al., 2001; P. R. Montague et al., 1996). Because DA neurons are anatomically clustered in the SN (Henny et al., 2012), and because microstimulation has been shown to enhance the activity of neurons that surround the electrode tip (Histed et al., 2009), one might expect to observe the greatest changes in win-same button when the microelectrode tip was positioned near DA neurons. Thus, we studied the relation between stimulation-related changes in win-same button and the properties of the neural activity recorded from the microelectrode during Stage 1. We extracted multi-unit spiking activity from each recording and extracted two features that are characteristic of DA activity—average waveform duration and the phasic post-reward response (see *Materials and Methods*; (Ungless & Grace, 2012; Zaghloul et al., 2009)). We found positive correlations between stimulation-related increases in win-same button and both the phasic post-reward response (Figure 6a, Pearson's $r = 0.69, p = 0.040$) and the mean waveform duration of recorded multi-unit activity (Figure 6b, Pearson's $r = 0.66$, $p = 0.053$). Multi-units recorded from the two subjects that showed the greatest increases in win-same button showed broad waveforms (0.85 ms, and 0.92 ms) and phasic post-reward bursts that were visible in the spike raster (+2.07 spikes/sec, and +1.43 spikes/sec; Figure 6c). These results suggest that stimulation-related increases in win-same button were greatest when the microelectrode was positioned near neural populations that displayed properties characteristic of DA neurons.

3.5 Discussion

We applied electrical microstimulation in substantia nigra (SN) of 11 patients undergoing deep brain stimulation (DBS) surgery for the treatment of Parkinson's disease (PD) as they performed a two-alternative probability learning task, where rewards were contingent on stimuli rather than actions. Subjects were required to learn stimulus-reward associations, rather than action-reward associations in order to perform well on the task. We found that SN microstimulation applied following reward trials disrupted learning compared to a control learning condition.

Phasic SN activity is functionally important for human reinforcement learning. By showing that SN microstimulation during the phasic post-reward interval alters performance during the task, our findings provide an important bridge between animal and human studies of reinforcement learning. Animal studies have shown that the phasic activity of DA neurons signal positive reward prediction errors (RPEs) that are sufficient to guide learning (Schultz et al., 1997; Bayer & Glimcher, 2005; Reynolds et al., 2001; Tsai et al., 2009), however, they may not generalize to human learning because animals in these studies have typically undergone long periods of intense training. On the other hand, human studies have not demonstrated a functional role for phasic DA activity in learning. Demonstrations of altered learning in patients with PD (Knowlton et al., 1996; Foerde, Race, Verfaellie, & Shohamy, 2013) and in association with pharmacological administration of DA agonists (M. J. Frank et al., 2004; Rutledge et al., 2009) may be driven by changes in tonic DA levels throughout the brain (that may alter learning through non-specific increase in motivation or arousal; (Niv et al., 2007)). Because SN stimulation has been shown to manipulate local neuronal activity HistEtal09,ClarEtal11, our finding that SN microstimulation during the phasic post-reward interval alters

learning provides direct evidence for the functional role of phasic SN activity in human reinforcement learning.

Relation to action-reward associations and DA activity. There are several explanations for the observed stimulation-related decrease in learning. One possibility is that microstimulation disrupted the encoding of RPEs, or increased the noise in the choice policy, both of which would result in increasingly random choices following stimulation trials. Alternatively, microstimulation may have strengthened competing action-reward associations, which would result in random item choices, but a bias towards repeating the same button press following reward trials ("win-same button").

We provide the following support for the hypothesis that stimulation enhances action-reward associations. First, we found a positive correlation between stimulation-related decreases in performance and stimulation-related increase in win-same button. Second, we showed (via simulations of the Q-learning model) that decreased learning rate or increased noise in the choice policy provide insufficient explanations of stimulation-related changes in behavior. Third, we showed that changes in the relative strength of action-reward associations in a hybrid action-stimulus (AQ) model can capture the major stimulation-related behavioral changes, including the positive correlation between stimulation-related decreases in accuracy and increased win-same button. Finally, we quantitatively fit the hybrid-AQ model to subjects' choice data and showed that stimulation-related decreases in accuracy were better explained by increases in the relative strength of action-reward associations than decreases in learning rate or increases in decision-making noise. Thus, SN microstimulation may have disrupted learning during the task by strengthening action-reward, rather than stimulus-reward associations.

One might expect strengthened action-reward associations following enhancement of phasic DA activity in the SN. Previous work has shown that SN DA neurons predominantly send their efferent projections to dorsal striatal regions which mediate action selection (S. N. Haber et al., 2000; Lau & Glimcher, 2008); thus, these neurons are hypothesized to preferentially strengthen action-reward associations during reinforcement learning (P. R. Montague et al., 1996; O'Doherty et al., 2004; M. Frank & Surmeier, 2009). Consistent with this hypothesis, we found that stimulation-related increases in win-same button were most prominent when the microelectrode was positioned near neuronal populations that demonstrated properties characteristic of DA neurons, particularly, broad waveforms and phasic post-reward responses (Zaghloul et al., 2009; Ungless & Grace, 2012). Because SN DA neurons are coupled via electrical junctions (Vandercasteele et al., 2005), stimulation near a small cluster of DA neurons might result in a spread of depolarization through a larger DA population. This interpretation is in agreement with a previous rodent study showing that microstimulation of certain SN subregions enhances action reinforcement and strengthens cortico-striatal synapses in a dopamine-dependent manner (Reynolds et al., 2001).

If SN DA neurons predominantly modulate action-reward associations, then their phasic responses should be more strongly modulated by the reward expectation associated with particular actions, rather than particular stimuli. This has not been directly tested in the human SN—the only previous demonstration of RPE-like responses from human SN DA neurons occurred during a reinforcement learning task with consistent stimulus-response mapping (Zaghloul et al., 2009). In that study, rewards were contingent on particular actions taken by the subjects, leaving open the possibility that SN DA responses were modulated by action-related reward expectancies, rather than stimulus-related reward expectancies.

Stimulation following negative feedback. Even though we observed reliable changes in learning performance when SN microstimulation was provided following positive feedback, we were unable to observe such changes when microstimulation was provided following negative feedback. These findings are consistent with previous studies which suggest that the DA system encodes positive RPEs more reliably than negative RPEs ((Bayer & Glimcher, 2005, 2007; Rutledge et al., 2009); although, see (M. J. Frank et al., 2004)). It is possible that microstimulation manipulated SN-mediated action-reward associations following negative outcomes, but that the SN's influence on learning was mitigated by the influence of a separate non-dopaminergic system that mediates learning from negative outcomes (e.g., serotonin;(Daw, Kakade, & Dayan, 2002)). Then, the behavioral changes following negative feedback stimulation might be subtle and may become evident with more data. Furthermore, because the effects of negative feedback stimulation were always tested after the effects of positive feedback stimulation, we cannot rule out a potential order effect. Future studies are needed to resolve this potential confound.

Limitations The interpretation that SN microstimulation strengthened action-reward associations by enhancing DA responses is supported by subjects' behavior following stimulation trials, functional properties of the neural population near the electrode, and is consistent with findings from previous studies. However, there are important limitations to consider. First, although we found a positive relation between stimulation-related decreases in performance and increases in win-same button, we were unable to find a reliable increase in win-same button across subjects. It may be the case that SN microstimulation had heterogeneous effects on our subjects—in some subjects it may have enhanced DA activity and strengthened action-reward associations, whereas in other subjects it may have

disrupted stimulus-reward associations by inhibiting RPE encoding (possibly by an enhancement of GABA-ergic neurons in the SN, which are known to provide inhibitory inputs onto DA neurons; (Tepper et al., 1995; Morita, Morishima, Sakai, & Kawaguchi, 2012; Pan et al., 2013)).

Second, it is important to consider the tendency of patients with PD to perseverate during cognitive tasks when interpreting our results (Cools, Barker, Sahakian, & Robbins, 2001). Rutledge et al. (2009) showed that patients with PD demonstrate choice perseveration during reinforcement learning, which was dependent on DA levels, but independent of reward history. Because stimulus-response mapping was consistent during their study, the observed perseverative effect may be specific to action selection rather than item choices. Thus, the stimulation-related increases win-same button that we observed in some of our subjects may also be explained by increased action perseveration. However, because action perseveration is not related to reward history, one would expect to observe a similar behavioral change following positive and negative feedback stimulation, which we did not observe.

Finally, the population we studied—patients undergoing DBS surgery for PD—is known to have degeneration of DA neurons in SN. Ideally, one would like to characterize the behavioral changes associated with SN microstimulation in healthy human subjects, but at present SN microstimulation may not be ethically conducted in any other human population. Certainly, this poses a challenge for interpreting findings concerning the functional role of SN neurons in patients who have degenerative disease. However, histological studies in PD patients (Damier et al., 1999a), and electrophysiological studies in rat models of PD (Hollerman & Grace, 1990; Zigmond et al., 1990), and humans (Zaghloul et al., 2009) indicate that a significant population of viable DA neurons remain in the parkinsonian SN. By

demonstrating altered reinforcement learning performance in association with SN microstimulation, our results suggest that these remaining neural processes may be functionally relevant for choice behavior.

Conclusions In this study, we show that manipulation of phasic SN activity via electrical microstimulation following rewards disrupted performance on a reinforcement learning task where rewards were contingent on stimuli, rather than actions. The greatest decreases in learning were observed when subjects showed an increased propensity to repeat the same action following rewards, suggesting that SN microstimulation strengthened action-reward associations, rather than stimulus-reward associations during the task. Although future studies are needed to rule out alternative explanations for the observed results such as disrupted RPE-encoding or increased action perseveration, our findings provide support for the hypothesis that SN DA neurons preferentially strengthen action-reward associations during reinforcement learning.

References

Adams, J. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin*, *101*(1).

Addison, P. S. (2002). *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance.* Bristol: Institute of Physics Publishing.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current opnion in neurobiology*, *12*(2), 169-177.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*, 6.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal ofMathematical Psychology*, *37*, 372-400.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Barto, A., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd international conference on development and learning.*

Bartra, O., McGuire, J., & Kable, J. (2013). The valuation system: A coordinate-based meta-analysis of bold fmri experiments examining the neural correlates

of subjective value. *NeuroImage*, *76*, 412-427.

Bayer, H., & Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Bayer, H., & Glimcher, P. (2007). Statistics of midbrain dopaminergic neuron spike trains in the awake primate. *Journal of Neurophysiology*, *98*(3), 1428-1439.

The behavior of organisms: An experimental analysis. (1938). Chicago.

Behrens, T., Woolrich, M. W., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214-1221.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Series B*, *57*, 289-300.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*(8), 2793-2798.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*, 442–477. doi: 10.1162/neco.2007.19.2.442

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*.

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *NeuroImage*, *85 Pt. 2*, 834–843.

Burke, J. F., Zaghloul, K. A., Jacobs, J., Williams, R. B., Sperling, M. R., Sharan, A. D., & Kahana, M. J. (2013). Synchronous and asynchronous theta and gamma activity during episodic memory formation. *Journal of Neuroscience*,

33(1), 292-304.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413.

Buzsáki, G. (2006). *Rhythms of the brain*. New York: Oxford University Press.

Buzsaki, G., Anastassiou, C., & Koch, C. (2012). The origin of extracellular fields and currents - eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, *13*, 407-419.

Carpenter, M., Nakano, K., & Kim, R. (1976). Nigrothalamic projections in the monkey demonstrated by autoradiographic technics. *Journal of Comparative Neurology*, *165*(4).

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Duzel, E., & Dolan, R. (2013). Dopamine restores reward prediction errors in old age. *Nature Neuroscience*, *16*(5), 648-653.

Clark, K., Armstrong, K., & Moore, T. (2011). Probing neural circuitry and function with electrical microstimulation. *Proceedings of the Royal Society B: Biological Sciences*.

Cohen, J., Haesler, S., Vong, L., Lowell, B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the Ventral Tegmental Area. *Nature*, *482*, 85-88.

Cools, R., Barker, R., Sahakian, B., & Robbins, T. (2001). Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. *Cerebral Cortex*, *11*, 1136–1143.

Dale, A. M., Fischl, B., & Sereno, M. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179-194.

Damier, P., Hirsch, E., Agid, Y., & Graybiel, A. M. (1999a). The substantia nigra of the human brain ii. patterns of loss of dopamine-containing neurons in

Parkinson's disease. *Brain*, *122*, 1437-1448.

Damier, P., Hirsch, E., Agid, Y., & Graybiel, A. M. (1999b). The substantia nigra of the human brain i. nigrosomes and the nigral matrix, a compartmental organization based on calbindin d28k immunohistochemistry. *Brain*, *122*(8), 1421-1436.

Daw, N., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4-6), 603-616.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876-879.

DeLong, M., Crutcher, M., & Georgopoulos, A. P. (1983). Relations between movement and single cell discharge in the substantia nigra of the behaving monkey. *Journal of Neuroscience*, *3*(8), 1599-1606.

Desikan, R., Segonne, B., Fischl, B., Quinn, B., Dickerson, B., Blacker, D., . . . Killiany, N. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968-80.

Doll, B., Shohamy, D., & Daw, N. (2014). Multiple memory systems as substrates for multiple decision systems. *Neurobiology of learning and memory*.

Duncan, S., & Barrett, L. (2007). The role of the amygdala in visual awareness. *Trends in cognitive sciences*, *11*(5), 190-192.

Engel, A. K., Moll, C. K. E., Fried, I., & Ojemann, G. A. (2005). Invasive recordings from the human brain–clinical insights and beyond. *Nature Reviews Neuroscience*, *6*, 35–47.

Estes, W. K. (1967). *Reinforcement in human learning*. Defense Technical Information Center.

Estes, W. K. (1986, Oct). Array models for category learning. *Cognitive Psychology*,

18(4), 500-549.

Fiorillo, C., Yun, S., & Song, M. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *Journal of Neuroscience*, *33*(11), 4693-709.

Fischl, B., Sereno, M., Tootell, R., & Dale, A. M. (1999). High-resolution inter-subject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*, 272-284.

Foerde, K., Race, E., Verfaellie, M., & Shohamy, D. (2013). A role for the medial temporal lobe in feedback-driven learning: Evidence from amnesia. *Journal of Neuroscience*, *33*(13), 5698-5704.

Frank, L. M., Stanley, G., & Brown, E. (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience*, *24*(35), 7681–7689.

Frank, M., Samanta, J., Moustafa, A., & Sherman, S. (2007). Hold your horses: Impulsivity , deep brain stimulation, and medication in parkinsonism. *Science*, *318*, 1309–1312.

Frank, M., & Surmeier, D. (2009). Do substantia nigra dopaminergic neurons differentiate between reward and punishment? *Journal of Molecular Cell Biology*, *1*, 15-16.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*, 1940–1943.

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, *33*(20), 8590 - 8595.

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, USA*, *108*(3), 15647-15654.

Gluck, M., & Bower, G. (1988). Evaluating an adaptive network model of human learning. *Journal Of Memory And Language*, *27*(2), 166-195.

Grattan, L., Rutledge, R., & Glimcher, P. (2011). Increased dopamine concentrations increase the perceived value of an action. In *Program No. 732.12. Society for Neuroscience Meeting Planner.* San Diego, CA.

Haber, S., & Knutson, B. (2009). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*(1), 4–26.

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, *20*(6), 2369-2382.

Handel, A., & Glimcher, P. (2000). Contextual modulation of substantia nigra pars reticulata neurons. *Journal of Neurophysiology*, *83*(5), 3042-3048.

Henny, P., Brown, M., Northrop, A., Faunes, M., Ungless, M., Magill, P., & Bolam, J. (2012). Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons. *Nature Neuroscience*, *15*(4), 613-619.

Hikosaka, O., & Wurtz, R. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. i. relation of visual and auditory responses to saccades. *Journal of Neurophyiology*, *49*(5), 1230-1253.

Histed, M., Bonin, V., & Reid, C. (2009). Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. *Neuron*, *63*, 508-522.

