DOPAMINE IN REWARD LEARNING AND NEUROPSYCHIATRIC DISORDERS

CONTEXT AND SALIENCE: THE ROLE OF DOPAMINE IN REWARD LEARNING AND NEUROPSYCHIATRIC DISORDERS

By TRENT TOULOUSE, B.S.

A Thesis submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

McMaster University

© Copyright by Trent Toulouse, January 2013

Doctor of Philosophy (2013) McMaster University

(Psychology) Hamilton, Ontario

TITLE: Context and Salience: The Role of Dopamine in Reward Learning and Neuropsychiatric Disorders

AUTHOR: Trent Toulouse, B.S. (McMaster University)

SUPERVISOR: Professor Suzanna Becker

NUMBER OF PAGES: ix, 123

Abstract

Evidence suggests that a change in the firing rate of dopamine (DA) cells is a major neurobiological correlate of learning. The Temporal Difference (TD) learning algorithm provides a popular account of the DA signal as conveying the error between expected and actual rewards. Other accounts have attempted to code the DA firing pattern as conveying surprise or salience. The DA mediated cells have also been implicated in several neuropsychological disorders such as obsessive compulsive disorder and schizophrenia. Compelling neuropsychological explanations of the DA signal also frame it as conveying salience. A model-based reinforcement learning algorithm using a salience signal analogous to dopamine neurons was built and used to model existing animal behavioral data.

Different reinforcement learning models were then compared under conditions of altered DA firing patterns. Several differing predictions of the TD model and the salience model were compared against animal behavioral data in an obsessive compulsive disorder (OCD) model using a dopamine agonist. The results show that the salience model predictions more accurately model actual animal behavior.

The role of context in the salience model is different than the standard TD-learning algorithm. Several predictions of the salience model for how people should respond to context shifts of differing salience were tested against known behavioral correlates of endogenous dopamine levels. As predicted, individuals with behavioral traits correlated with higher endogenous dopamine levels are far more sensitive to low salience context shifts than those with correlates to lower endogenous dopamine levels. This is a unique prediction of the salience model for the DA signal which allows for better integration of reinforcement learning models and neuropsychological frameworks for discussing the role of dopamine in learning, memory and behavior.

Acknowledgments

Primary thanks goes to my supervisor, Sue Becker, whose support and assistance with this research went above and beyond the call of duty. I would also like to extend thanks to my committee members, Scott Watter and Pat Bennett, for their assistance and ideas throughout my time at McMaster.

Finally I would like to extend special thanks to my friends and family, particularly my mother and father, for their unwavering support in helping me complete this thesis.

Table of Contents

Abstract……...iii

Acknowledgements…….iv

List of Figures……vii

1 Introduction…….....1

1.1 Anatomy of the VTA dopamine signal……...2

1.2 Biophysical basis of the DA signal……4

1.3 Cell recording data of DA cells……..5

1.4 Classes of reinforcement learning models…….7

1.5 Conflicting empirical data……..8

1.6 Neuropsychiatric explanations of the DA signal……9

2 Introduction…….13

2.1 Model details……22

2.2 Methods……26
2.2.1 Reward devaluation……...…26
2.2.2 Reward devaluation after extensive training ……29
2.2.3 Reward devaluation after moderate training…….29
2.2.4 Lesioning the dopaminergic pathway……30
2.2.5 DA sensitivity to reward probability…….30
2.2.6 DA signal with 100 percent reward delivery probability……..30
2.2.7 DA signal with 50 percent reward delivery probability……31

2.3 Results…….32
2.3.1 Reward devaluation after extensive training ……...32
2.3.2 Reward devaluation after moderate training ……...34
2.3.3 Lesioning the dopaminergic pathway……...36

2.4 DA sensitivity to reward probability……..38
2.4.1DA signal with 100 percent reward delivery probability……38
2.4.2 DA signal with 50 percent reward delivery probability……..39

2.5 Discussion……..40
2.5.1 Model-based reward devaluation……40
2.5.2 Context, learning and the hippocampus……..42
2.5.3 Dopamine as a reward probability signal ……...43
2.5.4 Limitations and future work……....45
2.5.5 Conclusions……46

3 Introduction……..48

3.1 Experiment 1……..55
3.1.1 Methods……...55
3.1.2 The Model……...58
3.1.3 Results……61

3.2 Experiment 2 ……64
3.2.1 Method……64
3.2.2 The Model……...65
3.2.3 Results……69

3.3 Discussion……..74

4 Introduction……...…76

4.1 Simulation ……80
4.1.1 Methods……..80
4.1.2 Results……81

4.2 Experiment 1……82
4.2.1 Method……...83
4.2.2 Results……84

4.4 Experiment 2……88
4.4.1 Methods……..89
4.4.2 Results……91

4.5 Discussion……..96

5 Overview…….101

5.1 Contributions to reinforcement learning models……..103

5.2 Contributions to obsessive compulsive disorder and animal models……...106

5.3 Contributions to understanding individual differences and endogenous dopamine.109

5.4 Conclusions……...111

References……...113

Appendix…….120

List of Figures

Figure 1: A schematic of the reward devaluation simulations used to compare models…… 28

Figure 2: Model results showing devaluation insensitivity after overtraining…….33

Figure 3: Model results looking at devaluation sensitivity with only moderate training……...35

Figure 4: Model results of reward devaluation with the DA signal set to zero……37

Figure 5: Plot showing the DA signal across time in a trial, and trial number for the standard conditioning paradigm used in Schultz……...……38

Figure 6: Plot showing the DA signals for 0.05 and 0.15 reward magnitudes delivered with 50 percent probability……39

Figure 7: Position recordings for ten rats across Quinpirole and saline trials……..57

Figure 8: Schematics for how subsections in the simulation correspond to the data and Markov tree breakdown for states and action choices……...59

Figure 9: A boxplot comparison of the model-based and TD-algorithm approaches…..62

Figure 10: Learning rate as measured from animal behavioral data showing a close match to the model-based approach……63

Figure 11: A schematic showing the Markov chain where the decision of action A or B does not alter reward beyond some base discount function..…….65

Figure 12: A comparison of the model-based and TD-algorithm approaches to path reinforcement…….69

Figure 13: Animal data showing the percent similarity in trips over time……...…70

Figure 14: Animal data showing that there is a low and high group for percent similarity between trips……71

Figure 15: Animal data looking at increased activity over time (a prediction of the TD-approach) versus percent similarity between trips……...73

Figure 16: Model data showing a comparison between low and high salience context shifts with normal DA levels……81

Figure 17: Model data showing a comparison between low and high salience context shifts with elevated DA levels……...82

Figure 18: Experimental data showing the correlation between SPQ scores and the accuracy difference between unexpected punishment versus unexpected reward…….85

Figure 19: Experimental data showing the relationship between SPQ score on non-reward-based, contextual learning…….87

Figure 20: Screen shots from the context salience based reversal-learning task.……....91

Figure 21: Empirical data showing the relationship between SPQ and two other potential correlates of increased endogenous dopamine…….....92

Figure 22: Empirical data showing the relationship between working memory and other potential correlates of endogenous dopamine levels……..93

Figure 23: Empirical data showing the relationship between accuracy on the low and high-context tasks and the set of potential endogenous dopamine level correlates…….95-96

Chapter 1

Introduction

This thesis examines the role of dopamine-mediated (DA) cells in the ventral tegmental area (VTA) in learning, memory and context processing, and how disruption of this system can explain elements of major psychological disorders. Interest in the dopamine-mediated VTA neurons has mostly been driven by work in reward learning. For all animals, including humans, reward-based learning is a powerful drive behind behavior — so powerful, in fact, that it drives addiction and can motivate self-destructive behaviors. Significant advances in understanding the neural mechanisms of reward learning have come from the machine learning field of reinforcement learning. Computational modelers have found strong links between the burst firing pattern of dopamine neurons and key mathematical terms in machine learning. However, there are competing theories in reinforcement learning regarding just what the dopamine cells are signaling and how they integrate into larger neural systems that motivate behavior. In the field of neuropsychology, disruption in the VTA dopamine cells has been identified as a potential source for symptoms in a range of psychological disorders, but the formulations of what the VTA dopamine cells are doing differ from the standard reinforcement learning hypotheses. The work here is an attempt to unify some of the conflicting theories in reinforcement learning and find a link between the conceptualization of dopamine in

neuropsychology and that of machine learning. Finding a unified view of what is actually conveyed by the VTA dopamine cells has potential to offer significant insights for reinforcement learning as well as increasing our understanding of the mechanisms behind a range of mental disorders.