Hollerman, J., & Grace, A. (1990). The effects of dopamine-depleting brain lesions on the electrophysiological activity of rat Substantia Nigra dopamine neurons. *Brain Research*, *533*, 203-212.

Hunga, Y., Smith, M., Bayle, D., Mills, T., Cheyne, & Taylor, M. J. (2010). Unattended emotional faces elicit early lateralized amygdala-frontal and fusiform

activations. *NeuroImage*, *50*(2), 727-733.

Jaggi, J., Umemura, A., Hurtig, H., Siderowf, A., Colcher, A., Stern, M., & Baltuch, G. (2004). Bilateral subthalamic stimulation of the subthalamic nucleus in Parkinson's disease: surgical efficacy and prediction of outcome. *Stereotactc & Functional Neurosurgery*, *82*, 104–14.

Joshua, M., Adler, A., Rosin, B., Vaadia, E., & Bergman, H. (2009). Encoding of probabilistic rewarding and aversive events by pallidal and nigral neurons. *Journal of Neurophysiology*, *101*, 758-772.

Kable, J., & Glimcher, P. (2009). The neurobiology of decision: Consensus and controversy. *Neuron*, *63*, 733-745.

Kahnt, T., Heinzle, J., Park, S., & Haynes, J. (2011). Decoding the formation of reward predictions across learning. *Journal of Neuroscience*, *31*(41), 14624-14630.

Kamin, L. (1969). Selective association and conditioning. In N. Mackintosh & W. Honig (Eds.), *Fundamental issues in instrumental learning* (pp. 42–64). Halifax, Canada: Dalhousie University Press.

Klucharev, V., Hytonen, R. M. S. A., K., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140-151.

Knowlton, B., Mangles, J., & Squire, L. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402.

Lafreniere-Roula, M., Hutchinson, W., Lozano, A., Hodaie, M., & Dostrovsky, J. (2009). Microstimulation-induced inhibition as a tool to aid targeting the ventral border of the subthalamic nucleus. *Journal of Neurosurgery*, *111*(4), 724-728.

Lau, B., & Glimcher, P. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451-463.

Lega, B. C., Kahana, M. J., Jaggi, J. L., Baltuch, G. H., & Zaghloul, K. A. (2011). Neuronal and oscillatory activity during reward processing in the human ventral striatum. *NeuroReport*, *22*(16), 795-800.

Lobb, C., Wilson, C., & Paladini, C. (2011). High-frequency, short-latency disinhibition bursting of midbrain dopaminergic neurons. *Journal of Neurophsyiology*, *105*, 2501-2511.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157.

Luscher, C., & Ungless, M. (2006). The mechanistic classification of addictive drugs. *PLoS Medicine*, *3*(11).

Ma, S., Rinne, J., Collan, Y., Roytta, M., & Rinne, U. (1996). A quantitative morphometrical study of neuron degeneration in the substantia nigra in Parkinson's disease. *Journal of the neurological sciences*, *140*(1-2), 40–45.

Maia, T., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154-162.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003, Jul). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, *19*(3), 1233–1239.

Manning, J. R., Jacobs, J., Fried, I., & Kahana, M. J. (2009). Broadband shifts in LFP power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience*, *29*(43), 13613 - 13620.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(11), 837-841.

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.

Menke, R., Jbabdi, S., Miller, K., Matthews, P., & Zarei, M. (2010). Connectivity-based segmentation of the Substantia Nigra in human and its implications in Parkinson's disease. *Neuroimage*, *52*, 1175-1180.

Montague, P., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417-448.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Morita, K., Morishima, M., Sakai, K., & Kawaguchi, Y. (2012). Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends in Neurosciences*, *35*(8), 457-467.

Morris, J., Ohman, A., & Dolan, R. (1999). A subcortical pathway to the right amygdala for mediating "unseen" fear. *Proceedings of the National Academy of Science USA*, *96*(4), 1680-1685.

Morrison, S., & Salzman, D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *Journal of Neuroscience*, *29*(37), 11471-11483.

Moyer, J., Danish, S., Keating, G., Finkel, L., & Baltuch, G. (2007). Implementation of dual simultaneous microelectrode recording systems during deep brain stimulation surgery for Parkinson's disease: Technical note. *Operative Neurosurgery Supplement I*, *60*, E177–78.

Nair-Roberts, R., Chatelain-Badie, S., Benson, E., White-Cooper, H., Bolam, J., & Ungless, M. (2008). Stereological estimates of dopaminergic, GABA-ergic, and glutamatergic neurons in the Ventral Tegmental Area, Substantia Nigra

and Retrorubal Field in the rat. *Journal of Neuroscience*, *152*(4), 1024-1031.

Nassar, M., Rumsey, K., Wilson, R., Parikh, K., Heasly, B., & Gold, J. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*, 1040-1046.

Neymotin, S., Lyton, W., A.O., O., & A.A., F. (2011). Measuring the quality of neuronal identification in ensemble recordings. *Journal of Neuroscience*, *31*(45), 16398-16409.

Niv, Y., Daw, N., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507-520.

Niv, Y., & Montague, P. (2009). Theoretical and empirical studies of learning. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (chap. 22). London: Academic Press.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452-454.

Otani, S., Daniel, H., Roisin, M., & Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cerebral Cortex*, *13*(11), 1251-1256.

Otto, R., Gershman, S., Markman, A., & Daw, N. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751-761.

Padoa-Schiopa, C., & Assad, J. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*, 223-226.

Pan, W. X., Brown, J., & Dudman, J. (2013). Neural signals of extinction in the inhibitory microcircuit of the ventral midbrain. *Nature Neuroscience*, *16*(1), 71-78.

Patel, S., Sheth, S., Gale, J. T., Greenberg, B., Dougherty, D., & Eskandar, E. N. (2012). Single-neuron responses in the human nucleus accumbens during a financial decision-making task. *Journal of Neuroscience*, *32*(21), 7311-5.

Paton, J., Belova, M., Morrison, S., & Salzman, C. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865-870.

Pavlov, I. (1927). *Conditioned reflexes*. New York, NY, US: Oxford University Press.

Pearce, J., & Hall, G. (1980). A model for pavlovian conditioning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–555.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. (2006). Dopamine-dependent prediction errors underpin reward-seeking behavior in humans. *Nature*, *442*, 1042-1045.

Platt, M., & Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233-238.

Poirier, L., Giguére, M., & Marchand, R. (1983). Comparative morphology of the substantia nigra and ventral tegmental area in the monkey, cat and rat. *Brain Research Bulletin*, *11*, 371-397.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(23), 1102–1107.

Ramayya, A. G., Misra, A., Baltuch, G. H., & Kahana, M. J. (2014). Microstimulation of the human substantia nigra following feedback alters reinforcement learning. *Journal of Neuroscience*, *34*(20), 6887–6895.

Ray, S., & Maunsell, J. (2011). Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biology*, *9*(4), e1000610.

Redish, A. D. (2013). The mind within the brain: How we make decisions and how those decisions go wrong. *Oxford University Press*.

Rescorla, R., & Wagner, A. (1972). A theory of pavolvian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Reynolds, J., Hyland, B., & Wickens, J. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67–70.

Roesch, M., Esber, G., Li, J., Daw, N., & Schoenbaum, G. (2012). Surprise! neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190-1200.

Rutledge, R., Dean, M., Caplin, A., & Glimcher, P. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*, *30*(40), 13525-13536.

Rutledge, R., Lazzaro, S., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *Journal of Neuroscience*, *29*(48), 15104-15114.

Sato, M., & Hikosaka, O. (2002). Role of primate substantia nigra pars reticulata in reward-oriented saccadic eye movement. *Journal of Neuroscience*, *22*(6), 2363-2373.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Schultz, W., & Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology*, *57*, 201-217.

Seymour, B., O'Doherty, J., Dayan, P., Koltzenburg, M., Jones, A., Dolan, R. J., ... Frackowiak, R. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664-667.

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, D. P., K.P., & Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: evidence from parkinson's disease. *Brain*, *135*, 1871-1883.

Steinberg, E., Keiflin, R., Boivin, J., Witten, I., Deisseroth, K., & Janak, P. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966-973.

Stephens, D. (1986). *Foraging theory*. Princeton Univ. Pr.

Sugrue, L., Corrado, G., & Newsome, W. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*, 363-375.

Sutton, R., & Barto, A. (1990). Time-derivative models of pavolovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Takahashi, Y., Roesch, M., Wilson, R., Toreson, K., O'Donnell, P., Niv, Y., & Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature Neuroscience*, *14*(12), 1590-1597.

Tepper, J., Martin, L., & Anderson, D. (1995). GABA-A receptor-mediated inhibition of rat Substantia Nigra dopaminergic neurons by pars reticulata projection neurons. *Journal of Neuroscience*, *15*(4), 3092-3103.

Thorndike, E. L. (1932). *The fundamentals of learning*. New York: Bureau of Publications, Teachers College.

Tsai, H., Zhang, F., Adamatidis, A., Stuber, S., Garret, Bonci, A., Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for

behavioral conditioning. *Science*, *324*(5930), 1080-1084.

Ungless, M., & Grace, A. (2012). Are you or aren't you? Challenges associated with physiologically identifying dopamine neurons. *Trends in Neurosciences*, *35*, 422-30.

Vandercasteele, M., Glowinski, J., & Venance, L. (2005). Electrical synapses between dopaminergic neurons of the substantia nigra pars compacta. *Journal of Neuroscience*, *25*(2), 291-298.

Vickery, T., Chun, M., & Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron*, *72*(1), 166-177.

Wallis, J. D., & Kennerley, S. (2011). Contrasting roles of reward signals in the orbitofrontal and anterior cingulate cortex. *Annals of the New York Academy of Sciences*, *1239*, 33-42.

Wang, Y., Zhang, Q., Ali, U., Gui, Z., Hui, Y., Chen, L., & Wang, T. (2010). Changes in firing rate and pattern of GABA-ergic neurons in subregions of the Substantia Nigra pars reticulata in rat models of Parkinson's Disease. *Brain Research*, *1324*, 54-63.

Waszczak, B., & Walters, J. (1983). Dopamine modulation of the effects of gamma-aminobutyric acid on Substantia Nigra pars reticulata neurons. *Science*, *220*(218-221).

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human Substantia Nigra neurons encode unexpected financial rewards. *Science*, *323*, 1496–1499.

Zaghloul, K. A., Lega, B. C., Weidemann, C. T., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal activity in the human Subthalamic Nucleus encodes decision conflict during action selection. *Journal of Neuroscience*, *32*(7), 2453–2460.

Zigmond, M., Abercrombie, E., Berger, T. W., Grace, A., & Stricker, E. (1990). Compensations after lesions of central dopaminergic neurons: some clinical and basic implications. *Trends in Neurosciences*, *13*, 290-296.

Subject	Age	Gender	△ accuracy	△ win-stay	△ win-same button	waveform duration	phasic spike response (sp/sec)
1	67	M	+0.12	−0.50	−0.17	0.77	−1.13
2	66	M	−0.36	−0.17	+0.21	0.78	0.34
3	66	M	−0.16	+0.025	−0.17	–	–
4	53	F	+0.08	+0.028	0	0.75	1.36
5	74	M	−0.32	−0.50	+0.20	0.84	−0.86
6	54	M	−0.68	−1.00	+0.53	0.85	2.07
7	56	M	−0.28	−0.67	+0.17	0.85	1.07
8	68	M	+0.04	−0.13	−0.29	0.73	−0.73
9	53	M	−0.08	0	+0.33	0.92	1.43
10	61	F	−0.20	−0.03	−0.03	0.87	0.57
11	66	F	−0.12	−0.13	−0.13	–	–

Table 3.1: Columns 4-6 describe behavioral changes during Stage 2. Columns 7-8 describe properties of multi-unit activity recorded during Stage 1. "-" indicates missing data. We were unable to obtain recordings from Subject #3 and did not identify spiking activity from Subject #11.

Condition	α	β	W_A	AQ-model pseudo-R^2 (AIC)	Q-model pseudo-R^2 (AIC)
SHAM	0.30 (± 0.12)	0.31 (± 0.11)	0.47 (± 0.14)	0.23 (369.7)	0.20 (361.3)
STIM$^+$	0.38 (± 0.11)	0.44 (± 0.11)	0.71 (± 0.12)	0.14 (404.7)	0.07 (412.8)

Table 3.2: Mean (± S.E.M) shown for various parameter values (columns 2-4) associated with the STIM$^+$ and SHAM pairs during Stage 2. We report pseudo-R^2 and Akaike information criterion (AIC) goodness-of-fit measures for the three-parameter AQ model (column 5) and the two-parameter Q model (column 6) for each condition (*Materials and Methods*).

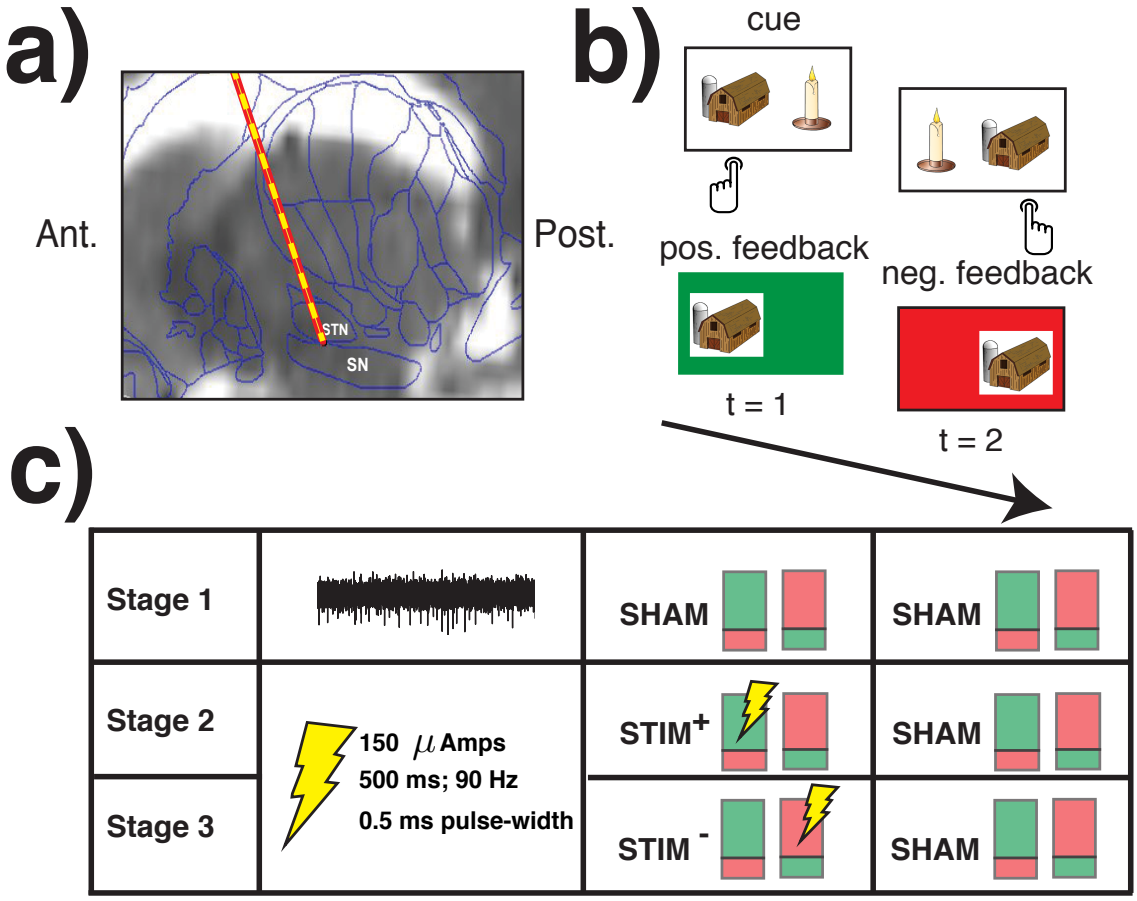

Figure 3.1: *A. Intra-operative targeting of substantia nigra.* During deep brain stimulation (DBS) surgery, a microelectrode in advanced into the substantia nigra (SN) to map the ventral border of the subthalamic nucleus (STN). An example pre-operative MRI scan (sagittal view) overlaid with a standard brain atlas and estimated microelectrode position is shown (Jaggi et al.,2004; Zaghloul et al.,2009). *B. Reinforcement learning task.* During surgery, 11 subjects performed a two-alternative probability learning task with inconsistent stimulus-response mapping. *C. Experimental design.* During each stage of the session (50 trials each), subjects sampled reward probabilities of two item pairs that were matched in relative reward rate. Each pair of colored rectangles depicts an item pair (the green and red shading within each rectangle indicates the probability of positive and negative feedback associated a particular item in the pair). During Stage 1, we obtained microelectrode recordings from the SN. An example 500-ms high-pass (> 300 Hz) filtered voltage trace is shown. During Stages 2 and 3, we applied electrical microstimulation through the recording microelectrode as depicted, but no longer obtained recordings (see *Materials and Methods*)

.

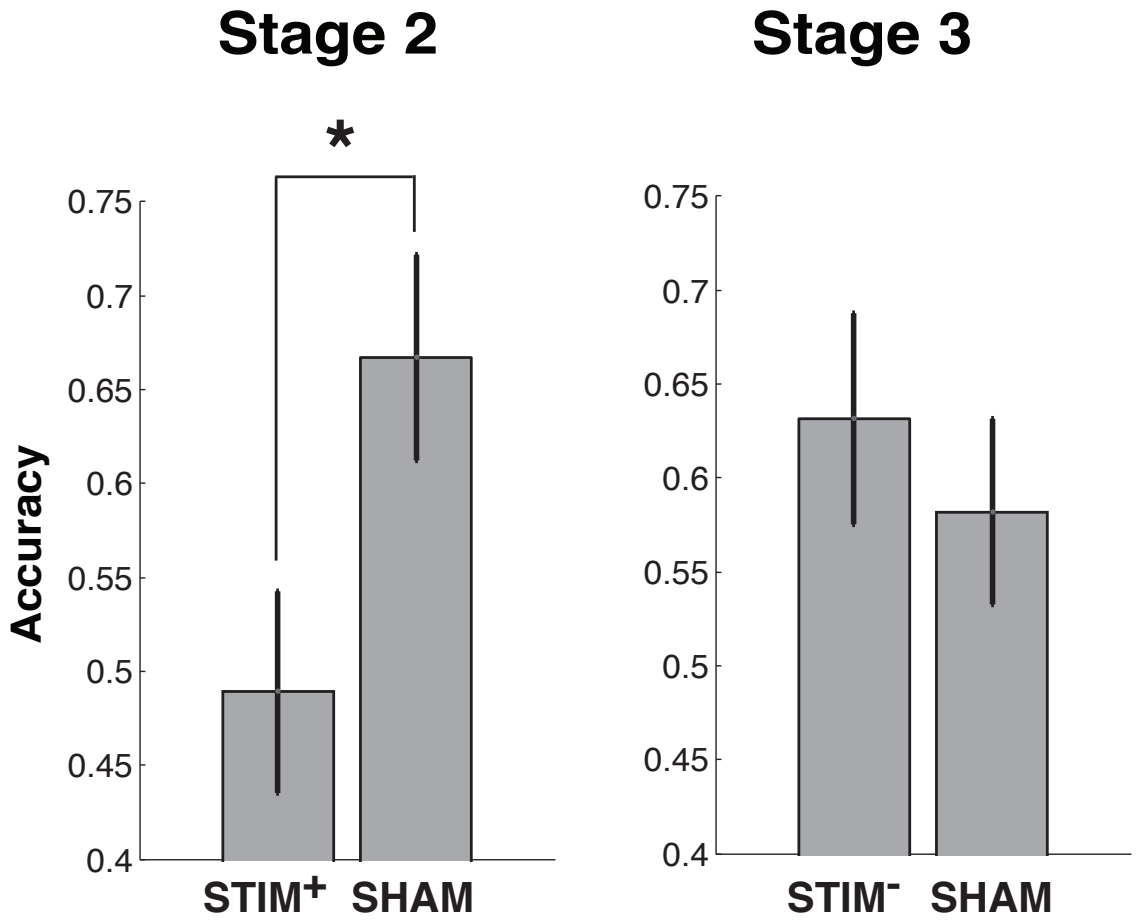

Figure 3.2: To index learning performance on a particular item pair, we computed the probability that subjects chose the item that was associated with a high reward-probability ("accuracy"). During Stage 2, subjects demonstrated lower accuracy on the $STIM^+$ pair compared to the SHAM pair. During Stage 3, we did not identify changes in accuracy between the $STIM^-$ and SHAM pairs. "*" indicates $p < 0.05$; error bars reflect standard error of the mean across subjects (n=11)

.

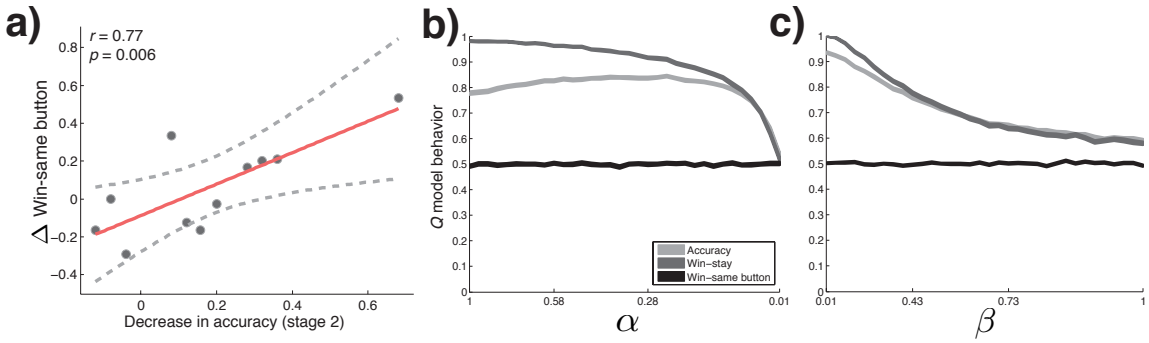

Figure 3.3: Stimulation-related decreases in accuracy were positively correlated with an increased bias towards repeating a button press following reward trials (win-same button; Pearson's $r = 0.77$, $p = 0.006$). Each dot represents a subject, the solid red line is the regression slope, and the dashed lines represent 95 % confidence intervals. *B,C. Q-learning model is insufficient to explain stimulation-related behavioral changes.* Simulated behavior of a standard two-parameter reinforcement learning algorithm (Q-model) on a two-alternative probability learning task with inconsistent stimulus-response mapping. Accuracy (light grey line), probability of repeating rewarded items (win-stay, dark grey line) and probability of repeating rewarded actions (win-same button, black line) are shown for decreasing learning rates (α; B) and increasing noise in the choice policy (β; C). Decreases in learning rate and increases in decision noise were accompanied by a decrease in accuracy and a decrease in win-stay, but no change in win-same button.

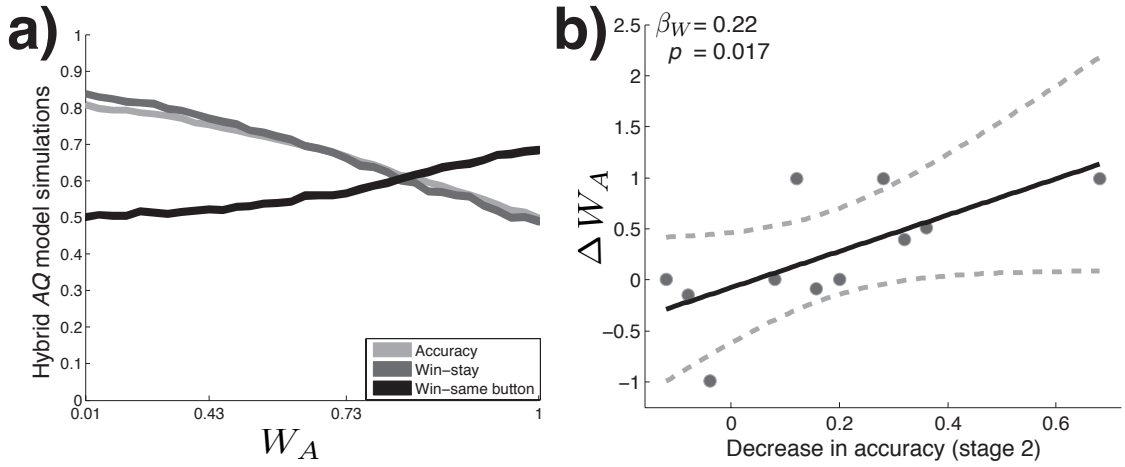

Figure 3.4: Simulated behavior of the three-parameter reinforcement learning algorithm (hybrid-AQ model) on a two-alternative probability learning task with inconsistent stimulus-response mapping. Accuracy (light grey line), probability of repeating rewarded items (win-stay, dark grey line) and probability of repeating rewarded actions (win-same button, black line) are shown for varying values of the action value weighting parameter (W_A). Strengthened action-reward associations were associated with decreases in accuracy, win-stay, and increases in win-same button. *B. Stimulation-related behavioral changes can be explained by strengthened action-reward associations.* We quantitatively fit the hybrid-AQ model to subjects' behavior on the STIM$^+$ and SHAM pair during Stage 2. We found that stimulation-related decreases in accuracy showed a significant positive relation with increases in W_A, but not α, or β. See main text for statistics.

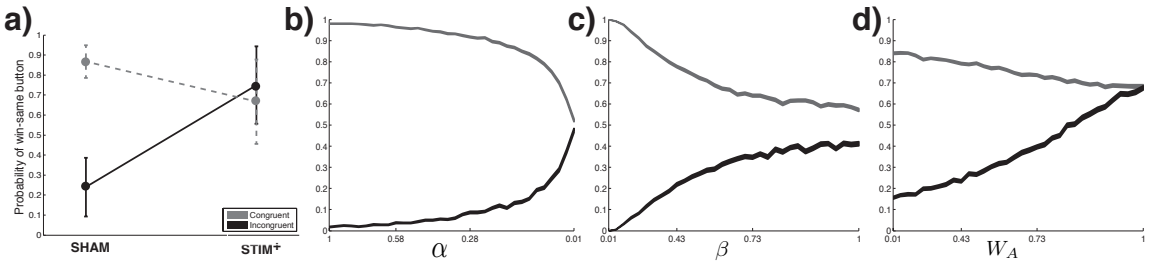

Figure 3.5: A. Subjects who showed stimulation-related increases in win-same button (n = 5) showed asymmetric changes during congruent (grey) and incongruent (black) trials when comparing STIM$^+$ and SHAM trials. *B,C.* Simulated behavior of a $Q-$learning model shows symmetric changes in win-same button during congruent and incongruent trials. *D.* Strengthened action-reward associations in the hybrid-AQ learning model results in asymmetric changes in win-same button.

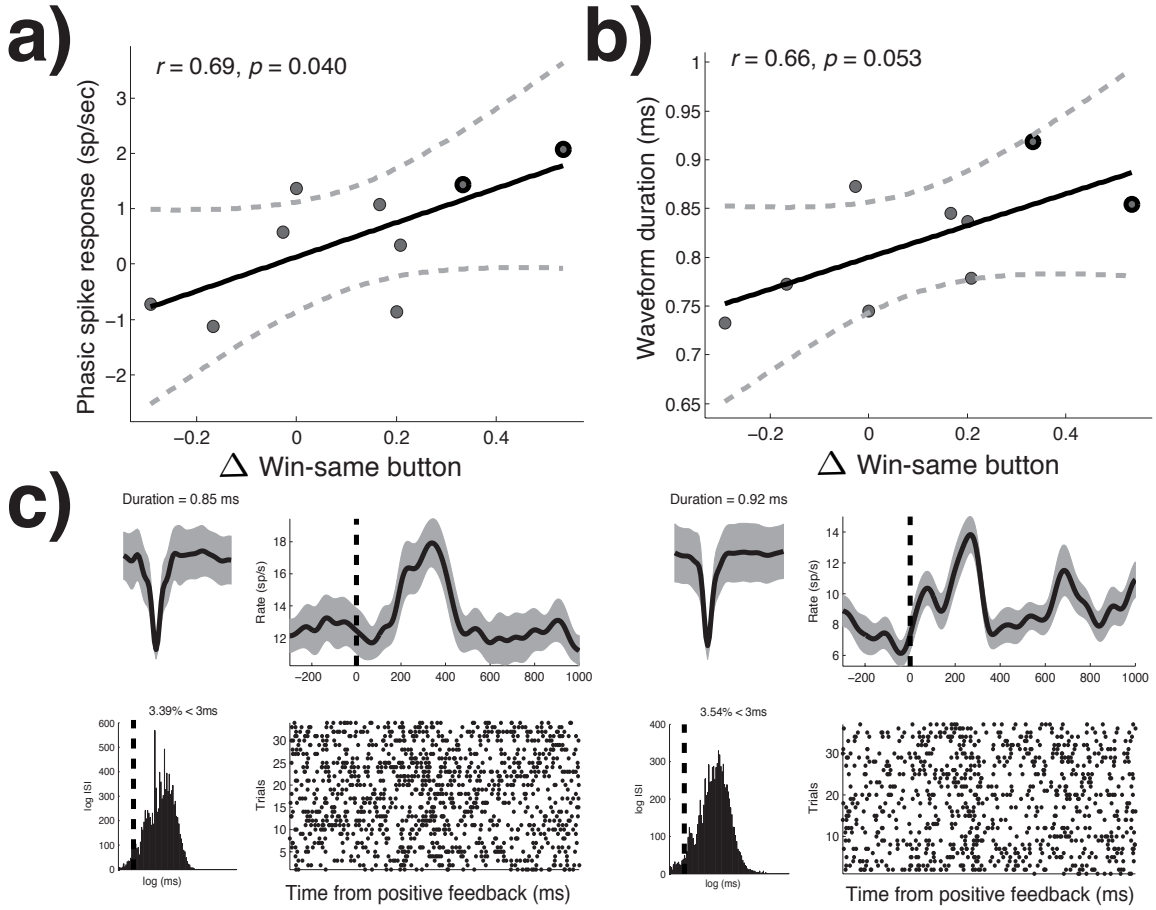

Figure 3.6: Stimulation-related increases in win-same button were positively correlated with post-reward phasic responses (*A.*) and the mean waveform duration (*B.*) of multi-unit activity recorded during Stage 1. Each dot represents a subject, the solid red line is the regression slope, and the dashed lines represent 95 % confidence intervals. 9 of the 11 subjects contributed to this analysis (we were unable to obtain recordings from subject #3, and we did not identify spiking activity from subject #11, see *Materials and Methods*). *C.* Example waveforms and post-reward phasic responses of unit activity from the two subjects who showed the greatest increases in win-same button (outlined in red in panels A and B). For each unit, we show the average waveform (top left, gray shading marks the standard deviation), the inter-spike interval (bottom left, red line marks 3 ms), the average post-reward firing response (top right, smoothed with a Gaussian kernel of half-width = 75 ms; gray shading indicates standard error of mean), and the spike raster following reward trials. Dashed red line indicates reward onset.

Chapter 4

Intracranial high-frequency activity reveals distributed representations of unexpected outcomes during reinforcement learning

Ashwin G. Ramayya and Michael J. Kahana (2014) *In preparation*.