Anatomy of the VTA dopamine signal

Dopaminergic neurons have been strongly implicated in reinforcement learning. There are two major regions of the brain containing the cell bodies of the dopaminergic neurons, the VTA and the subtantia nigra (SN). The VTA dopaminergic pathway is of particular interest here because of its role in reinforcement learning whereas the SN is more strongly implicated in motor control. The VTA is a group of neurons located at the bottom of the midbrain close to the midline. The VTA projects to a wide array of brain structures, most of which have reciprocal connections. The excitatory glutamatergic inputs into the VTA drive the dopamine neurons to fire in brief bursts. These bursts are the primary source of dopamine for areas such as the medial prefrontal cortex and the nucleus accumbens. This phasic dopamine firing pattern has gained considerable attention from researchers as a potential mediator of reinforcement learning, motivated behavior and neuropsychiatric disorders. Schultz et al. (1997) demonstrated that the temporal pattern of phasic firing of VTA dopamine neurons seems to correspond to the reward-prediction error term in the Temporal-Difference (TD) learning procedure, a

highly influential reinforcement learning algorithm. This finding suggested that the main role of the dopaminergic bursting is in mediating learning. On the other hand, Berridge and Robinson (1998) showed that disruption of the VTA disrupted goal-directed behavior. Therefore they suggested that, in addition to any role dopamine may play in learning, it has immediate effects on motivated behavior, and suggested a role for dopamine in signaling "incentive salience." Dysregulation of the mesolimbic dopamine system has also been shown to play a major role in schizophrenia, adult attention deficit disorder (ADHD), and obsessive compulsive disorder (OCD). Work by Kapur (see Kapur, 2003) attempting to understand the specific role of this dopamine dysregulation in schizophrenia attributes the burst firing to a signal of "aberrant salience," leading to spurious learning and motivated behavior, and driving much of the positive symptomatology such as delusions and paranoia. In addition to reward-prediction error, incentive salience, and aberrant salience, the burst firing of the VTA dopaminergic neurons has been linked to a signal of aversion, general uncertainty and novelty (for review see Symmonds et al., 2011). Although there is a lot of overlap between these proposed mechanisms, there are some important differences. Understanding exactly what the dopamine cells are signaling has significant implications for a wide range of research and clinical approaches to neuropsychiatric disorders.

Biophysical basis of the DA signal

The excitatory afferent connections to the VTA were traditionally thought to be somewhat limited, originating mostly from the prefrontal cortex, bed nucleus of the stria terminalis, and laterodorsal and pedunculopontine tegmental nuclei (Sesack & Pickel, 1992; Charara et al., 1996; Omelchenko & Sesack, 2005). However, evidence has emerged that the VTA actually has significant reciprocal connections throughout the brain, from the prefrontal cortex to the brainstem. Geisler et al. (2007) found glutamatergic inputs from the Claustrum/dorsal endopiriform area, central gray, cuneiform nucleus, dorsal raphe, lateral hypothalamic area, lateral habenula, lateral preoptic area, medial hypothalamus, medial preoptic area, median raphe, medial septum/diagonal band complex, parabrachial nucleus, prefrontal cortex, pedunculopontine and laterodorsal tegmental nuclei, prelimbic cortex, reticular formation, sublenticular substantia innominata, ventral endopiriform area, and the ventral pallidum into the VTA. Most of these connections are reciprocal; however, the accumbens, a major efferent connection, does not actually project back to the VTA.

Approximately 55-65 percent of the cells within the VTA are dopaminergic neurons with the remainder primarily being GABAergic inhibitory cells (Cohen et al., 2012). These dopaminergic cells exhibit diverse burst firing either through excitatory input directly or inhibition of the GABAergic cells in the VTA (Cohen et al., 2012). This burst firing is the primary source of dopamine for the prefrontal cortex and the nucleus

accumbens and is often referred to as the dopaminergic signal. The known role of the prefrontal cortex and the nucleus accumbens in reward learning and motivated behavior sparked interest from these fields in the VTA.

Another highly interesting connection loop is between the VTA and the hippocampus. Luo et al. (2011) identified a circuit that links the CA3 of the dorsal hippocampus to the VTA. They found that the CA3 output modulated the strength of the DA signal and that disruption of this circuit actually stopped cocaine-seeking behavior in rats following context cues related to previous cocaine consumption. Martig and Mizumori (2011) examined efferent connections from the VTA to the hippocampus. They looked at changes in place fields in a reward learning task that correlated with changes in reward, reward position and disruption of the VTA inputs. Disruption of the VTA increased errors significantly and reorganized the place fields in CA1 and CA2. Reward position change, but not magnitude change, was actually able to rescue place field organization after VTA disruption. Together this evidence suggests an important relationship between context, memory and the DA signal.

Cell recording data of DA cells

One of the first systematic explorations of the relationship between DA cell firing rates and reward was done by Schultz et al. (1997) in a series of DA cell recordings in the VTA of monkeys in a classical conditioning paradigm. In a typical trial, a symbol was

presented on a screen followed by a delay, after which the monkey received some juice. Early on in the conditioning task the DA cells would fire in response to receiving the primary reward, the juice, but after multiple trials the DA cells stopped responding to the primary reward and instead began responding to the earliest predictor – the symbol preceding the reward. This conditioning effect suggested that the DA bursts were conveying a positive reward-prediction error when an initially unexpected reward was delivered; as the reward became predictable, the error signal and the DA bursting in response to the reward diminished. On the other hand, after extensive learning had taken place, if juice was not delivered after the symbol was presented, the cell firing rate actually fell below baseline, suggesting that the DA bursting (or lack thereof) also conveyed a negative reward-prediction error when less reward was received than predicted.

The proposed explanation for this firing pattern was that the DA signal was firing in response to unpredicted reward. Early on, when the animal did not have a learned association between the symbol and the juice reward, the DA cells would fire in response to reward delivery. After having learned the association, there was no response to the primary reward, but failure to get the predicted reward would decrease the firing rate.

This basic pattern of DA cell responsiveness has been replicated multiple times in various animal models (Tsai et al., 2009). The same basic pattern has also been identified in humans using fMRI. By convolving the timing of the expected DA signal with the

hemodynamic response function, researchers have identified the ventral striatum (VS) as responding to a range of rewards. This includes attractive faces (Senior, 2003), erotic images (Sabatinelli et al., 2007; Walter et al., 2008), favorable social interactions (Zink et al., 2008), monetary rewards (Knutson & Bossaerts, 2007), and pleasant tastes (O'Doherty et al., 2002). Carlson et al. (2011) combined ERP and fMRI to show that the VS activation closely matches the same cell recording data found in animals.

Classes of reinforcement learning models

Some of the most significant insights into what the DA signal actually conveys and its role in learning have come from the field of machine learning. There are two main approaches to modeling reinforcement learning, model-free and model-based algorithms. Undoubtedly the most influential of these machine learning algorithms is Sutton and Barto's (1998) TD learning algorithm, which successfully accounts for many aspects of animal learning. This reward-prediction error term actually behaves very similarly to the cell firing rates as measured by Schultz et al. (1997). Thus, it has been hypothesized that the neural basis of reinforcement learning may be well modeled by TD-learning.

There also are model-based reinforcement learning algorithms, where the animal constructs a detailed internal model of the world. This model is constructed by learning to associate states with actions and state transition probabilities, learning which states are associated with direct receipt of reinforcement, and learning which states are terminal

states (e.g. the food location at the end of a maze). In this case, action selection involves forward simulation of the world model to evaluate chains of state-action pairs and their outcomes. The DA signal has been postulated to convey surprise and salience in these model-based accounts (see Smith et al., 2006).

Conflicting empirical data

Although the Schultz cell recording data has shown to be robust across species, some cell recording data suggest that the firing patterns of the DA cells in response to reward are somewhat more complicated. Cell recording evidence suggests that the DA signal may be sensitive more to reward probability than reward magnitude as the Schzult data originally suggested (Tobler et al., 2005).

Another conflicting piece of data comes from operant conditioning tasks involving reward devaluation. Model-based accounts predict immediate sensitivity to reward devaluation whereas TD-learning predicts devaluation only after repeated pairings with the action performed. The effects of devaluation reported in the literature are mixed, and depend upon context and the amount of training.

The goal of the work in Chapter 2 is to develop a reinforcement learning algorithm that can explain the conflicting empirical data that confronts existing algorithms. By adding memory and context sensitivity to a model-based approach the devaluation data can be explained. This model-based approach uses the DA signal as a

measure of surprise and salience based on a Bayesian measure of the probability of events occurring. This method also is able to explain both the standard Schultz cell recording data and the data showing sensitivity to reward probability.

Neuropsychiatric explanations of the DA signal

Dysregulation of the DA cells in the VTA has become a major interest in neuropsychiatric explanations for multiple mental disorders. Interest in the role of upregulated dopamine in psychosis started with observations of acute psychotic symptoms in usually non-psychotic individuals after ingestion of dopamine agonists (Harris & Batki, 2010). The same drugs also worsen or cause reemergence of symptoms in patients whose psychosis was otherwise controlled or in remission (Angrist et al., 1980). In vivo measures of dopamine have also showed significant increases in schizophrenics particularly those showing psychotic symptoms (Lindstrom et al., 1999; Abi-Dargham et al., 1998).