4.1 Abstract

Theories of reinforcement learning suggest that individuals alter their decisions based on unexpected outcomes. Whereas monkey single-unit studies have demonstrated distributed representations of unexpected outcomes in several regions, the extent to which such representations exist in the human brain is not known. Here, we obtained intracranial electroencephalography (iEEG) recordings from the cortex and medial temporal lobe (MTL) of 39 patients undergoing surgical monitoring

for drug-refractory epilepsy as they performed a two-alternative reinforcement learning task. We identified putative outcome valence-encoding contacts based on changes in high-frequency activity (HFA, 70-200 Hz), a known indicator of local firing rates. We related the activity of these putative valence signals to trial-by-trial model-based estimates of reward expectation and identified patterns of activity consistent with unexpected reward and penalty representations, respectively. Unexpected reward representations were frequently observed in right occipito-temporo-prefrontal regions, and the strength of their expectancy-related changes in activity was correlated with subjects' tendency to select the high-probability item during the task. These results demonstrate the existence of distributed unexpected outcome representations in the human brain that are functionally related to learning.

4.2 Introduction

Prominent theories of reinforcement learning posit that individuals alter their decisions based on unexpected rewards and penalties (Rescorla & Wagner, 1972; Sutton & Barto, 1990). Unexpected rewards are thought to strengthen associations between recently active neural populations, and increase the future probability of making a rewarding decision, whereas unexpected penalties are thought to weaken these associations and decrease the future probability of making a penalizing decision (P. R. Montague et al., 1996). Thus, to understand the neural basis of reinforcement learning, it is crucial to characterize the manner in which unexpected outcomes observed during reinforcement learning are neurally represented.

Several human neuroimaging studies have identified regionally-clustered hemodynamic changes that encode unexpected outcome signals in select cortical and

striatal regions (Berns et al., 2001; McClure et al., 2003; O'Doherty et al., 2004; Rutledge et al., 2010). Because hemodynamic changes sample activity from large neural populations (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001), these regionally-clustered representations likely arise due to correlated inputs into the region. For example, the most prominent unexpected reward representations are observed in regions that receive prominent inputs from midbrain dopaminergic neurons (P. Montague, King-Casas, & Cohen, 2006; Kable & Glimcher, 2009; S. Haber & Knutson, 2009), a neural population that has been shown to be functionally important for reward-based learning in animals (Glimcher, 2011) and humans (Zaghloul et al., 2009; Ramayya et al., 2014).

However, recent work suggest that these regionally-clustered signals may only represent a subset of unexpected outcome representations in the human brain. Single-neuron recordings in non-human primates that have demonstrated that several cortical regions demonstrate diverse encoding schemes (Padoa-Schiopa & Assad, 2006; Morrison & Salzman, 2009; Wallis & Kennerley, 2011); some neurons encode unexpected rewards, whereas others may encode unexpected penalties. Information encoded by such heterogeneous populations may not be detected when averaging activity within a region, as typically done in functional neuroimaging studies, but may be evident when using methods that are sensitive to diverse changes within a region. A recent multi-voxel pattern analysis of neuroimaging data demonstrated that it is possible to decode outcome valence information from almost all cortical and subcortical structures, most of which were not known to encode outcome valence based on prior univariate neuroimaging studies (Vickery et al., 2011).

Thus, it is now known that information about outcome valence is widely represented throughout the human brain, however, the extent to which these

widespread valence representations represent unexpected outcome representations is not known. In this study, we obtained intracranial electroencephalography (iEEG) recordings from the cortex and medial temporal lobe (MTL) of 39 patients with drug-refractory epilepsy as they performed a two-alternative probability learning task. We studied changes in high-frequency activity (HFA; 70-200 Hz) at individual electrode contacts, an iEEG feature that has been shown to be correlated with the average firing activity of local neurons (Manning et al., 2009; Ray & Maunsell, 2011). These changes provides a spatio-temporally precise representation of local neuronal activity and may allow for the detection of diverse changes within a region (Bouchard, Mesgarani, Johnson, & Chang, 2013). We sought to identify putative valence signals distributed across the brain and relate their activity to reward expectation, so as to shed light on their functional relevance for learning.

4.3 Methods

Subjects. Patients with drug-refractory epilepsy underwent a surgical procedure in which grid, strip, and depth electrodes were implanted so as to localize epileptogenic regions. Data were collected from Thomas Jefferson University Hospital (TJUH) and the Hospital of University of Pennsylvania (HUP) in collaboration with the neurology and neurosurgery departments at each institution. Our research protocol was approved by the Institutional Review Board at each hospital and informed consent was obtained from the participants and their guardians. In total, we recorded neural activity from 39 subjects (12 female, 7 left-handed, mean age 37 years).

Reinforcement learning task. Subjects performed a two-alternative probability learning task which has been previously used to study reinforcement learning and value-based decision making (Figure 4.1; (M. J. Frank et al., 2004; M. Frank et al., 2007; Zaghloul et al., 2012)). During the task, subjects selected between pairs of Japanese characters ("items") and received positive or negative feedback following each choice. Subjects were informed that one item in each pair carried a high probability of positive feedback than the other item pair, and that their goal was to select items that maximized their probability of obtaining positive feedback. On a given trial, the items were simultaneously presented on the screen; one on the left side and one on the right side of the screen. They were presented on a dark grey background in white font. The items remained on the screen until subjects made a response by pressing the left "SHIFT" button on a keyboard (which selected the item on the left or right side of the screen, respectively). Once a response was registered by the computer, the selected item was highlighted in blue, and feedback was immediately provided. In the event of positive feedback, the selection screen turned green, and an audible ring of a cash register was presented. In the event of negative feedback, the selection screen turned red, and an error tone was presented. The screen remained colored for 2 seconds. There was a 0-400 ms jitter between successive trials. Items were randomly arranged on the left or right side of the screen from trial to trial.

During a session, subjects were presented with up to three novel pairs to encourage learning throughout the session. Distinct item pairs were presented in a randomly interleaved manner; each item pair carried a distinct relative reward rate (80/20, 70/30, or 60/40). Reward rates associated with each item were determined randomly prior to each session and fixed throughout the experiment. Each session began with the exclusive presentation of a single item pair (random selection of

a relative reward rate). If participants met a minimum performance criteria on the given item pair over a block of 10 trials (i.e., accuracy >= 60% for 80/20 or 70/30 pairs, or >= 50% for the 60/40 pair), a second item pair was introduced and randomly interleaved along with the first item pair. A third item pair was only introduced in subjects that met the performance criteria on the two item pairs already introduced. Participants performed a total of 107 sessions (each subject performed an average of 2.82 sessions), with an average of 130 trials per session.

iEEG recordings. Clinical circumstances alone determined electrode number and placement. Subdural (grids and strips) and depth contacts were spaced 10 mm and 8 mm apart, respectively. iEEG was recorded using a Nihon-Kohden (TJUH) or Nicolet (HUP) EEG system. Based on the amplifier and the discretion of the clinical team, signals were sampled at either 512, 1024, or 2000 Hz. Signals were converted to a bipolar montage by differencing the signals between each pair of immediately adjacent electrodes on grid, strip, or depth electrodes; the resulting bipolar signals were treated as new virtual electrodes (henceforth referred to as "contacts" throughout the text), originating from the midpoint between each electrode pair (Burke et al., 2013). Analog pulses synchronized the electrophysiological recordings with behavioral events.

Extracting high-frequency activity from iEEG recordings We convolved clips of iEEG (1000 ms before feedback onset to 2000 ms after onset, plus a 1000 ms flanking buffer) with 30 complex valued Morlet wavelets (wave number 7) with center frequencies logarithmically spaced from 70 to 200 Hz (Addison, 2002). We first squared and then log-transformed the wavelet convolutions, resulting in a continuous representation of log-power surrounding each feedback presentation.

We averaged these log-power traces in 200 ms epochs with 190 ms overlap surrounding feedback presentation (-1000-2000 ms), yielding 281 total time intervals surrounding feedback presentation. To identify high-frequency activity (HFA), we averaged power across all frequencies (ranging from 70 to 200 Hz). We z-transformed HFA power values within each session by the mean and standard deviation of task-related HFA recorded from that session (0-500 ms post-stimulus, -750-0 ms pre-choice, and 0-2000 ms post-feedback). We henceforth refer to z-transformed HFA values as "HFA".

Assessing HFA differences between positive and negative outcomes. For each contact, we identified temporally-contiguous HFA differences between positive and negative feedback by performing a cluster-based permutation procedure that accounts for multiple comparisons (Maris & Oostenveld, 2007). As suggested by Maris and Oostenveld (2007), we began by performing an unpaired t-test at each time interval comparing HFA distributions associated with all positive and negative feedback trials performed by the subject. Using an uncorrected $p = 0.05$ as a threshold, we identified the largest cluster of temporally adjacent windows that showed positive t-statistics (greater HFA following positive compared to negative outcomes), and the largest cluster of temporally adjacent windows that showed negative t-statistics (greater HFA following negative compared to positive outcomes). By taking the sum within each of these clusters, we computed a positive and negative "cluster statistic", respectively. To assign significance to each of these cluster statistics, we generated a null distribution of cluster statistics based on 1000 iterations of shuffled data (on each iteration, positive and negative feedback labels were randomly assigned to HFA traces from each trial). Based on where each cluster-statistic fell on the null distribution, we generated a one-tailed $p-$value for

each effect. We considered an effect to be significant if it was associated with a cluster-based p-value < 0.05, thus, the false-positive rate of obtaining each effect at 5%.

Assessing the frequency of a particular effect across subjects To assess whether a particular effect more frequently observed by chance across subjects, we performed the following procedure ("counts t-test"). In each subject, we counted the number of significant contacts that we observed ("true counts"), and generated a binomial distribution of counts values expected by chance ("null counts distribution"), based on the number of available contacts in that subject and the false-positive rate associated with the test. We obtained a z-scored counts value in each subject by comparing the true counts value to the null counts distribution. We then assessed whether distribution of z-scores across subjects deviated from zero via a one-sample paired t-test; positive t-statistics suggest that the effect was more frequently observed than chance, and negative t-statistics suggest that the effect was less frequently observed by chance. When comparing the frequencies of two-effects across subjects (e.g., reward and penalty effects), we performed a paired counts t-test in the following manner. Within each subject, we obtained z-scored counts values for reward and penalty effects based on the null counts distribution as described earlier, and compared the distributions of reward- and penalty-related z-values across subjects (via paired-t-test). Positive z-values indicate that reward effects occurred more frequently than penalty effects, whereas negative values indicate that penalty effects were more frequently than reward effects. We corrected for multiple comparisons using a false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995).

Electrode Localization. Surface electrodes (strips and depths) were manually identified on subject's post-operative CT scans and transformed to a common cortical surface representation to allow for comparisons across subjects. We employed FreeSurfer's software routines (Dale, Fischl, & Sereno, 1999) to generate a cortical surface representation that was representative of our patient population–individuals undergoing intracranial EEG monitoring for drug-refractory epilepsy. We did this by generating cortical surface reconstructions for a large group of patients who volunteered to participate in our research studies (n = 62). This group included subjects who participated in the current study and those who participated in previous studies conducted by our group (e.g., (Burke et al., 2013)), and for whom a pre-operative MRI was available from which a cortical surface could be modeled. We aggregated these surfaces to generate an average cortical surface representation, that was co-registered to the MNI152 brain (Fischl, Sereno, Tootell, & Dale, 1999). Each point on this surface representation was automatically assigned an anatomical label based on a manually-labeled anatomical atlas (Desikan et al., 2006). To map electrode coordinates from the CT scan onto the cortical surface, we registered each post-operative CT scan to the average cortical surface using a rigid-body 6 degrees-of-freedom affine transformation algorithm, and manually adjusted each transform such that electrodes were positioned as close to the cortical surface as possible. Finally, electrodes were "snapped" to the cortical surface by moving each coordinates to the nearest point on the gyral surface (the maximum deviation allowed was 20 mm). We assigned an anatomical label to each bipolar pair of electrodes based on the location on the cortical surface that was closest to the midpoint between the two contacts. Depth electrodes were manually localized by a neuroradiologist using a post-operative MRI scan. To visualize these depth contacts in a common anatomical space, we transformed them to MNI-coordinates

using the same CT-to-average surface transformation described above, however, we did not snap the electrodes to the cortical surface. Depth contacts were visualized on a MNI-brain slice generated using the WFU pick atlas toolbox (Maldjian, Laurienti, Kraft, & Burdette, 2003).

Estimating reward expectation To obtain trial-by-trial estimates of reward expectation, we fit a standard reinforcement learning model to subjects behavioral data. Because our goal was to model choice behavior during learning, we only considered behavioral data from item pairs where subjects demonstrated evidence of learning (> 70% accuracy on last 10 trials, and > 50% accuracy overall). The Q-model maintains independent estimates of reward expectation (Q) values for each option i at each time t (Sutton & Barto, 1990). A choice is probabilistically generated on each trial by comparing the Q values of available options on that trial using the following logistic function: $P_i(t) = \frac{\exp(Q_i(t)/\beta)}{\sum_j \exp(Q_j(t)/\beta)}$. β is a free parameter for inverse gain in the softmax logistic function (which accommodates noise in the choice process or different relative tendencies for exploration vs. exploitation; (Daw et al., 2006)). Once an item is selected by the model, feedback is received, and Q values are updated using the following learning rule: $Q_i(t+1) = Q_i(t) + \alpha[R(t) - Q_i(t)]$, where $R(t) = 1$ for correct feedback, $R(t) = 0$ for incorrect feedback and α is the learning rate parameter that adjusts the manner in which previous reinforcements influence current Q values. Large α values (upper bound = 1) heavily weight recent outcomes when estimating Q, whereas small α values (lower bound = 0) incorporate reinforcements from many previous trials. We identified the best-fitting parameters for each subject by performing a grid-search through the two dimensional parameter space (α, learning rate, and β, noise in the choice policy, 0.01 to 1, with a step size of 0.1) and selected the set of parameters that minimized the mean squared error be-

tween the model's predictions of subject's choices ($i*$), and subjects' actual choices. To quantify the model's goodness-of-fit, we compared each subject's mean squared error value to a null distribution of mean squared errors generated for that subject's data based on a random guessing model ($P = 0.5$ for all choices, 10000 iterations). Based on this comparison, we obtained a p-value describing the false-positive rate associated with the observed mean squared error for that subject. In all subjects, the best-fitting parameters provided a better prediction of subjects choice behavior than the random guessing model (FDR-corrected p's < 0.001). We describe mean best-fitting parameters, goodness-of-fit data in Table 4.1.

4.4 Results

Behavioral results. 39 subjects selected between pairs of Japanese characters ("items") and received positive or negative feedback following each choice (Figure 4.1a). Subjects were informed that one item in each pair carried a higher reward probability than the other, and that their goal was to maximize their probability of obtaining positive feedback. During each session, subjects were presented with multiple item pairs in an interleaved manner, with each item pair carrying distinct relative reward rates (see *Materials and Methods*). We found that subjects demonstrated a tendency towards choosing the high-probability item during the last 10 trials of an item pair ($t(38) = 7.24, p < 0.001$) that was greater than that observed during the first 10 trials of an item pair ($t(38) = 5.11, p < 0.001$; Figure 4.1b). These data suggest that subjects demonstrated learning during the task.

To assess the importance of rewards and penalties for learning, we studied subjects' choice behavior following rewards and penalties during the first 10 trials. To index learning from rewards and penalties, we studied the frequency

that subjects repeated the same choice following rewards ("win-stay") and the frequency that subjects altered their decision following penalties ("lose-switch"). Subjects demonstrated a mean win-stay of 0.75, that was more frequent than chance ($t(38) = 7.89, p < 0.001$), but demonstrated a mean lose-switch of 0.54, that did not deviate from chance ($p > 0.2$). We tested whether individual differences in performance (overall frequency of choosing the high-probability item, "accuracy") were dependent on win-stay or lose-switch during the first 10 trials using a linear regression model. We observed a positive relation between accuracy and win-stay ($t(38) = 3.30, p = 0.001$), but did not observe a significant relation between accuracy and lose-switch ($p = 0.18$). These results suggest that subjects' learning during the task was predominantly driven by choice behavior following rewards.

Identifying putative outcome valence signals Theories of reinforcement learning posit that individuals alter their decisions based on unexpected rewards and penalties (Rescorla & Wagner, 1972; Sutton & Barto, 1990). To characterize the neural representations of these cognitive signals, we first identified neural populations that demonstrated distinct activity following positive and negative outcomes. We refer to these signals as "putative valence" because they may reflect neural populations that encode differences in outcome valence, but could also be driven by low-level sensory features, or salience, factors that we did not explicitly control in the experiment. We obtained intracranial electroencephalograpy (iEEG) recordings from 4,266 surface and depth electrode located in throughout the cortex and MTL (Figure 4.1c). We focused our analyses on high-frequency activity (HFA; 70-200 Hz), an iEEG feature that has been correlated with local neural firing rates (Manning et al., 2009; Ray & Maunsell, 2011), and thereby provides a spatio-temporally precise measure of local neuronal activity (Buzsaki, Anastassiou, & Koch, 2012; Burke et

al., 2014). Rather than averaging activity within regions of interest, we studied HFA changes at individual electrode contacts so as to extract information from regions that may demonstrate heterogeneous representations of outcome valence and reward expectation.

We identified contacts that showed significant temporally-contiguous HFA differences between positive and negative feedback (cluster-based permutation procedure; *Materials and Methods*).. The false-positive rate associated with identifying a significant effect at a particular contact was set to 5%. We found that 2,150 contacts (50%) demonstrated HFA differences between positive and negative outcomes; 874 contacts (20%) showed positive effects (relatively greater HFA following positive feedback) and 1,031 contacts (24%) showed negative effects (relatively greater HFA following negative feedback, Figure 4.4a). We also observed a small subset of contacts (n = 245) that demonstrated both positive and negative effects during distinct time intervals. To assess whether a particular effect was more frequently observed across subjects than expected by chance, we performed an across-subject t-test on z-transformed counts values ("counts t-test," *Materials and Methods*). Across subjects, we observed positive and negative contacts at above-chance frequencies ($t(38) > 8.94, p < 0.001$, each associated with a false-positive rate of 5%). We focus the remainder of our analyses on contacts that exclusively showed a positive or a negative effect (henceforth, "valence-encoding contacts"). Consistent with recent neuroimaging studies (Vickery et al., 2011), we found that valence-encoding contacts were widely distributed and generally interspersed throughout the cortex and MTL (see *Supplemental Information*).

Relating putative valence signals to reward expectation To assess the functional relevance of these putative valence signals for learning, we studied the relation

between HFA and reward expectation during time interval that we observed significant differences between positive and negative feedback (based on our cluster-based permutation procedure, *Materials and Methods*). Because our goal was to study neural processes related to learning, we only considered neural and behavioral data from item pairs where subjects demonstrated evidence of learning (> 70% accuracy on last 10 trials, and > 50% accuracy overall). 1,345 valence-encoding contacts (from 26 subjects) were recorded during trials that that met this criteria. To obtain trial-by-trial estimates of reward expectation, we fit a standard-reinforcement learning model to each subjects' behavioral data ((Sutton & Barto, 1990; Glimcher, 2011); *Materials and Methods*; Table 4.1). Because distinct item pairs were presented in an interleaved manner, reward expectation estimates were dissociated from time during the task (Figure 4.2a). We studied the relation between HFA and reward expectation, separately following positive and negative feedback, using the following regression model. $Y = \beta_0 + \beta_Q Q + \beta_t T$, where Y is a vector containing HFA values, Q is a vector containing expectation values. T tracked number of times a given item pair had been previously presented so as to account for any novelty-related changes in HFA. We considered a contact to show an expectation-related effect if there was a significant β_Q coefficient (t-statistic, $p < 0.05$) associated with HFA following positive or negative feedback. Several example contacts that showed expecation-related changes in activity are shown in Figure 4.2b.

α	β	mean sq. error	mean sq. error (null)
0.20 (± 0.04)	0.23 (± 0.04)	0.14 (± 0.01)	0.26 (± 0.01)

Table 4.1: Summary of Q model fits. Mean (± s.e.m across subjects) shown for best-fitting parameter values and goodness-of-fit measures (see emphMaterials and Methods).

[Table 1 about here.]

If unexpected outcome representations are prominent in the human brain, as suggested by theoretical studies, one would expect to observe opposing relations between HFA and reward expectation following positive and negative feedback. Following positive feedback, HFA should demonstrate a negative relation with reward expectation, indicating that post-reward HFA is greater when reward expectation is low (unexpected rewards), compared to when reward expectation is high (expected rewards). In contrast, following negative feedback, HFA should show a positive relation with reward expectation, indicating that post-penalty HFA is greater when reward expectation is high (unexpected penalties), compared to when it is low (expected penalties). Consistent with these predictions, we observed two expectation-related patterns of activity more frequently than expected by chance (counts t-test, FDR-corrected $p < 0.05$, Figure 4.3a); contacts that demonstrated negative β_Q values following positive feedback (17%; $t(25) = 3.65, p = 0.001$), and contacts that demonstrated positive β_Q values following negative feedback (9.6%; $t(25) = 3.39, p = 0.002$). We refer to these groups of contacts as "UR" and "UP" contacts, because they encoded unexpected rewards and penalties, respectively. We observed little overlap between these groups of contacts as only 1.7% of valence-encoding contacts demonstrated both UR and UP activity. We include these contacts in both categories (our main results were unchanged when considering contacts that exclusively encoded UR and UP; data not shown).

UR and UP contacts may represent neural signals that guide learning following rewards and penalties, respectively. Because subjects' behavioral data suggested that learning was predominantly related to choice behavior following rewards, one might expect that the strength of expectation-related changes in UR contacts was related to subjects' performance during the task. To measure the strength of UR representations in each subject, we averaged the t-statistics associated with β_Q

during positive feedback among all UR contacts recorded from that subject. Across subjects, we observed a significant correlation between accuracy and the strength of UR contacts ($r = 0.56, p = 0.006$, Figure 4.3b), suggesting that UR representations were functionally related to learning. However, we did not observe such a relation between accuracy and the strength of UP representations ($p > 0.5$). These results are consistent with behavioral results suggesting that individual differences in performance were related to subjects' choice behavior following rewards, but not penalties.

Based on previous studies in animals, one might expect to observe unexpected outcome representations to be regionally distributed throughout the human cortex. To assess whether this was the case, we studied the proportion of valence-encoding contacts that demonstrated UR and UP responses in several ROIs (Figure 4.3c). We only included regions where we identified valence-encoding contacts from at least 5 subjects. We found that both UR and UP contacts were distributed across several regions. We observed UR contacts more frequently than expected by chance in a group of right hemisphere regions including occipital, fusiform, temporal, and vlPFC (t's$> 3.41, p$'s< 0.007, FDR-corrected p's< 0.05). We observed trends towards frequently observing UP contacts in the right sensorimotor, parietal and temporal regions (t's$> 1.81, p$'s< 0.1). Thus, both UR and UP contacts were regionally-distributed throughout several regions and frequently observed in the right hemisphere.

UR and UP contacts may reflect activity from neural populations that predominantly encode positive and negative reward prediction errors, that signal outcomes that are better or worse than expected, respectively (Rescorla & Wagner, 1972). If this is the case, then one might expect UR contacts to demonstrate greater overall activity following rewards compared to penalties, and UP contacts to demonstrate

greater activity following penalties compared to rewards. To assess whether this is the case, we studied the frequency of positive and negative valence effects among UR and UP contacts. We found that the majority of UR (72%) and UP (72%) contacts demonstrated negative valence effects, where activity was greater following penalties compared to rewards. Negative valence effects were generally more frequent among UP and UR contacts than among valence-encoding contacts that did not show UP or UR effects (54%; counts t-test, $t(23) = 2.08, p = 0.049$). Thus, unexpected outcome representations typically showed greater overall activity for penalties compared to rewards.

4.5 Discussion

During reinforcement learning, it is thought that individuals alter their decisions based on unexpected outcomes (Rescorla & Wagner, 1972; Sutton & Barto, 1990). We wanted to study the manner unexpected outcomes obtained during reinforcement learning are represented the human brain. Whereas prior single-unit studies in monkeys suggest that unexpected outcome representations may be distributed in several regions (Wallis & Kennerley, 2011), human functional imaging studies have typically averaged activity within brain regions to identify regionally-clustered representations of unexpected outcomes (Berns et al., 2001; McClure et al., 2003; O'Doherty et al., 2004; P. Montague et al., 2006). To bridge the gap between these previous findings, we wanted to assess whether there exist regionally-distributed representations of unexpected outcomes in the human brain.

We obtained iEEG recordings from 39 patients with drug-refractory epilepsy as they performed a two-alternative probability learning task. We studied changes in HFA, an iEEG feature that provides a spatio-temporally precise representation of

local neuronal activity (Manning et al., 2009; Ray & Maunsell, 2011). Rather than averaging activity within regions of interest, we studied HFA changes at individual electrode contacts so as to extract information from regions that may demonstrate heterogeneous representations of outcome valence and reward expectation. Previous studies have shown that HFA at nearby electrode contacts may demonstrate heterogeneous patterns of activity and may represent information beyond that represented by the average activity within a region (Bouchard et al., 2013). We found that electrode contacts distributed throughout the cortex and medial temporal lobe demonstrated reliable differences between positive and negative outcomes (*Supplemental Information*). These results are more consistent with recent multi-voxel-pattern-analyses of functional neuroimaging data that have demonstrated ubiquitous coding of outcome valence throughout the cortex and MTL (Vickery et al., 2011), rather than traditional functional neuroimaging studies that average activity within nearby regions (Bartra et al., 2013). We refer to these signals as "putative valence" because they may largely reflect neural populations that encode differences in outcome valence, but could also be driven by low-level sensory features, or salience, factors that we did not explicitly control in the experiment.

Our main goal was to assess the prevalence of unexpected outcome representations among these putative valence-encoding contacts. We assessed the relation of HFA recorded from each putative valence-encoding contact to trial-by-trial estimates of reward expectation obtained from by a reinforcement learning model to subjects' choice behavior. As predicted by prior theoretical work (Sutton & Barto, 1990), we found that the most prevalent patterns of activity were consistent with representations of unexpected rewards (UR) and unexpected penalties (UP), respectively. These signals may reflect neural processes that guide learning from rewards and penalties, respectively. We found that the strength of UR representations was

correlated with subjects' performance during the task, which is consistent with our finding that subjects' performance on the task is mainly related to their ability to learn from rewards. We did not observe a correlation between the strength of UP representations and performance, however, this may reflect the fact that subjects' choice behavior following penalties was not related to performance during the task. One possibility is that subjects directly rely on these signal to encode unexpected rewards, and that subjects' ability to learn from rewards improves with the strength of this signal. Alternatively, it may be the case that subjects' performance on the task results in increased signal strength, thus making the neural signal easier to detect. Then, the fidelity of the error signal would be driven by subjects' performance on the task, rather than the other way around. Future studies that apply electrical microstimulation in a clinical setting to a particular valence-encoding signal may be needed to resolve this issue (Ramayya et al., 2014).

We found that UR and UP contacts were distributed across several regions, many of which were not previously identified by neuroimaging studies (Berns et al., 2001; McClure et al., 2003; O'Doherty et al., 2004; P. Montague et al., 2006; Rutledge et al., 2010). These results suggest that unexpected outcome representations are encoded by neural populations that are widely distributed throughout the brain. We found that UR and UP signals were typically observed in distinct electrode contacts, suggesting that these signals were typically encoded by distinct neural populations. One possibility is that UR and UP signals are generated as a result of inputs from low-level neurotransmitter systems that project widely throughout the brain. For example, dopamine and serotonin systems that have been previously implicated in reward and penalty-based learning, respectively (Schultz et al., 1997; Daw et al., 2002). The regionally-segregated cortical projections of such neurotransmitter systems may explain the segregation of UP and UR

representations.

Moreover, unexpected outcome signals were frequently observed in the right hemisphere; UR representations were typically observed in a distributed set of right-hemisphere regions, including occipital, fusiform, temporal, and ventrolateral prefrontal regions. These regions are typically engaged by emotionally salient visual stimuli, that are often associated with negative valence, are typically associated with activation in the amygdala (Morris, Ohman, & Dolan, 1999; Adolphs, 2002; Duncan & Barrett, 2007; Hunga et al., 2010). Consistent with these findings, valence-encoding contacts in these regions typically demonstrated greater activity following penalties compared to rewards (*Supplemental Information*). How might one explain the presence of unexpected reward signals in these regions? One possibility is that these neural populations multiplex multiple feedback signals, a positive RPE that signals unexpected rewards and a negative valence signal that encodes incoming penalties. Alternatively, these contacts may represent an idiosyncratic salience representation, whereby negative outcomes are most salient, regardless of their associated expectation, and the salience of positive outcomes decreases as they become more expected. Previous findings have shown that the amygdala encodes unsigned prediction errors, that signal the surprise associated with incoming feedback (Roesch, Esber, Li, Daw, & Schoenbaum, 2012). Such signals may guide learning by enhancing learning rates following surprising feedback trials (Pearce & Hall, 1980; Behrens, Woolrich, Walton, & Rushworth, 2007; Roesch et al., 2012; Nassar et al., 2012). Neural signals with finer spatial resolution may be needed to investigate the origin of unexpected reward signals in these regions.

Conclusions In conclusion, we found that reward and penalty representations were both widely represented in the cortex and MTL. Regionally-distributed sub-

sets of these representations were modulated by reward expectation in a manner consistent with unexpected rewards and penalties, respectively. Unexpected reward representations were prominently observed in right occipito-temporo-prefrontal regions and were correlated with subjects' performance during the task, suggesting a functional relevance for learning. These results demonstrate that unexpected outcomes are encoded by regionally-distributed neural populations during human reinforcement learning. Future studies should investigate the emergence of these signals, and study the manner in which they alter subsequent decisions.

4.6 Supplemental Data

Spatio-temporal properties of putative outcome valence signals First, we characterized the spatio-temporal properties of reward and penalty signals throughout the cortex and MTL. We registered electrode contacts from each subject to a common anatomical space (*Materials and Methods*), and assessed whether they were more frequently observed than chance in various regions of interest (ROI). We only studied ROIs for which we recorded neural data from at least 5 subjects (Table**??**; Figure4.4a). In 15 of the 21 ROIs that met this criteria, we found that subjects showed both reward and penalty contacts more frequently than expected by chance (counts t-test, FDR-corrected p's< 0.05). In 4 of the remaining 6 ROIs, subjects either showed reward and penalty contacts at above-chance levels (counts t-test, FDR-corrected p's< 0.05). When we directly compared the frequency of reward and penalty contacts in the various ROIs (paired counts t-test, see *Materials and Methods*), we observed a bias towards reward contacts in a distributed set of left-hemisphere regions (MTL, OFC, and parietal regions), and a bias to-

wards negative contacts in a distributed set of right-hemisphere regions (occipital and dlPFC; counts t-test, FDR-corrected p's< 0.05). Thus, we found that valence-encoding contacts were widely distributed and generally interspersed throughout the cortex and MTL, but also observed regional biases towards reward and penalty representations in the left and right hemisphere, respectively.

Even if positive and negative outcome representations are both present in a particular ROI, it may be the case that they are locally clustered within that ROI. To assess whether this was the case, we performed a cortical surface searchlight analysis by assessing whether reward and penalty contacts were more frequently observed in 12.5 mm spheres centered at each vertex of the cortical surface (Figure 4.4b). We considered all spheres that contained electrodes from at least 5 subjects, and classified a region as showing a significant reward or penalty effect based on a counts t-test (uncorrected $p < 0.05$). We applied less conservative statistical criteria than the previous anatomical analysis to fully examine the regional patterns of reward and penalty outcome representations. We found that reward and penalty representations were interspersed in several lateral temporal, parietal and lateral prefrontal regions. However, we observed segregated reward and penalty signals in surrounding regions, including medial prefrontal and medial temporal surface. In the latter two regions, we generally observed positive outcome representations in anterior regions (e.g., frontal pole, OFC; hippocampus, entorhinal cortex), and negative outcome representations in posterior regions (eg., posterior superior frontal gyrus, paracentral lobule; posterior fusiform). Thus, we observed overlapping reward and penalty representations bilaterally in a group of lateral temporo-parieto-frontal regions, but observed segregated representations in surrounding regions.