Kapur (2003) theorized the increased endogenous dopamine present during psychosis was causing “aberrant salience.” This was based on the motivational salience model of the DA system developed by Berridge and Robinson (1998). Burst firing in a normally-functioning DA system will signal salient cues that attention should be focused on and serve to motivate behavior to maximize reward or minimize punishment. When the DA system is dysregulated these signals of salience are out of sync with any

normally-linked context. Kapur (2003) outlined three basic stages of the development of delusions: stage one is a heightened state of awareness, stage two is a strong drive to make sense of this new awareness, and, finally, in stage three a person experiences relief when the new awareness is explained through the development of a delusion. Hallucinations often appear in stage three.

This multistage progression of the positive symptoms of schizophrenia fits well with the idea that the disorder is driven by a chaotic signal of salience unconnected with perceptual experience. The DA signal is telling the individual that something important and related to them is happening. This salience signal normally is a cue to motivate behavior so the individual will be driven to find an explanation for the experience and find some action they can take in response. Given a strong enough signal, and enough time, an individual will find an explanation for it, either through random correlations in the environment or delusional thought processes. This multistage process, driven by an aberrant salience cue, is also supported by evidence that psychosis rarely develops after a single administration of an amphetamine or other stimulant, but rather after multiple ingestions (Yui et al., 1999).

The idea of disruption of the DA signal causing aberrant salience can also be applied to other mental disorders. OCD was once thought to be mostly driven by dysregulation of the serotonin system but now there is mounting evidence that dopamine is playing a major, if not the major, role in its symptoms.

For example, Szechtman and Woody (2004) proposed a model of OCD based on disruption of a “security motivation system.” Their conception of OCD fits well with the incentive salience model of dopamine. The DA firing is signaling surprise and salience, the exact opposite of the perception required to terminate activation of the security motivation system. In fact, a signal of surprise and salience is likely to activate the system in and of itself.

Szechtman & Woody (2004) has developed an OCD model in animals using the injection of a D2/D3 agonist Quinpirole. The behavior of animals under the Quinpirole, in contrast to saline injection, gives us a great opportunity to look at differences between the hypothesized roles of dopamine. The reward-prediction error model of TD-learning makes different predictions about how behavior will change based on injection type when compared to the aberrant salience model. The second major section of this thesis (Chapter 3) looks at the behavior predicted by both models and compares it to the actual behavior of the animals. Overall we find evidence that a simple reward-prediction model does not explain behavior and that some form of salience is being signaled by the DA system.

While psychosis and other mental disorders emerge with extreme dopaminergic dysregulation, there is strong evidence that smaller differences in endogenous dopamine levels that exist among individuals have measurable effects on cognition (Cools et al., 2009). There are many aspects of normal day-to-day thinking that might find their explanation in these random fluctuations of the DA system. Magical thinking and

superstitious behavior are similar, if less severe, to the same kind of thought processes we see in the development of psychosis: heightened salience becoming associated with a random environmental factor or contextual cue. Personality disorders such as schizotypy are also likely to be linked to a dysregulated DA system and may find their explanations in slight differences in endogenous dopamine levels.

If these traits are linked to differences in endogenous dopamine we should be able to correlate measures of traits to behavior on learning tasks known to be affected by dopamine. We can also use them to help look at differences between the proposed reward-prediction error role of dopamine and that of aberrant salience. The last major section of this thesis (Chapter 4) looks at measures of personality traits and belief systems that might correlate with endogenous dopamine levels and compares them to behavior on tasks known to be sensitive to dopamine as well as several context-dependent reinforcement learning tasks that let us further compare the reward-prediction model of dopamine to that of our probabilistic model of contextual memory and incentive salience. Once more we find that the reward-prediction model is not adequate to explain the results and some signal of salience is likely to be linked to the DA system.

Chapter 2

Introduction

Understanding the principles governing animal learning is a challenging task. Formal theories of learning from both psychology and machine learning have had a profound influence on our current views as to how animals learn to think and act. Early attempts by associationist theorists to explain performance in simple learning paradigms such as Pavlovian and instrumental conditioning focused mostly on the direct relationship between stimulus and response with little interest in the underlying mechanisms within the organism itself (Thorndike, 1932), though some theories postulated roles for motivational factors such as drive reduction or need satisfaction (Hull, 1943). More powerful models of reinforcement learning and optimal control from the fields of artificial intelligence and machine learning have gone well beyond these simple conditioning paradigms to explain sequential action selection in the face of infrequent and probabilistic reward. Undoubtedly the most influential of these is Sutton and Barto's (1998) TD learning algorithm, which models many aspects of animal learning. At the heart of the TD-learning algorithm is the assumption that one maintains an ongoing estimate of expected reward, and that learning is driven by the reward-prediction error (TD-error), or difference between expected and received reward, on a moment-to-

moment basis. One of the most exciting developments in this field has been the discovery that the phasic firing of dopaminergic neurons in the ventral tegmental area correlates with this TD-error signal (Schultz et al., 1997; Schultz, 1998; Dayan & Abbot, 2001). Thus it has been hypothesized that the neural basis of reinforcement learning may be well modeled by TD-learning. However, not all data are consistent with the TD-error theory of the dopamine signal. For example, phasic bursting of dopamine neurons is observed not only when an unexpected reward is delivered, or when a CS signals a reward, but also when an unexpected or novel stimulus occurs (Horvitz, 2000); moreover, the dopamine burst occurs even before the animal has had time to make an eye movement toward the stimulus to analyze what it is (Redgrave et al., 1999). Also, dopamine has immediate effects on behavior in addition to its effects on learning (Horvitz, 2000; Redgrave et al., 1999; Berridge & Robinson, 1998), suggesting it is conveying more than just an error signal that drives learning of reward expectations. For example, enhancing dopamine affects attention, working memory and the probability of certain motivated behaviors (Phillips et al., 2008), thus appearing to influence the motivational state of the animal. An alternative view is that the change in firing rates of dopamine cells is not signaling a prediction error, but rather a measure of the probability of a reward, thereby conveying surprise and salience (Berridge & Robinson, 1998; Gray, 1982; Smith et al., 2006). This formulation of the dopamine signal is in line with work by Berridge and Robinson that described dopamine cell firing rates as corresponding to incentive salience, signaling

motivation or “wanting” of reward rather than its hedonic value or learning new likes and dislikes. This notion has been used in a model-based account of instrumental conditioning put forward by Smith et al. (2006).

There are two major ways in which modelers have attempted to capture the way animals learn what series of actions should be performed to get maximal reinforcement. One way is to learn estimated state values (“cached values”), as in the TD-learning algorithm. For example, a location in a maze one step from a rewarded goal state, after learning, is associated with a much higher value than the starting position or a location at a dead end. More generally, the “value” of being in a given state is defined as the total amount of reward expected if an optimal series of actions was taken from that point onwards, with distal rewards being devalued relative to proximal rewards. During the learning process, the model incrementally builds up a value estimate for each state. This simplifies the problem of which action to choose next. Rather than considering all possible chains of actions leading to reward, the animal simply chooses the action leading to the next state with the highest value. In contrast, in model-based reinforcement learning, the animal constructs a detailed internal model of the world by learning to associate states with actions and state transition probabilities, which states are associated with direct receipt of reinforcement, and which are terminal states (e.g. the end-point of a maze). In this case, action selection involves forward simulation of the world model to evaluate chains of state-action pairs and their outcomes, rather than the one-step look-up

of cached values associated with environmental states used in TD algorithms. As mentioned above, the TD-learning model postulates that dopamine signals the reward-prediction error (TD-error), a term which drives the learning of value estimates. In contrast, in the model-based approach of Smith et al. (2006), the dopamine signal is interpreted as conveying the relative surprise and incentive salience of an outcome.

Two key situations in which model-based reinforcement learning and TD-learning accounts make different predictions are 1) when the reward magnitude changes even though the overall probability of receiving the reward is unchanged, and 2) in a reward devaluation paradigm; for example, when a cue normally associated with food delivery in a hungry state is instead given while the animal is satiated.

In the case of changes in reward magnitude with constant reward probability, TD accounts of dopamine predict an increase in dopamine signal strength due to the increase in the amount of unexpected reward, whereas the model-based account of Smith et al. (2006) predicts no change in dopamine signaling as the probability of receiving a reward is unchanged, even though the amount of reward may vary. Tobler et al. (2005) measured the firing rates of neurons in the ventral tegmental area of Macaque monkeys in response to a reward delivery. In each of three conditions, a different visual cue served as the conditioned stimulus, while a juice reward of different magnitudes ranging from 0.05 ml to 0.5 ml served as the unconditioned stimulus. The reward was delivered in each condition fifty percent of the time. Therefore, the magnitude of the expected reward

varied across conditions by up to tenfold but the probability of receiving a reward was 0.5 across all conditions. If a dopamine neuron's firing rate signals a reward-prediction error, then the magnitude of change in firing rate should increase with the magnitude of the reward. However, Tobler et al. found that the firing rate did not change across conditions in response to the reward delivery, which suggests that the dopamine cells are actually sensitive to reward probability. The response to the CS did, however, increase with the magnitude of the reward. This may mean there is still some sensitivity to magnitude or, as we suggest in our model of the DA signal, the estimated reward probability is always based on the total maximum reward expected over time.