We next assessed whether there were regional differences in timing among re-

ward and penalty signals distributed throughout the brain. For each reward and penalty contact, we studied the time during which we observe peak outcome-related HFA differences. We compared the peak difference-times of reward (and penalty) contacts in each region that they were frequently observed to the mean peak-time observed across all reward (and penalty) contacts (one-sample t-test, Figure 4.5). We did not observe any timing differences that survived multiple comparisons-correction (FDR-corrected p's > 0.2), but observed several trends towards significance (uncorrected p's< 0.1). We observed relatively early peak-times among reward contacts in the left sensorimotor and fusiform regions, and among penalty contacts in the left dlPFC. We observed relatively late peak-times among penalty contacts in the left sensorimotor and temporal regions. Thus, apart from these weak regional differences in timing, we found that reward and penalty contacts generally showed similar temporal dynamics across various brain regions.

We found that reward and penalty signals were both frequently observed in most regions of interest, which suggest that information about outcome valence is widely represented throughout the cortex and MTL. These results are consistent with recent multi-voxel-pattern-analyses of human neuroimaging data demonstrating that outcome valence information can be decoded from almost all cortical and subcortical structures (Vickery et al., 2011). Our results build on this line of work by characterizing the spatio-temporal properties of reward and penalty signaling in the cortex and MTL. We found that reward and penalty signals were interspersed in lateral temporo-patieto-frontal regions, but locally segregated in surrounding regions. We observed several regional biases towards reward and penalty representations. We found that reward contacts were more prevalent several left hemisphere regions (orbitofrontal cortex, medial temporal lobe and parietal lobe), whereas penalty contacts were more prevalent in several right hemisphere

regions (dorsolateral prefrontal and occipital cortex). These results suggest that representations of incoming rewards and penalties are widely distributed throughout the cortex and MTL, but locally segregated in several regions. We observed few regional differences in timing between reward and penalty signals distributed throughout the brain. Although it is difficult to interpret a negative effect, the observed temporal dynamics suggesting that reward and penalty representations do not evolve as a cascade from low-level posterior sensory cortices to higher-order prefrontal cortices, but rather emerge during similar time intervals throughout the brain. These results are consistent with neurobiological models positing that feedback signals are simultaneously transmitted throughout the brain via widespread projections from deep structures (P. R. Montague et al., 1996; Glimcher, 2011).

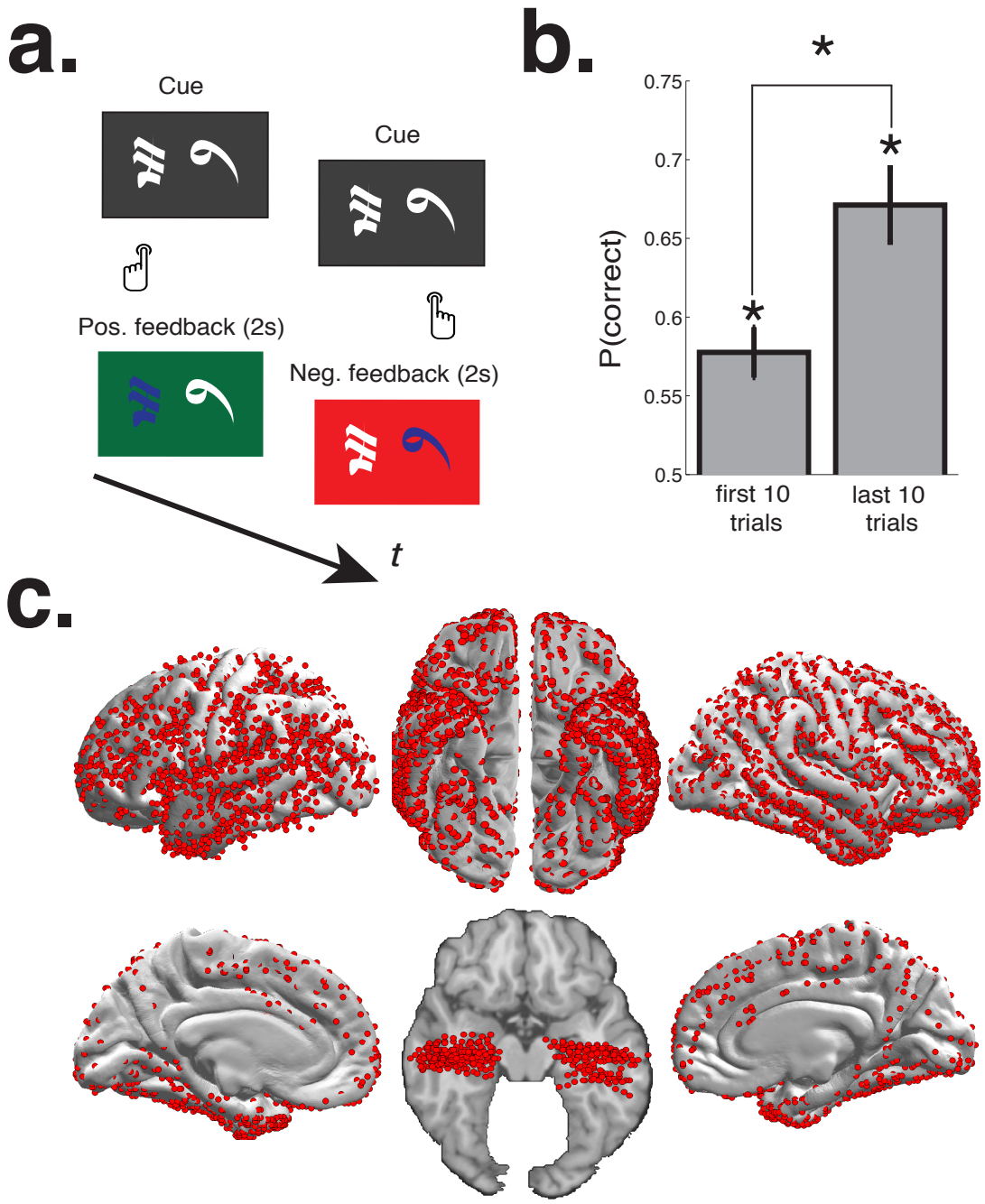

Figure 4.1: **a.** Subjects selected between pairs of Japanese characters on a computer screen and probabilistically received positive or negative audio-visual feedback following each choice. **b.** Average tendency towards selecting the high-probability item during the first and last 10 trials of each item pair. Error bars represent s.e.m across subjects. **c.** iEEG electrodes from each subject were localized to a common anatomical space (see *Materials and Methods*). We show strip and grid contacts on the cortical surface, and depth electrodes targeting the medial temporal lobe on the axial slice. On rare occasions, depth electrodes were placed in the frontal and parietal lobes to supplement surface recordings (not shown).

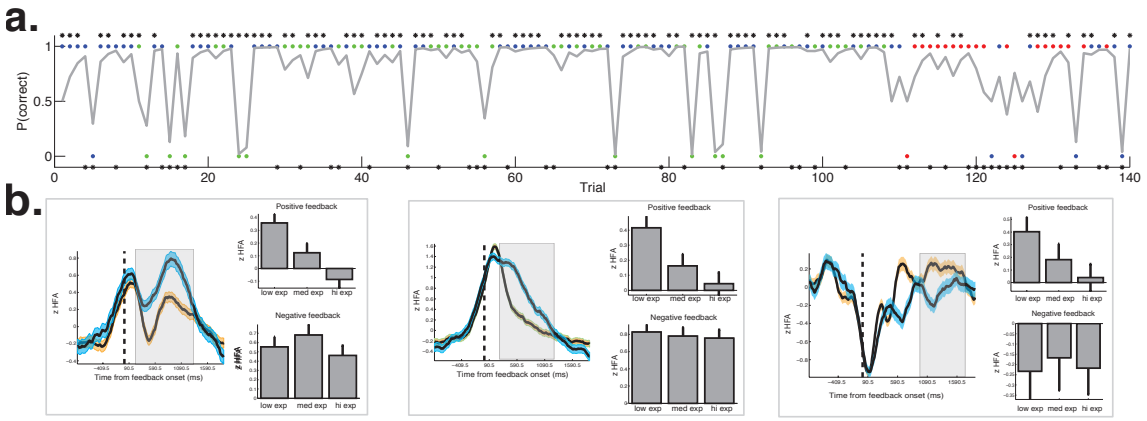

Figure 4.2: a. Behavioral data from one example session. On the top of the figure, dots indicate when the subject chose the high-probability item. Color of the dots indicate the item pair that was presented (blue - 80/20, green - 70/30, red - 60/40). Asterisks indicates when positive feedback was provided following each choice. Bottom of the figure, dots indicate when the subject chose the low-probability item (color-scheme same as the top), whereas asterisks indicate when negative feedback was provided following each choice. Grey line indicates model-predictions of subjects' choices. **b.** Three example contacts recorded from this subject that showed expectation-related changes in activity. Shaded box indicates the time during which we observed significant HFA differences between positive (orange) and negative (blue) outcomes. During this time interval, we studied post-reward and post-penalty changes in HFA during varying degrees of reward expectation.

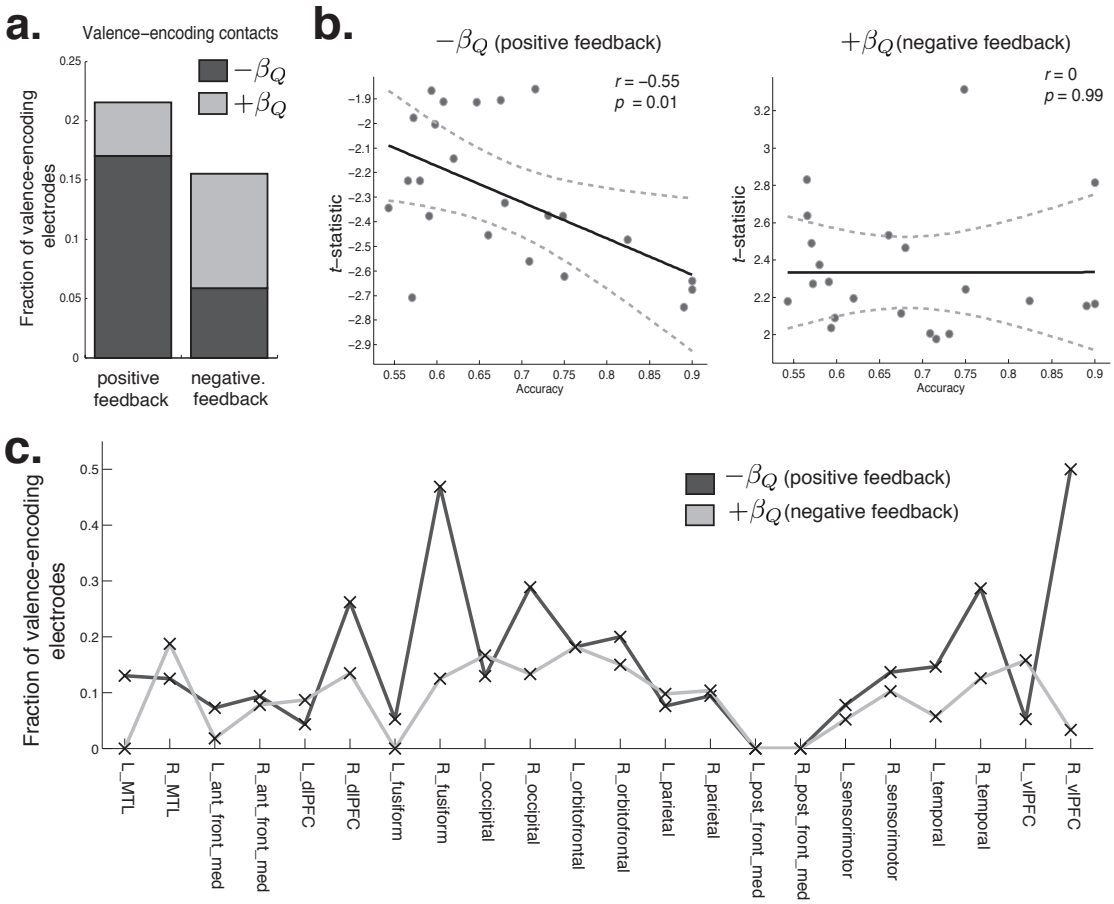

Figure 4.3: a. Fraction of valence-encoding contacts that demonstrated significant relations with reward expectation ($p < 0.05$). Following positive feedback, we observed $-\beta_Q$ and following negative feedback, $+\beta_Q$ more frequently than expected by chance. We refer to these patterns as "unexpected reward" and "unexpected penalty" contacts, respectively. See main text for statistics. **b.** Correlating the strength of expectation-related changes with subjects' performance. **c.** Anatomical distribution of unexpected reward and penalty contacts. In several ROIs, we show the fraction of valence-encoding contacts that showed unexpected reward and penalty signals (dark grey, light grey, respectively). We only included regions from which we observed valence-encoding contacts from at least five subjects. **d.** Percentage unexpected reward contacts and unexpected penalty contacts that demonstrated greater overall activity for rewards and penalties.

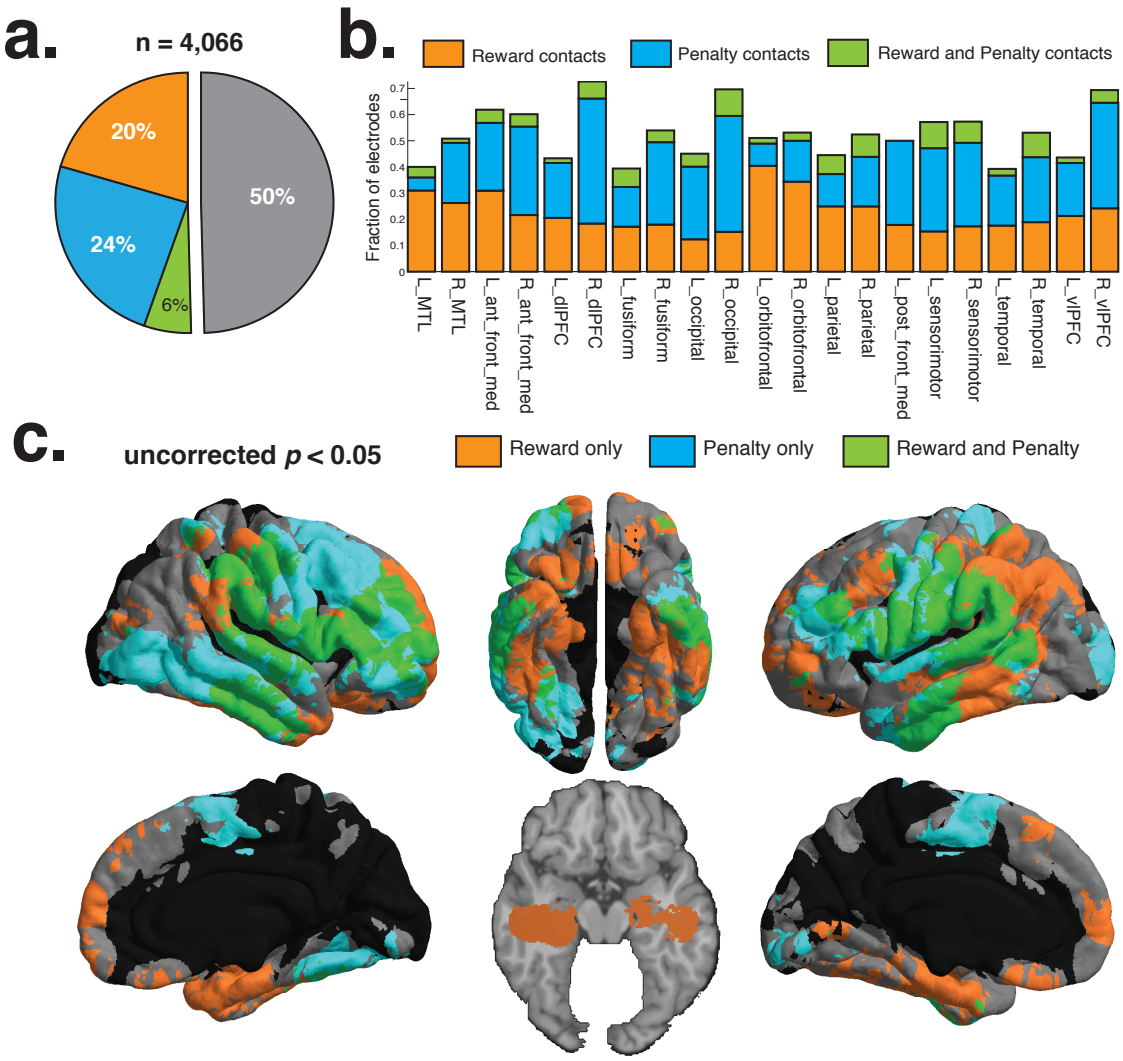

Figure 4.4: a. Fraction of reward (orange) and penalty (blue) contacts among all recorded contacts **b.** Fraction of positive and negative electrodes in each ROI. "*" indicates regions where subjects showed positive or negative effects more frequently than chance ($p < 0.05$, FDR-corrected). **c.** Searchlight-analysis. 12.5 mm spheres were centered on the vertices of the cortical surface and axial slice. We indicate regions where we more frequently observed reward and penalty contacts than expected by chance ($p < 0.05$, uncorrected). Orange - rewards contacts only, blue- penalty contacts only, green - overlapping reward and penalty contacts. We did not include individual contacts that demonstrated both reward and penalty contacts. Regions with neural data from 5 or fewer subjects are colored black. See main text for statistics.