In the reward devaluation paradigm, as mentioned above, an action that previously led to a highly valued reward (e.g. food delivered when hungry) is then experienced in a devalued form (e.g. food delivered when satiated). If the animal is sensitive to this reward devaluation, it should then become less likely to choose that action to receive the devalued reward. Model-based and TD-learning accounts make different predictions as to whether sensitivity to reward devaluation is immediate or must be learned. The TD-learning account of this phenomenon relies on the animal repeatedly experiencing the pairing of the previously rewarded action with the devalued reward with an action choice before the prior state leading to that action can be devalued. In other words, some number of learning trials is required to update the value associated with the state that previously triggered this action. In contrast, model-based approaches predict

that the animal does not need to experience the new reward paired with the action in order to adjust action policy. The animal only needs to experience the devalued reward once, and update its association between the goal state and the reward. Action choice always involves forward simulation of the model up to the point of reaching the reward, at which point, if the reward is devalued, that action path will be less favorable. The critical question, therefore, is whether the animal shows sensitivity to reward devaluation immediately, as predicted by model-based accounts, or only after experiencing the devalued reward repeatedly paired with the action, as predicted by TD-learning.

Reward devaluation has been studied in animals using instrumental learning tasks. Animals are trained to perform an action to get a reward, and then in another context that same reward is devalued either by pairing it with a negative stimulus or altering the state of the animal (e.g. its degree of satiety). The animal is then returned to the original task and run under extinction conditions. If the animal immediately performs the learned response less frequently than one that did not experience the devalued reward, then immediate sensitivity to reward devaluation can be inferred. If it is performed just as frequently, then sensitivity to devaluation is not immediate, and must be learned. The effects of devaluation reported in the literature are mixed, and depend upon context and amount of training. For example, moderately trained rats show immediate sensitivity to devaluation, consistent with model-based accounts. On the other hand, extensively trained rats are insensitive to devaluation (Daw et al., 2005; Dickinson & Balleine, 1994;

Dickinson & Balleine, 2004). The lack of sensitivity to devaluation in the extensively trained animals could be taken as support of the TD-learning account. However, the devaluation effect is heavily modulated by context; specifically, it is only in the case of a change in context that animals sometimes show insensitivity to devaluation (Dickinson & Balleine, 2004), a finding that is not well handled by any current model.

Some computational modelers have attempted to reconcile the conflicting behavioral data from reward devaluation experiments by proposing that animals may use two different systems, a slower but more accurate model-based system and a faster but more approximate cached value system. While some have proposed these systems as working cooperatively, more recently it has been proposed that these systems are independent and actually compete with each other. Daw et al. (2005) demonstrated that this competitive system can explain much of the behavioral data on reward devaluation, if the more computationally costly model-based algorithm is used early on when uncertainty is high, and after training when uncertainty is low the more efficient cached value algorithm is used. However, this account relies on TD-learning and, as mentioned earlier, this account does not match the entire cell recording data available.

We propose instead a single system that explains the observed responses of dopamine neurons across a wide range of conditions. We build on the model-based approach of Smith et al. (2006) by incorporating a contextual memory system into the model. When the model encounters an outcome that is within the expected distribution of

outcomes it has experienced in the past, it simply generalizes its past knowledge and acts according to what has worked well in the past. On the other hand, if the experienced outcome is well outside its expectations, the model forms a context-specific memory for this event and can treat it as a special case in future. Thus, the proposed algorithm employs a model-based approach to instrumental learning which integrates context as a major element of learning reward value. In our account the conflicting data on devaluation is not due to shifts between two different neurological systems, but rather one system that can discriminate between generalizable differences in expected and actual consequences, and changes that are context-specific. The dopamine signal contributes to this differentiation by signaling significant departures from expected reward probability, thus serving as a signal of surprise and salience. This approach can explain the conflicting behavioral data on reward devaluation within a single system. Moreover, our proposed role for dopamine is a better match for cell recording data that shows sensitivity of dopaminergic cells to reward probability, salience and novelty.

Our approach expands on previous model-based accounts by integrating a contextual memory system into a Bayesian reinforcement learning framework. We also build on the notion of dopamine as a signal of incentive or motivational salience (Berridge & Robinson, 1998) by proposing a specific computational utility for this signal. Most reinforcement learning methods restrict the information associated with a terminal (goal) state to just the magnitude of the reward received. However, empirical evidence

suggests that animals use contextual cues to associate the same reward with different values; for example, in the context of a satiated state, a food reward has less value. Moreover, two different rewards (e.g. food vs. water) may have different incentive values in a given context. These contextual cues can be used to either generalize an experience across multiple contexts or to associate a particular experience to a particular context. To capture this ability, the model updates two internal estimates of a particular reward's value. The first is a general reward value that gets updated with exposure to that reward across all contexts. This general reward value can be used when there is limited experience with the specific reward, or when it is encountered in a new context. A second value is updated only within the specific context in which the reward is being experienced. This allows for learning how a reward value differs between contexts.

We propose an expanded role for the dopamine signal, extending the framework of Smith et al. (2006) which views dopamine as a signal of surprise and significance. In our model, dopamine neurons exhibit increased firing rates in inverse proportion to the probability of the reward value received. The lower the probability of a given reward value, the larger the increase in firing rate. In our model, the size of the DA signal controls the relative update of the general and context-specific reward values. A large change in firing rate acts as a cue that the experience of the reward in the current context is outside the expectations of that same reward in other contexts and, as such, learning is isolated to the context-specific reward value estimate. A smaller DA signal means that the

experience of the reward is likely context-independent and, as such, learning is weighted more towards the general reward value estimate. The expected reward probability is easily computed since our algorithm uses a Bayesian inference approach to reinforcement learning (see Appendix for details). There is growing empirical evidence that animal learning and decision making appear to follow Bayesian optimality principles (Knill & Pouget, 2004; Courville et al., 2006).

My model is not unique in attempting to explain devaluation through state-dependent learning. Reddish et al. (2007) proposed a similar algorithm introducing a state-classification system into a temporal-difference learning paradigm. According to this model all learning is done based on the basic Sutton and Barto TD algorithm. But each observation is linked to the same state. Then during extinction when there is a context shift and no reinforcer there is no devaluation because it is in a new context. While this context development shares some overlap with the model that I have developed there are some distinct differences.

The first is that Reddish et al. (2007) still rely completely on a model-free TD approach to learning. The dopamine signal is still interpreted as purely a prediction-error. The Reddish et al. model then will also have difficulty explaining DA cell recording data showing that the DA signal seems sensitive to probability rather than purely prediction error. Another problem is that the Reddish et al. model does not provide for a way to generalize across context. The lack of devaluation during the extinction trial is caused by

simply not having a reinforcer present in that context. Therefore, this model can not readily explain the fact that only moderate exposure to the reinforcer leads to devaluation. This same problem presents itself with latent inhibition where the model incorrectly predicts that there will be no effect for prior exposure to the unconditioned stimulus because there is a context change. Again this is a failure of the ability of the model to generalize across context.

Greshman et al. (2010) expanded upon the Reddish et al. model by defining context in terms of a clustering algorithm. Context is not so specifically defined but rather observations of the environment (including reinforcers) are clustered together based on common traits. This is an improvement over the Reddish et al. model because it allows for generalizing across similar context. Greshman et al. (2010) is able to account for a wider variety of reinforcement data such as latent inhibition. However, the clustering prevents any kind of generalization between significantly different contexts. It also predicts that devaluation should be a factor of only context not training experience. Therefore, the differential effects of devaluation based on the degree of exposure to the reinforcer is not readily predicted by this model. It also uses a reward-prediction error term for the dopamine signal, which fails to explain data showing that the DA signal likely encodes probability.

Model details

One of the major innovations of our model is the use of the dopamine signal to differentiate whether the estimate of a reward's value should be updated in a context-invariant or context-specific manner. This approach allows our model to explain a wide range of reward devaluation behavioral data within a single system. A typical devaluation experiment involves training the animal while hungry to perform one of two tasks, one which gives access to sugar water, and another which gives the animal a food pellet. After a set number of training sessions the animal is then exposed to the food pellet in another context where it is devalued either by causing sickness after consumption of the pellet, or by exposing it to the pellet after being satiated. This experiment can be simulated by using two different Markov Decisions Trees, each tree corresponding to different contextual cues (see Figure 1). The first tree corresponds to the instrumental learning task taking place in context A. In the initial state, either of two actions can be performed, each leading to a different reward value, one corresponding to the "food" reward and another to the "water" reward. Hunger is simulated by more strongly valuing the "food" relative to the water reward. The second tree corresponds to the devaluation period taking place in context B after a moderate amount of training in context A, and simply consists of an initial state leading to a single terminal state where the "food" reward is delivered but has less value. The third tree corresponds to the devaluation period taking place in context B after extensive training in context A.

The model is first exposed to the learning task in context A. During each iteration of the task the model updates its estimate of the probability of the transition between the initial state and each of the terminal states, as well as its estimate of the reward value at each corresponding terminal state. Since this is the first exposure to each reward type, the model updates its estimates of reward value without any contextual cues being taken into account. After a set number of iterations, with less than 8 being used for moderate training and 15 or more for extensive training, the model is shifted to context B. After each iteration it updates its estimates of the transition probability between the initial state and the terminal state. However, since this is the second context that the food pellet has been encountered in, the DA signal becomes important in modifying how the reward estimates are updated.

When a reward is experienced, it is easy to calculate the log probability of that reward value, based on the predictive probability distribution. We propose that the DA signal can be modeled as follows:

$$\Delta f_i = \begin{cases} s \cdot \log\left(p\left(R \mid \max_a r\right)\right) & \text{If R < E[R]} \\ -s \cdot \log\left(p\left(R \mid \max_a r\right)\right) & \textit{Otherwise} \end{cases}$$

Thus, the change in phasic firing rate of a dopamine neuron, $€f_i$, is equal to the log of the probability of the reward experienced in a state relative to the total reward expected if the optimal action policy is followed. This is multiplied by a binary term *s* which is equal to 1 when a state is either significant or surprising, and 0 otherwise. A state is surprising

when the expected experience in a state differs from the actual experience; this could be because of a novel cue or because an expected reward was different from what was predicted. A state is significant if associated with a cue or action that predicts future reward. The sign of the DA signal depends on whether the experienced reward in a state is greater or less than the expected reward at that state. As in TD-learning, this formulation of the DA signal matches a wide range of dopamine cell recording data. For experimental designs that keep reward probability at 100 percent, such as the Schultz (1998) data, our formulation essentially reduces to TD-error. However, for reward delivery schedules that are not 100 percent, such as the Tobler et al. (2005) data, our formulation produces a different firing pattern in response to reward delivery that shows sensitivity to reward probability rather than reward magnitude.

If the current context in which the reward is being experienced is different from the previous context in which it was experienced, the model actually updates two different estimates of the reward value. The first is the estimate of the reward across all contexts, which was also updated during exposure in context A. The second is a reward value associated only with the current context B. The degree to which these two values are updated is weighted by the size of the DA signal. If the DA signal is large then the model heavily weighs the current experienced reward in updating the estimate for the reward value in the current context, but greatly diminishes its effect on the estimated reward across contexts. If the DA signal is small then the current experienced reward

value is used to update estimates for the reward across all contexts and little context-dependent learning takes place.

When only moderate exposure to context A has occurred the probability distribution for different reward estimates is flat, and even large deviations from previous experience do not produce a significant DA signal. Exposure to context B then causes the estimates of the reward value to decrease across contexts. After exposure to a large number of training iterations in context A the probability distribution for reward estimates is narrow and even slight deviations cause a large DA signal, so exposure to context B will isolate most of the learning to a context-dependent estimate of reward value, and no reward devaluation will be experienced in the first context. This performance of the model should closely match animal behavioral data in devaluation paradigms, where moderately trained rats are sensitive to reward devaluation, whereas overtraining abolishes this sensitivity. In addition, the dependence of contextual learning on the DA signal can also explain data from animals where the DA signal is abolished. In animals that have efferent connections from the ventral tegmental area lesioned, reward devaluation is shown even after significant overtraining. This can be modeled by simply setting the DA signal to zero. This forces any exposure to the reward type to only update context-independent estimates of that reward value.

Methods

Two sets of simulations were run to compare our model against behavioral and cell recording data. The first set focused on the reward devaluation literature, looking at devaluation sensitivity as it relates to amount of training and the presence of the DA signal. The second set of simulations compared simulated firing rates of dopamine neurons using our formulation of the DA signal to both the classic dopamine cell recording data of Shultz et al. (1997) as well as more recent data where reward probability is varied.

Reward devaluation

In order to test the model in regards to devaluation, two Markov Decision Processes (MDPs) were set up that correspond to two different contexts. In the first there are two possible actions that lead to two different terminal states with different rewards. The second MDP has only the terminal state, which has a reward that is the same kind as one of the rewards from the first MDP but is devalued.

The model shown in Figure 1 was initially trained in context A and then in context B. The reward in state S4 (context B) is the same kind as given in state S2 (context A) except that it has been devalued by having a reduced mean value and standard deviation. The number of training sessions in context A, the mean and standard deviations of the reward's value function, and the change in the reward's value function for context B are the major free parameters that define the devaluation paradigm. A schematic of the

training process using MDPs is shown in Figure 1. This set-up mimics many behavioral experiments where an animal is trained to perform and choose between two actions, each associated with a different reward, e.g. sucrose water or a food pellet. After training the animal is exposed to one of the two rewards, e.g. the food pellet, in another setting in which it is devalued. The animal is then put back into the initial setting and tested to see if its action choice reflects the devalued reward.

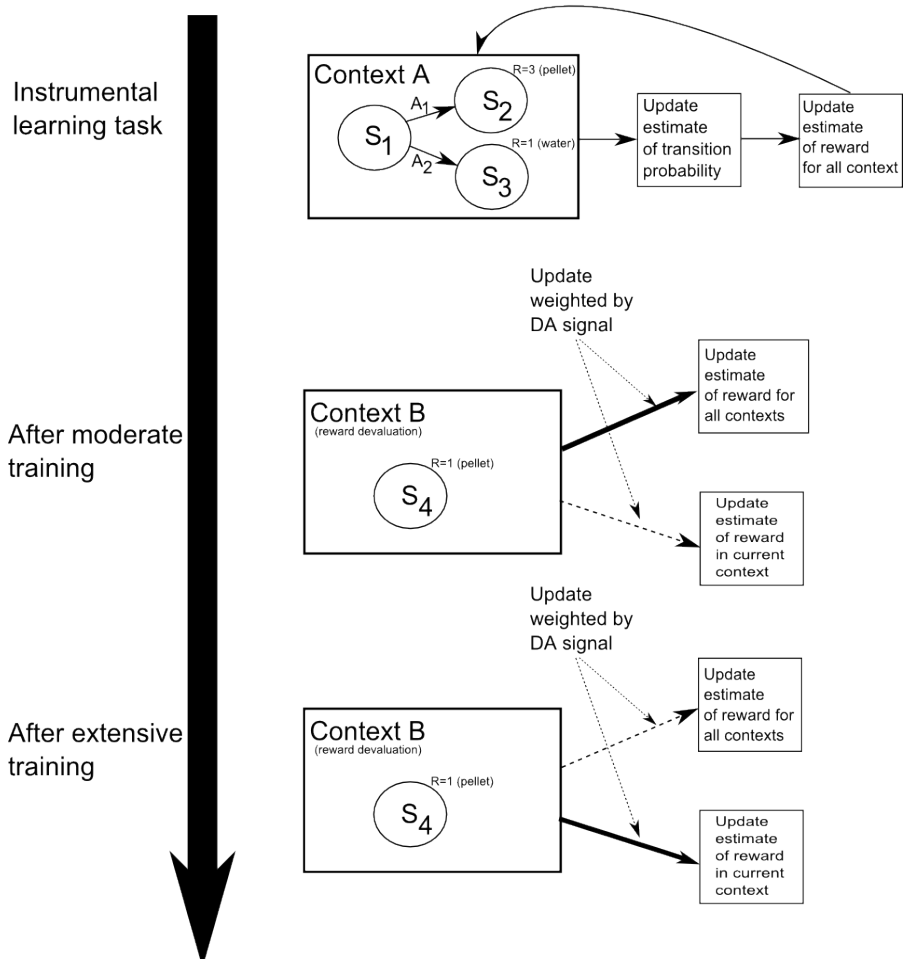


Figure 1: A schematic of the devaluation simulations used to compare models. The model is first trained in context A where it learns two action choices each of which leads to a different reward. Depending on the simulation the model is moved to context B after either only moderate training or extensive training. In context B the first reward is devalued, and the DA signal is then used to weight the updating of the new estimate between context-dependent or independent values.

Simulations of the model were run for several different training regimes. The mean value, μ, of the first reward was 2 and of the devalued version of this reward was -2. The mean value of the second reward was 1. The standard deviation, σ, for all the rewards was varied between simulations at 5-50 percent of μ. The standard deviation only had a strong effect at the most extreme ends of the scale; for the results presented here it was set to 10 percent of μ. The first experiment tested the model's sensitivity to devaluation after extensive training. The DA signal was used as a cue to create a separate contextual memory when the reward value was far outside the expected range.