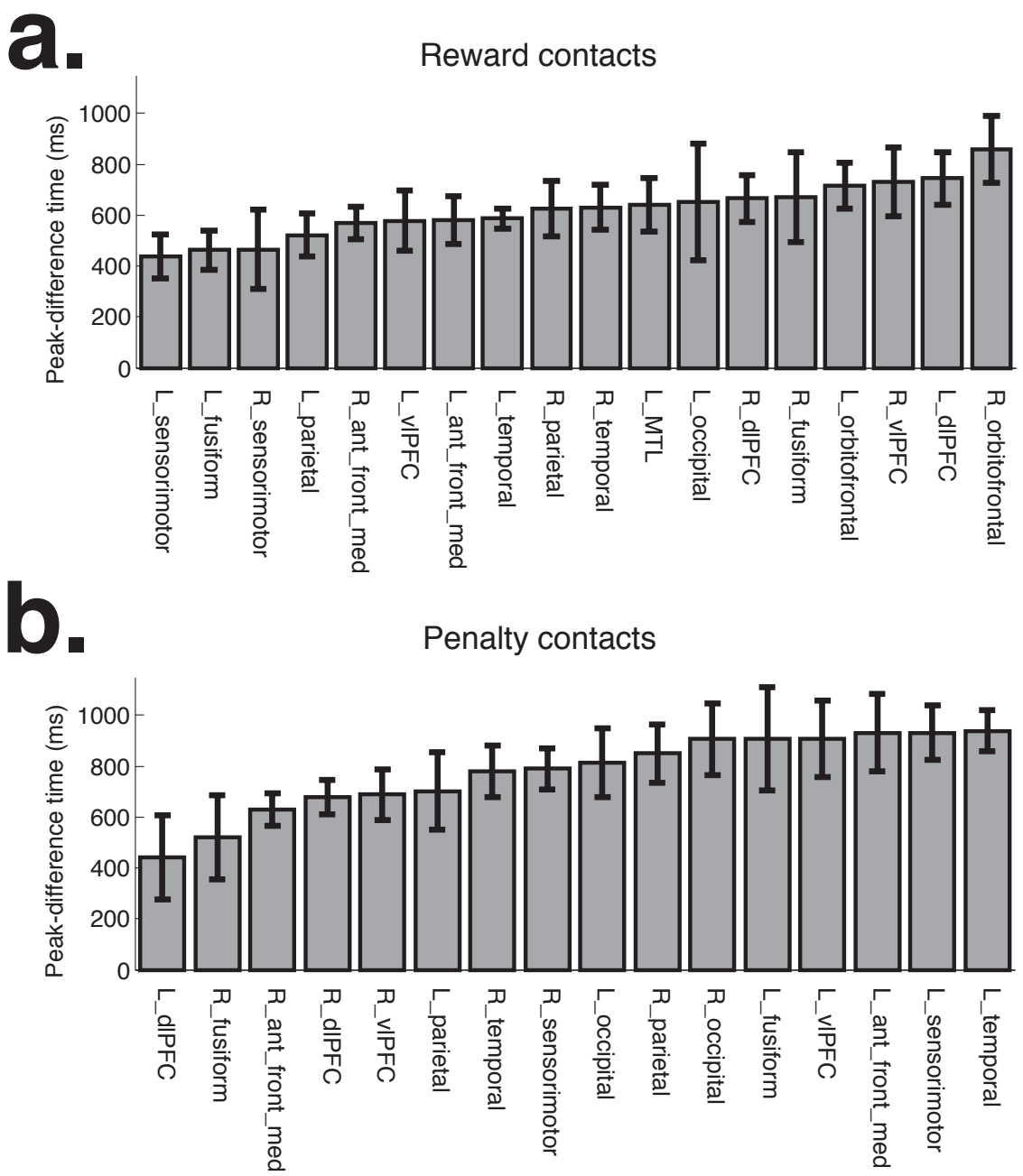

Figure 4.5: a,b. Mean times of peak-differences for reward and penalty contacts in regions that they were frequently observed. Error bars indicate s.e.m across subjects. Horizontal line indicates mean times among all reward and penalty contacts, respectively.

Chapter 5

General discussion

5.1 Conclusions

In this dissertation, we present results from three studies to shed novel insights on the neural basis of human reinforcement learning. In the first two studies, we shed light on the functional properties for dopaminegic neurons in the substantia nigra during reinforcement learning by studying patients undergoing deep brain stimulation surgery for Parkinson's disease. In Chapter 2, we analyze microelectrode recordings from the SN and provide electrophysiological evidence that putative DA neurons are functionally distinct from other neurons within the region. In Chapter 3, we study the effects of electrical microstimulation of the human SN on reinforcement learning. We show that manipulating the phasic activity of DA neurons during reinforcement learning via electrical stimulation can alter subjects performance during the task. These results demonstrate the first causal evidence for role of phasic DA activity during human RL. More specifically, our results suggest that SN DA neurons demonstrate a RPE signal that is specialized for training physical actions, a function that is consistent with the anatomical connectivity of

SN DA neurons (S. N. Haber et al., 2000), and previous neuroimaging studies (O'Doherty et al., 2004). This study demonstrates the first evidence that electrical microstimulation can be applied in a clinical setting to alter human reinforcement learning.

In Chapter 4, we study reinforcement learning in patients with drug-refractory epilepsy undergoing intracranial electroencephalography (iEEG) monitoring for resective surgery. We studied changes in high-frequency activity (HFA, 70-200 Hz), a known indicator of local firing rates, at electrode contacts distributed throughout the cortex and medial temporal lobe. By analyzing HFA changes separately at each electrode contact, we sought to identify heterogeneous representations of obtained and expected rewards. We replicated the main result from a recent multi-variate functional neuroimaging study (Vickery et al., 2011) by showed that valence information was widely represented in the cortex and MTL. We went beyond this study by showing that a regionally-distributed subset of these valence-representations were modulated by reward expectation. As predicted by prior theoretical work (Sutton & Barto, 1990), we found that that the most prominent patterns of activity were consistent with representations of unexpected rewards and penalties, signals that may guide following rewards and penalties, respectively. The strength of unexpected reward representations was correlated with subjects' performance during the task, suggesting a functional relevance for learning. Unexpected reward signals were prominently observed in right occipito-temporo-frontal regions. Thus, whereas valence information may be widely represented throughout the cortex and medial temporal lobe, a distributed subset of these signals (prominently observed in the right hemisphere) may represent unexpected outcome representations that are functionally relevant for learning.

Together, these results describe the existence of two distinct neural represen-

tations of learning signals during reinforcement learning. Midbrain DA neurons may represent a relatively homogeneous implementation of RPEs that are sufficient to modulate learning (Glimcher, 2011). The results from Chapter 3 suggest that there may be a functionally topographic arrangement of DA neurons within the midbrain; those in the ventromedial midbrain (ventral tegmental area) may be specialized for updating stimulus values (Tsai et al., 2009), whereas those in the dorsolateral region (substantia nigra) may be specialized for updating action values. Such regions may offer a practical opportunity to obtain physiological control of specific reinforcement learning using processes using methods such as electrical microstimulation. On the other hand, neural activity throughout the cortex may encode unexpected outcomes in a more heterogeneous manner. Although this heterogeneity may result in a more information-rich neural representation, it may also result in a neural representation that is more difficult to control via electrical stimulation. For these reasons, a practical strategy to obtain physiological control over human reinforcement learning may be to decode cognitive variables from the cortex via multi-site recordings, and influence behavior by manipulating cognitive representations in low-level nuclei (e.g., midbrain DA neurons).

5.2 Future directions

5.2.1 Testing a functional specialization in SN DA neurons

A direct follow-up to the study reported in Chapter 3 is to investigate the precise manner in which DA neurons guide human reinforcement learning. As suggested by our study, SN DA neurons maybe particularly important for reinforcing rewarded actions. To test for such a functional specialization, we will study the

activity of SN DA neurons during a probability learning task which manipulates the consistency of stimulus-response mapping. When there is consistent mapping between stimuli and responses ("pure-mapping"), rewards will be contingent on particular actions, whereas when there is inconsistent mapping between stimuli and responses ("mixed-mapping"), rewards will be contingent on stimuli, but decoupled from actions. By studying the dependence of DA neuronal responses and microstimulation-related behavioral changes on stimulus-response consistency, we will assess two competing hypotheses – 1) that SN DA neurons are functionally specialized to reinforce rewarded actions, and 2) that SN DA neurons are not specialized for action learning, but can also strengthen associations between rewards and preceding stimuli.

5.2.2 Low-frequency functional connectivity analyses of intracranial EEG

In Chapter 4, we demonstrate the existence of distributed neural representations that encode distinct information (e.g., unexpected penalty vs. unexpected reward signals) based on changes in HFA. Because HFA reflect local firing rates (Manning et al., 2009), an important question is to assess whether these distributed neural representations are synchronized across the brain by low-frequency rhythms (Buzsáki, 2006; Burke et al., 2013). Such rhythms may provide a neural substrate by which regionally-distributed neural representations may be coordinated together in a temporally-precise manner. If these representations reflect inputs from midbrain DA neurons, such temporal coordination would be necessary to allow for spike-time-depenent-plasticity and altered associative learning.

5.2.3 Studying unsigned prediction errors in Intra-cranial EEG

The results from Chapter 4 raised the possibility that a subset of unexpected outcome representations may represent unsigned prediction errors that assign salience to incoming feedback. Certain theories of reinforcement learning posit that such unsigned prediction errors inform subjects about how behaviorally salient the obtained feedback is, thus modulating its "associability". The more unexpected the obtained outcome, the larger the unsigned prediction error, and the larger the change in associative strength (Pearce & Hall, 1980). Additionally, such signals may be used to modulate learning rates in environments where there are varying degrees of uncertainty (Behrens et al., 2007; Nassar et al., 2012). Such signals may be computed in the amygdala based on reward prediction errors encoded by DA neurons and transmitted to several prefrontal regions, including the anterior cingulate (Roesch et al., 2012). To directly study these neural representations, one could study individual's choice behavior in a changing environment, and identify neural signals that are related to trial-by-trial updates of learning rate (Nassar et al., 2012).

5.2.4 The relation between reinforcement learning and episodic memory: Towards a comprehensive model of human learning

Since the days of Thorndike and Estes, there has been debate about whether human associative learning is mediated by a Law of Effect principles, where stimulus-response associations are retrospectively strengthened by obtained rewards, or whether it is mediated by principles of episodic memory, where associations are formed based on temporal contiguity and used to make projections of the future and

make decisions in a goal-directed manner (Thorndike, 1932; Estes, 1967). Over the past 30 years or so, evidence has emerged that these systems may both exist within the brain and work in a competitive manner to generate decisions (Redish, 2013). Several theories of multiple learning systems have emerged, from non-quantitative frameworks such as procedural and declarative systems, to quantitative models formalizing multiple systems of learning. For example, COVIS (competition of verbal and implicit systems) represents such a formalism as applied to category learning, whereby decision-bound and rule-based modules compete to form a decision (Ashby & Maddox, 2005). Also, model-free and model-based reinforcement learning models represent retrospective, and prospective approaches to alter reward expectations during reinforcement learning (Sutton & Barto, 1990). Recent studies have demonstrated direct links between model-based approches have been and episodic memory processes (Gershman, Schapiro, Hupbach, & Norman, 2013; Doll, Shohamy, & Daw, 2014). In another line of research addressing the issue of multiple learning systems, it has been shown that simply delaying the timing of feedback presentation during a two-alternative probability learning task may shift learning from a law-of-effect to a more episodic system (Foerde et al., 2013).

There are two major goals for future research. First, there must be an effort to reconcile the similarities and differences between the various reinforcement learning models that have been proposed so far as explanations of human behavior. The insights from this effort should be used to generate a general model of human learning, that can explain behavior on a wide-range of learning tasks (e.g., both probabilistic classification and list learning). This theory of human learning should be used to make novel predictions regarding the interaction between the multiple learning systems that have not previously been tested. Quantitative model fits to subjects' behavioral data should be used to characterize individual differences in

learning, and also describe group-level differences between healthy and patient populations. Second, there must be a research effort to map components of this theory to the human brain. A place to start may be to simultaneously record neural activity from regions that are known to be important for law-of-effect learning (e.g., midbrain DA neurons) and episodic memory processes (e.g., medial temporal lobe neurons) as subjects perform a task that is sensitive to individual differences in the degree to which retrospective and prospective learning strategies (Otto, Gershman, Markman, & Daw, 2013). Ultimately, this a comprehensive theory of human learning may be used as a framework to study failure modes that occur in psychiatric disease, so as to identify the dysfunctional neural systems on an individual basis and guide clinical therapy (Redish, 2013; Maia & Frank, 2011). There is much work to be done.

References

Adams, J. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin*, *101*(1).

Addison, P. S. (2002). *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance.* Bristol: Institute of Physics Publishing.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current opnion in neurobiology*, *12*(2), 169-177.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*, 6.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal ofMathematical Psychology*, *37*, 372-400.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Barto, A., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd international conference on development and learning.*

Bartra, O., McGuire, J., & Kable, J. (2013). The valuation system: A coordinate-based meta-analysis of bold fmri experiments examining the neural correlates

of subjective value. *NeuroImage*, *76*, 412-427.

Bayer, H., & Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Bayer, H., & Glimcher, P. (2007). Statistics of midbrain dopaminergic neuron spike trains in the awake primate. *Journal of Neurophysiology*, *98*(3), 1428-1439.

The behavior of organisms: An experimental analysis. (1938). Chicago.

Behrens, T., Woolrich, M. W., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214-1221.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Series B*, *57*, 289-300.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*(8), 2793-2798.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*, 442–477. doi: 10.1162/neco.2007.19.2.442

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*.

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *NeuroImage*, *85 Pt. 2*, 834–843.