Reward devaluation after extensive training

For extensive training, the model was trained on the MDP shown in context A (Figure 1) for 10-12 iterations, with one action choice per iteration A. The model was then trained for an equal number of iterations on the MDP shown in context B to simulate devaluation. After training in each context the likelihood of each action being performed in context A was calculated, to assess sensitivity to devaluation. A change in this likelihood after context B exposure signals that reward devaluation has occurred.

Reward devaluation after moderate training

The second simulation was designed to capture devaluation sensitivity in only moderately trained animals. In this case, the devaluation of the reward is still within the expected range of the estimated value so the updated reward value should be context independent. For moderate training the simulation was run through the first MDP until action A was selected a preset number of times (5-8). Then an equal number of iterations were run in the second MDP. Once more the likelihood of each action choice in context A was calculated after each training session, to test for devaluation.

Lesioning the dopaminergic pathway

The final simulation was designed to investigate the effects of lesioning the efferent dopaminergic connections from the ventral tegmental area. In animals this creates devaluation sensitivity regardless of the amount of training experienced. To simulate this we used the same method as simulation 1a but we set and kept the DA signal at zero.

DA sensitivity to reward probability

In the final set of simulations, we assessed how well our formulation of the dopamine signal accounts for the firing patterns of dopaminergic neurons, including the

classic data of Shultz et al. (1997) which are well modeled by the TD-error model, as well as more recent data which present a challenge for the TD-error account of dopamine.

DA signal with 100 percent reward delivery probability

The first set of simulations was designed to measure the DA signal in a conditioning task similar to those used by Schultz et al. (1997). The conditioning task was modeled as an MDP where advancing through states is dependent on time rather than action choice. A cue is presented after the initial state and then when 10 seconds have passed (with each second being represented by one state in the MDP) a reward is delivered. This constitutes a single trial, with tentrials being run in total. The DA signal is calculated at each second in the trial. The simulation was also run with the reward not being delivered in the last trial to test the DA signal for changes in response to unexpected reward delivery failure.

DA signal with 50 percent reward delivery probability

The next simulation focused on a conditioning task from Tobler et al. (2005) which showed that if reward delivery probability is set at a constant 50 percent the DA response to the reward does not change even if the magnitude of the reward value varies by ten-fold, which is not predicted by the standard TD-error approach. To model this we used the same style of MDP from simulation 3.2.1. There were two possible reward

magnitudes of 0.05 and 0.15, and each reward had a 50 percent chance of being delivered at the end of each trial. A total of 20 trials were used and the DA signal was calculated for each second of the trial.

Results

In each of the reward devaluation simulations the following statistics were tracked: the actual reward delivered at the terminal state, the model's estimate of the reward, and the 95 percent confidence interval for the estimated reward value. The confidence interval is a good guide for sensitivity to reward devaluation. If the confidence interval is wide (such as early on in training) any experience of the devalued reward will be far less surprising than if the confidence interval is narrow (such as after extensive training).

Reward devaluation after extensive training

The results from the simulations that used extensive training before exposure to the devalued reward are shown in Figure 2. The most important features of the reward learning graphs are the confidence interval and whether any learning takes in the general estimate after exposure to the devalued reward. After extensive training, the confidence in the estimate of the reward value was initially wide but narrowed quickly. When the devalued reward was presented the general estimate for the reward stopped being updated, and instead the new learned value was stored in a separate context-specific memory. This context-dependent learning was triggered reliably after exposure to 10-15 iterations of action A, as long as the variance of the reward was less than 30 percent of the mean. Since any change in reward expectation after exposure to the devalued reward

was stored as a separate memory, the likelihood of each action choice in context A did not change after exposure to context B. Thus, like extensively trained animals, the extensively trained model showed insensitivity to devaluation.

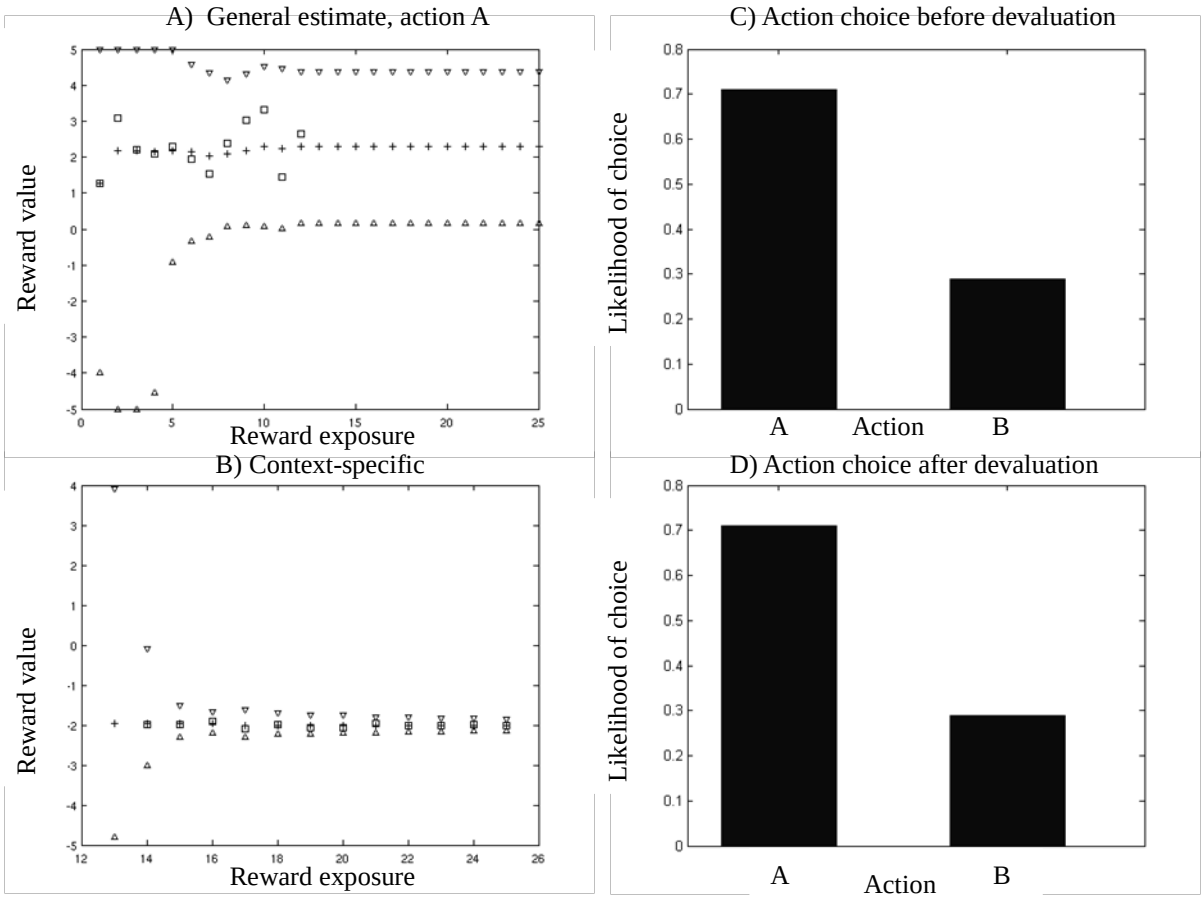


Figure 2: Model results showing devaluation insensitivity after overtraining. Plot A shows the actual reward received (squares), the model's estimate of reward (pluses), the 95 percent confidence interval (triangles) of this estimate, and the general estimate each time action A is chosen. After the 12th trial, the model is moved to context B. Importantly, the estimated reward value does not change in this context after the switch. Plot B is the reward estimate after exposure to context B; the devaluation of the reward is stored

separately in this memory rather than being generalized as in the original context A. Plot C shows the likelihoods of choosing actions A and B after 12 iterations, at which point action A is highly favored. Plot D shows the likelihood of choosing actions A and B after exposure to context B, showing no change in likelihood and thus that devaluation did not occur.

Reward devaluation after moderate training

The results for the second simulation designed to assess devaluation sensitivity after only moderate exposure to the reward in the first MDP are shown in Figure 3. Again the most interesting features of the learning curve are the confidence intervals and whether the general estimate changes after exposure to the devalued reward. In this case, the confidence interval is always wide initially, and does not reach a point of narrowing before devaluation takes place. The devalued reward is therefore not seen as lying outside the model's expectation and the general estimate for the reward value decreases. With only moderate exposure to the reward in context A the confidence in the estimate is low enough that exposure to the devalued reward is not surprising so learning in context B is used to update the general estimate for that reward. Once training has stopped with the devalued reward, the value of action A is less than action B, so the model's likelihood of choosing action A goes down significantly compared to its likelihood before exposure to the devalued reward. This is evidence of devaluation sensitivity. Variance plays a key role in how many trials it takes to eliminate devaluation sensitivity. When variance is low (less than 5 percent of the mean reward value) it can take as few as 5 trials to produce

devaluation insensitivity. As the variance increases so does the number of trials; when the variance is at 10-15 percent of the mean reward value it takes 8-12 trials before devaluation insensitivity emerges. Really large variances can produce a situation where devaluation sensitivity will always be present.

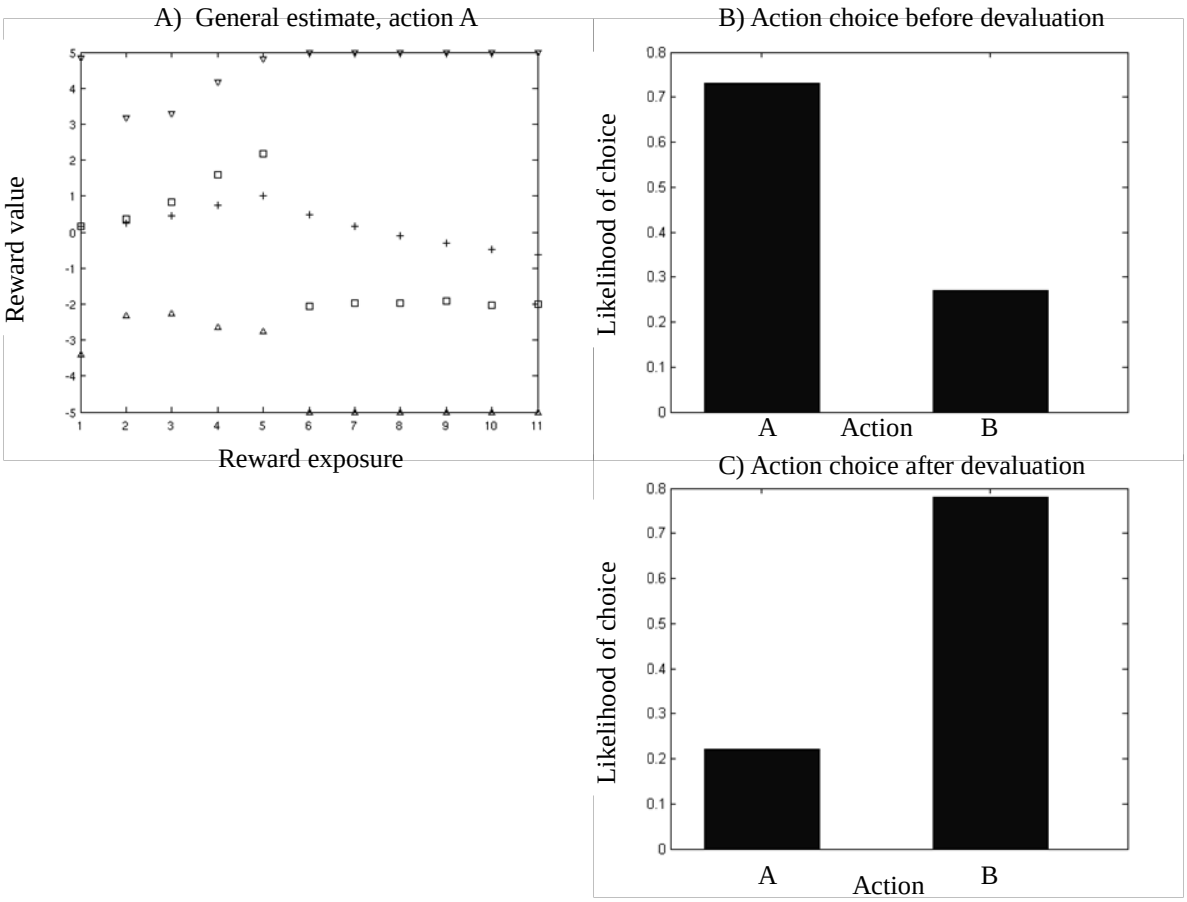


Figure 3: Model results looking at devaluation sensitivity with only moderate training. Plot A shows the actual reward received (squares), the model's estimate of reward (pluses), and the 95 percent confidence interval (triangles) of the general estimate each time action A is chosen. After the 5th trial, the model is moved to context B. Since the

change in reward value is not surprising it is used to update estimates in the original context. Plot B shows the likelihood of choosing action A and B after 5 iterations, showing action A is highly favored. Plot C shows the likelihood of choosing action A and B after exposure to context B. In this case, the change in likelihood is evidence that reward devaluation has occurred.

Lesioning the dopaminergic pathway

Figure 4 presents the results from the final reward devaluation simulation, designed to test the effect of lesioning the DA signal, which in animals causes reward devaluation sensitivity to remain despite extensive training. This is simulated by setting and keeping the DA signal at zero. The main point of interest on this learning graph is when the exposure to the devalued reward starts at trial 12. The reward is far outside the confidence interval but, without the DA signal present to weight learning only in a context-specific manner, the general estimate for the reward value goes down. The estimated value of action A is now less than action B so the likelihood of choosing action A goes down. This is evidence for reward devaluation sensitivity. This sensitivity will remain no matter how much training is given in the original context as long as the DA signal is zero.

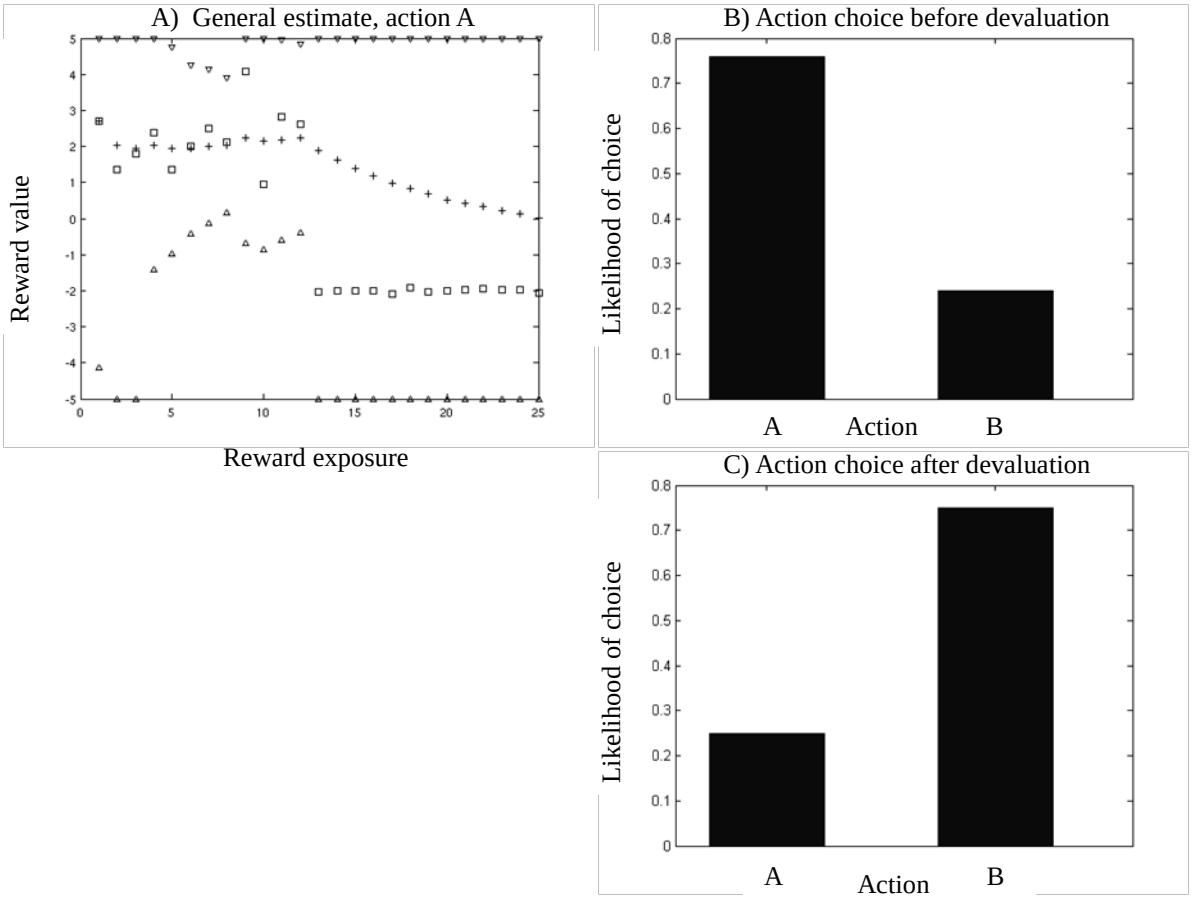


Figure 4: Model results of reward devaluation with the DA signal set to zero. Plot A shows the actual reward received (squares), the model's estimate of reward (pluses), and the 95 percent confidence interval (triangles) of the general estimate each time action A is chosen. After the 12th trial, the model is moved to context B. Since the DA signal is zero no context-dependent learning takes place. Plot B shows the likelihood of choosing action A and B after 12 iterations, showing action A is highly favored. Plot C shows the likelihood of choosing action A and B after exposure to context B; in this case, the change in likelihood is evidence that reward devaluation has occurred.