Burke, J. F., Zaghloul, K. A., Jacobs, J., Williams, R. B., Sperling, M. R., Sharan, A. D., & Kahana, M. J. (2013). Synchronous and asynchronous theta and gamma activity during episodic memory formation. *Journal of Neuroscience*,

33(1), 292-304.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413.

Buzsáki, G. (2006). *Rhythms of the brain*. New York: Oxford University Press.

Buzsaki, G., Anastassiou, C., & Koch, C. (2012). The origin of extracellular fields and currents - eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, *13*, 407-419.

Carpenter, M., Nakano, K., & Kim, R. (1976). Nigrothalamic projections in the monkey demonstrated by autoradiographic technics. *Journal of Comparative Neurology*, *165*(4).

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Duzel, E., & Dolan, R. (2013). Dopamine restores reward prediction errors in old age. *Nature Neuroscience*, *16*(5), 648-653.

Clark, K., Armstrong, K., & Moore, T. (2011). Probing neural circuitry and function with electrical microstimulation. *Proceedings of the Royal Society B: Biological Sciences*.

Cohen, J., Haesler, S., Vong, L., Lowell, B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the Ventral Tegmental Area. *Nature*, *482*, 85-88.

Cools, R., Barker, R., Sahakian, B., & Robbins, T. (2001). Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. *Cerebral Cortex*, *11*, 1136–1143.

Dale, A. M., Fischl, B., & Sereno, M. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179-194.

Damier, P., Hirsch, E., Agid, Y., & Graybiel, A. M. (1999a). The substantia nigra of the human brain ii. patterns of loss of dopamine-containing neurons in

Parkinson's disease. *Brain*, *122*, 1437-1448.

Damier, P., Hirsch, E., Agid, Y., & Graybiel, A. M. (1999b). The substantia nigra of the human brain i. nigrosomes and the nigral matrix, a compartmental organization based on calbindin d28k immunohistochemistry. *Brain*, *122*(8), 1421-1436.

Daw, N., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4-6), 603-616.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876-879.

DeLong, M., Crutcher, M., & Georgopoulos, A. P. (1983). Relations between movement and single cell discharge in the substantia nigra of the behaving monkey. *Journal of Neuroscience*, *3*(8), 1599-1606.

Desikan, R., Segonne, B., Fischl, B., Quinn, B., Dickerson, B., Blacker, D., . . . Killiany, N. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968-80.

Doll, B., Shohamy, D., & Daw, N. (2014). Multiple memory systems as substrates for multiple decision systems. *Neurobiology of learning and memory*.

Duncan, S., & Barrett, L. (2007). The role of the amygdala in visual awareness. *Trends in cognitive sciences*, *11*(5), 190-192.

Engel, A. K., Moll, C. K. E., Fried, I., & Ojemann, G. A. (2005). Invasive recordings from the human brain–clinical insights and beyond. *Nature Reviews Neuroscience*, *6*, 35–47.

Estes, W. K. (1967). *Reinforcement in human learning*. Defense Technical Information Center.

Estes, W. K. (1986, Oct). Array models for category learning. *Cognitive Psychology*,

18(4), 500-549.

Fiorillo, C., Yun, S., & Song, M. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *Journal of Neuroscience*, *33*(11), 4693-709.

Fischl, B., Sereno, M., Tootell, R., & Dale, A. M. (1999). High-resolution inter-subject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*, 272-284.

Foerde, K., Race, E., Verfaellie, M., & Shohamy, D. (2013). A role for the medial temporal lobe in feedback-driven learning: Evidence from amnesia. *Journal of Neuroscience*, *33*(13), 5698-5704.

Frank, L. M., Stanley, G., & Brown, E. (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience*, *24*(35), 7681–7689.

Frank, M., Samanta, J., Moustafa, A., & Sherman, S. (2007). Hold your horses: Impulsivity , deep brain stimulation, and medication in parkinsonism. *Science*, *318*, 1309–1312.

Frank, M., & Surmeier, D. (2009). Do substantia nigra dopaminergic neurons differentiate between reward and punishment? *Journal of Molecular Cell Biology*, *1*, 15-16.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*, 1940–1943.

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, *33*(20), 8590 - 8595.

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, USA*, *108*(3), 15647-15654.

Gluck, M., & Bower, G. (1988). Evaluating an adaptive network model of human learning. *Journal Of Memory And Language*, *27*(2), 166-195.

Grattan, L., Rutledge, R., & Glimcher, P. (2011). Increased dopamine concentrations increase the perceived value of an action. In *Program No. 732.12. Society for Neuroscience Meeting Planner.* San Diego, CA.

Haber, S., & Knutson, B. (2009). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*(1), 4–26.

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, *20*(6), 2369-2382.

Handel, A., & Glimcher, P. (2000). Contextual modulation of substantia nigra pars reticulata neurons. *Journal of Neurophysiology*, *83*(5), 3042-3048.

Henny, P., Brown, M., Northrop, A., Faunes, M., Ungless, M., Magill, P., & Bolam, J. (2012). Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons. *Nature Neuroscience*, *15*(4), 613-619.

Hikosaka, O., & Wurtz, R. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. i. relation of visual and auditory responses to saccades. *Journal of Neurophyiology*, *49*(5), 1230-1253.

Histed, M., Bonin, V., & Reid, C. (2009). Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. *Neuron*, *63*, 508-522.

Hollerman, J., & Grace, A. (1990). The effects of dopamine-depleting brain lesions on the electrophysiological activity of rat Substantia Nigra dopamine neurons. *Brain Research*, *533*, 203-212.

Hunga, Y., Smith, M., Bayle, D., Mills, T., Cheyne, & Taylor, M. J. (2010). Unattended emotional faces elicit early lateralized amygdala-frontal and fusiform

activations. *NeuroImage*, *50*(2), 727-733.

Jaggi, J., Umemura, A., Hurtig, H., Siderowf, A., Colcher, A., Stern, M., & Baltuch, G. (2004). Bilateral subthalamic stimulation of the subthalamic nucleus in Parkinson's disease: surgical efficacy and prediction of outcome. *Stereotactc & Functional Neurosurgery*, *82*, 104–14.

Joshua, M., Adler, A., Rosin, B., Vaadia, E., & Bergman, H. (2009). Encoding of probabilistic rewarding and aversive events by pallidal and nigral neurons. *Journal of Neurophysiology*, *101*, 758-772.

Kable, J., & Glimcher, P. (2009). The neurobiology of decision: Consensus and controversy. *Neuron*, *63*, 733-745.

Kahnt, T., Heinzle, J., Park, S., & Haynes, J. (2011). Decoding the formation of reward predictions across learning. *Journal of Neuroscience*, *31*(41), 14624-14630.

Kamin, L. (1969). Selective association and conditioning. In N. Mackintosh & W. Honig (Eds.), *Fundamental issues in instrumental learning* (pp. 42–64). Halifax, Canada: Dalhousie University Press.

Klucharev, V., Hytonen, R. M. S. A., K., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140-151.

Knowlton, B., Mangles, J., & Squire, L. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402.

Lafreniere-Roula, M., Hutchinson, W., Lozano, A., Hodaie, M., & Dostrovsky, J. (2009). Microstimulation-induced inhibition as a tool to aid targeting the ventral border of the subthalamic nucleus. *Journal of Neurosurgery*, *111*(4), 724-728.

Lau, B., & Glimcher, P. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451-463.

Lega, B. C., Kahana, M. J., Jaggi, J. L., Baltuch, G. H., & Zaghloul, K. A. (2011). Neuronal and oscillatory activity during reward processing in the human ventral striatum. *NeuroReport*, *22*(16), 795-800.

Lobb, C., Wilson, C., & Paladini, C. (2011). High-frequency, short-latency disinhibition bursting of midbrain dopaminergic neurons. *Journal of Neurophsyiology*, *105*, 2501-2511.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157.

Luscher, C., & Ungless, M. (2006). The mechanistic classification of addictive drugs. *PLoS Medicine*, *3*(11).

Ma, S., Rinne, J., Collan, Y., Roytta, M., & Rinne, U. (1996). A quantitative morphometrical study of neuron degeneration in the substantia nigra in Parkinson's disease. *Journal of the neurological sciences*, *140*(1-2), 40–45.

Maia, T., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154-162.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003, Jul). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, *19*(3), 1233–1239.

Manning, J. R., Jacobs, J., Fried, I., & Kahana, M. J. (2009). Broadband shifts in LFP power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience*, *29*(43), 13613 - 13620.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(11), 837-841.

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.

Menke, R., Jbabdi, S., Miller, K., Matthews, P., & Zarei, M. (2010). Connectivity-based segmentation of the Substantia Nigra in human and its implications in Parkinson's disease. *Neuroimage*, *52*, 1175-1180.

Montague, P., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417-448.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Morita, K., Morishima, M., Sakai, K., & Kawaguchi, Y. (2012). Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends in Neurosciences*, *35*(8), 457-467.

Morris, J., Ohman, A., & Dolan, R. (1999). A subcortical pathway to the right amygdala for mediating "unseen" fear. *Proceedings of the National Academy of Science USA*, *96*(4), 1680-1685.

Morrison, S., & Salzman, D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *Journal of Neuroscience*, *29*(37), 11471-11483.

Moyer, J., Danish, S., Keating, G., Finkel, L., & Baltuch, G. (2007). Implementation of dual simultaneous microelectrode recording systems during deep brain stimulation surgery for Parkinson's disease: Technical note. *Operative Neurosurgery Supplement I*, *60*, E177–78.

Nair-Roberts, R., Chatelain-Badie, S., Benson, E., White-Cooper, H., Bolam, J., & Ungless, M. (2008). Stereological estimates of dopaminergic, GABA-ergic, and glutamatergic neurons in the Ventral Tegmental Area, Substantia Nigra

and Retrorubal Field in the rat. *Journal of Neuroscience*, *152*(4), 1024-1031.

Nassar, M., Rumsey, K., Wilson, R., Parikh, K., Heasly, B., & Gold, J. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*, 1040-1046.

Neymotin, S., Lyton, W., A.O., O., & A.A., F. (2011). Measuring the quality of neuronal identification in ensemble recordings. *Journal of Neuroscience*, *31*(45), 16398-16409.

Niv, Y., Daw, N., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507-520.

Niv, Y., & Montague, P. (2009). Theoretical and empirical studies of learning. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (chap. 22). London: Academic Press.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452-454.

Otani, S., Daniel, H., Roisin, M., & Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cerebral Cortex*, *13*(11), 1251-1256.

Otto, R., Gershman, S., Markman, A., & Daw, N. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751-761.

Padoa-Schiopa, C., & Assad, J. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*, 223-226.

Pan, W. X., Brown, J., & Dudman, J. (2013). Neural signals of extinction in the inhibitory microcircuit of the ventral midbrain. *Nature Neuroscience*, *16*(1), 71-78.

Patel, S., Sheth, S., Gale, J. T., Greenberg, B., Dougherty, D., & Eskandar, E. N. (2012). Single-neuron responses in the human nucleus accumbens during a financial decision-making task. *Journal of Neuroscience*, *32*(21), 7311-5.

Paton, J., Belova, M., Morrison, S., & Salzman, C. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865-870.

Pavlov, I. (1927). *Conditioned reflexes*. New York, NY, US: Oxford University Press.

Pearce, J., & Hall, G. (1980). A model for pavlovian conditioning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–555.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. (2006). Dopamine-dependent prediction errors underpin reward-seeking behavior in humans. *Nature*, *442*, 1042-1045.

Platt, M., & Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233-238.

Poirier, L., Giguére, M., & Marchand, R. (1983). Comparative morphology of the substantia nigra and ventral tegmental area in the monkey, cat and rat. *Brain Research Bulletin*, *11*, 371-397.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(23), 1102–1107.

Ramayya, A. G., Misra, A., Baltuch, G. H., & Kahana, M. J. (2014). Microstimulation of the human substantia nigra following feedback alters reinforcement learning. *Journal of Neuroscience*, *34*(20), 6887–6895.

Ray, S., & Maunsell, J. (2011). Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biology*, *9*(4), e1000610.

Redish, A. D. (2013). The mind within the brain: How we make decisions and how those decisions go wrong. *Oxford University Press*.

Rescorla, R., & Wagner, A. (1972). A theory of pavolvian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Reynolds, J., Hyland, B., & Wickens, J. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67–70.

Roesch, M., Esber, G., Li, J., Daw, N., & Schoenbaum, G. (2012). Surprise! neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190-1200.

Rutledge, R., Dean, M., Caplin, A., & Glimcher, P. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*, *30*(40), 13525-13536.

Rutledge, R., Lazzaro, S., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *Journal of Neuroscience*, *29*(48), 15104-15114.

Sato, M., & Hikosaka, O. (2002). Role of primate substantia nigra pars reticulata in reward-oriented saccadic eye movement. *Journal of Neuroscience*, *22*(6), 2363-2373.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Schultz, W., & Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology*, *57*, 201-217.

Seymour, B., O'Doherty, J., Dayan, P., Koltzenburg, M., Jones, A., Dolan, R. J., ... Frackowiak, R. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664-667.

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, D. P., K.P., & Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: evidence from parkinson's disease. *Brain*, *135*, 1871-1883.

Steinberg, E., Keiflin, R., Boivin, J., Witten, I., Deisseroth, K., & Janak, P. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966-973.

Stephens, D. (1986). *Foraging theory*. Princeton Univ. Pr.

Sugrue, L., Corrado, G., & Newsome, W. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*, 363-375.

Sutton, R., & Barto, A. (1990). Time-derivative models of pavolovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Takahashi, Y., Roesch, M., Wilson, R., Toreson, K., O'Donnell, P., Niv, Y., & Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature Neuroscience*, *14*(12), 1590-1597.

Tepper, J., Martin, L., & Anderson, D. (1995). GABA-A receptor-mediated inhibition of rat Substantia Nigra dopaminergic neurons by pars reticulata projection neurons. *Journal of Neuroscience*, *15*(4), 3092-3103.

Thorndike, E. L. (1932). *The fundamentals of learning*. New York: Bureau of Publications, Teachers College.

Tsai, H., Zhang, F., Adamatidis, A., Stuber, S., Garret, Bonci, A., Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for

behavioral conditioning. *Science*, *324*(5930), 1080-1084.

Ungless, M., & Grace, A. (2012). Are you or aren't you? Challenges associated with physiologically identifying dopamine neurons. *Trends in Neurosciences*, *35*, 422-30.

Vandercasteele, M., Glowinski, J., & Venance, L. (2005). Electrical synapses between dopaminergic neurons of the substantia nigra pars compacta. *Journal of Neuroscience*, *25*(2), 291-298.

Vickery, T., Chun, M., & Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron*, *72*(1), 166-177.

Wallis, J. D., & Kennerley, S. (2011). Contrasting roles of reward signals in the orbitofrontal and anterior cingulate cortex. *Annals of the New York Academy of Sciences*, *1239*, 33-42.

Wang, Y., Zhang, Q., Ali, U., Gui, Z., Hui, Y., Chen, L., & Wang, T. (2010). Changes in firing rate and pattern of GABA-ergic neurons in subregions of the Substantia Nigra pars reticulata in rat models of Parkinson's Disease. *Brain Research*, *1324*, 54-63.

Waszczak, B., & Walters, J. (1983). Dopamine modulation of the effects of gamma-aminobutyric acid on Substantia Nigra pars reticulata neurons. *Science*, *220*(218-221).

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human Substantia Nigra neurons encode unexpected financial rewards. *Science*, *323*, 1496–1499.

Zaghloul, K. A., Lega, B. C., Weidemann, C. T., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal activity in the human Subthalamic Nucleus encodes decision conflict during action selection. *Journal of Neuroscience*, *32*(7), 2453–2460.

Zigmond, M., Abercrombie, E., Berger, T. W., Grace, A., & Stricker, E. (1990). Compensations after lesions of central dopaminergic neurons: some clinical and basic implications. *Trends in Neurosciences*, *13*, 290-296.