DA sensitivity to reward probability

DA signal with 100 percent reward delivery probability

The measured size of the DA signal of our model for the simulation using the Schultz (1998) conditioning paradigm where the reward delivery probability is 100 percent is shown in Figure 5. Initially the DA signal only responds to the delivery of an unexpected rewarding US but over the trials the response is transferred instead to the CS which predicts the coming rewarding US. When the same simulation is run, but with no reward being given in the last trial, we see a drop in the DA signal. These results confirm that our formulation of the dopamine signal can account for cell recording data in traditional conditioning tasks as well as the TD-error approach does.

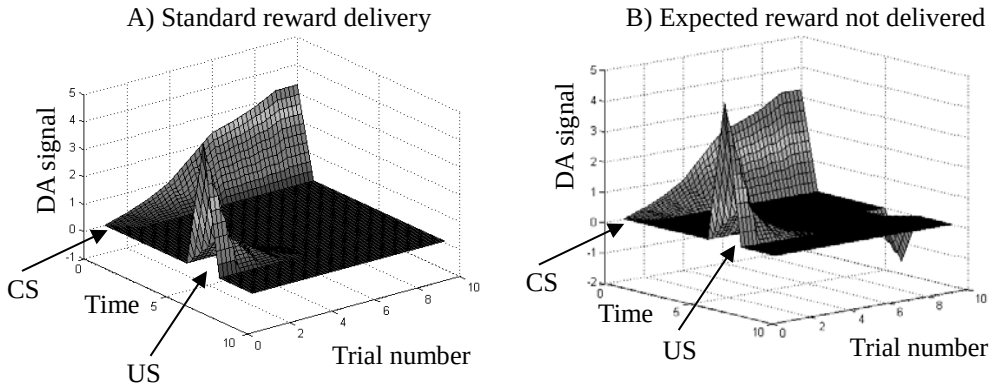


Figure 5: Plot A shows the DA signal across time in a trial, and trial number for the standard conditioning paradigm used in Schultz (1998). The unexpected reward produces a signal that eventually transfers to the CS. Plot B is the same data but, with the last trial not delivering the expected reward, shows a drop in signal.

DA signal with 50 percent reward delivery probability

The measured size of the DA signal in simulations based on the procedure of Tobler et al. (2005) are shown in Figure 6, where reward magnitude increased but the delivery probability of 50 percent stayed constant. The plots in Figure 6 are only from trials in which reward delivery occurred. For both magnitudes the DA signal responds early to the US and is sensitive to reward size. Over time the DA signals in response to the CS, and this signal as well is sensitive to reward size. However, the DA signal after reward delivery quickly settles to the same value for both reward sizes. These results show that our formulation of the dopamine signal behaves similarly to the actual cell recording data, which the TD-error approach cannot account for.

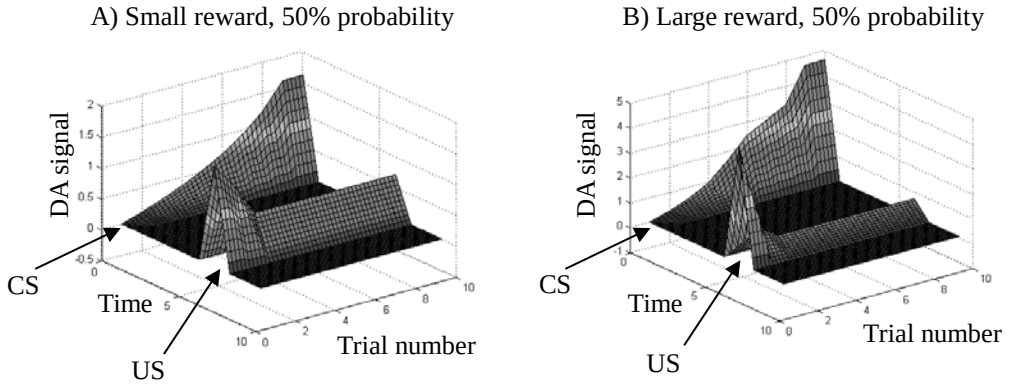


Figure 6: Plot showing the DA signals for 0.05 and 0.15 reward magnitudes delivered with 50 percent probability. Plot A shows the DA signal for the 0.05 reward magnitude delivered with 50 percent probability; plot B is the 0.15 magnitude reward delivered with

50 percent probability. The signal is sensitive to reward magnitude at the CS but not when the US is delivered.

Discussion

We have focused our simulations of our model on explaining a wide range of seemingly contradictory data on dopaminergic signaling and function using a model-based account of reinforcement learning. The key to our model is the formulation of the change in firing rates of dopamine cells in the ventral tegmental area as conveying surprise and salience and using that signal to mediate the formation of context-specific memories of reward values. This formulation of the dopamine signal matches a wider array of cell recording data than does the popular TD-error interpretation.

Model-based reward devaluation

We have demonstrated that model-based approaches can account for the animal reward devaluation literature. Previous model-based approaches had no way to account for devaluation insensitivity after extensive training, while cached value-based models such as TD-learning had a hard time explaining devaluation sensitivity early in training. One approach to this problem put forward by Daw et al. (2005) was to propose two different learning systems in competition with each other: a model-based learning system using pre-frontal cortex structures, and a cached value-based system using the striatal dopaminergic circuits. The model-based system is used early on as it is likely to be more

accurate than the cached value-based system. However, after extensive exposure to the task, the cached value system is just as accurate and computationally more efficient and so is more likely to be used. This explains the devaluation literature as moderately trained animals are still using a model-based approach while extensively trained animals are using the cached value approach.

The model-based approach proposed here differs from that of Daw et al. (2005) in several key ways. By augmenting a model-based reinforcement learning account with a context-specific memory system, our account provides a single system framework that operates at both early and late phases of training. The transition from relying more heavily on context-specific knowledge to knowledge that generalizes across learning experiences falls naturally out of our Bayesian framework, where expected probability distributions over events converge as more experience is acquired. Daw et al.'s model also views the DA signal differently; in their model dopamine only plays a role in the cached value system, and not in the model-based system. In contrast, consistent with the widespread dopaminergic projections throughout the brain, and most prominently to the prefrontal cortex (Oades & Halliday, 1987), our model incorporates the dopaminergic signal within a model-based learning framework, where it acts as a salience signal to modulate context-specific learning. This integration of the dopamine signal with the formation of contextual memories allows the model to develop devaluation insensitivity and explains the more subtle puzzle of the difference in devaluation between moderate

and extensive training within a single learning system. Our approach then tightly links devaluation sensitivity to the DA signal and to confidence in reward estimates. An interesting prediction is that the amount of training needed to produce devaluation insensitivity should be correlated with the variance in the value of the reward delivered. Low variance rewards should lead to devaluation insensitivity much more quickly than high variance rewards. With significant variance in the reward animals should remain devaluation-sensitive even after extensive training. This integration of contextual memory, dopamine and learning offers an opportunity to more deeply expand upon and explore the role of memory in reinforcement learning.

Context, learning and the hippocampus

A key part of our model is the formation of context-specific memories linked to key salient events such as large violations of expectation in reward value. The most likely neural locus for this context-specific learning and memory is the hippocampus. A large body of literature supports a key role for the hippocampus in context-specific memory formation; for example, in contextual fear conditioning (see, for example, Maren et al., 1998; Phillips & LeDoux, 1992). Moreover, there are prominent dopaminergic projections into the hippocampus (Oades & Halliday, 1987) and dopamine plays a key role in hippocampal memory acquisition and retention. Dopamine agonists increase the encoding of recently acquired information, as well as its retention in the hippocampus,

while antagonists inhibit encoding and retention (for a review see Lisman & Otmakhova, 2001). These data are consistent with the role played by dopamine in our model. While a general estimate of reward value can just be cached and applied across many contexts, context-specific value estimates (e.g. food reward delivered in a satiated state) have to be encoded as separate memories with contextual cues retained over considerable time periods. The increase in DA activity in response to reward expectation violations could modulate the encoding of such context-specific memories in the hippocampus.

Dopamine as a reward probability signal

Conceptualizing the dopamine signal as conveying the probability of the experienced reward can capture both the classic cell recording data of Shultz et al. (1997), as well as more recent cell recording data that the TD-error model cannot. Our formulation specifically predicts that the change in firing rate of dopamine cells should be related to the probability of the experienced reward relative to the total expected reward. This is very different from the TD-error account which postulates that phasic dopamine firing conveys the reward-prediction error. Our simulations showed that our model was thereby able to account for the findings of Tobler et al. (2005) when magnitude varies ten-fold but probability stays constant. Future work extending the findings of Tobler et al. to multiple-action and multiple-cue conditioning tasks could further validate the accuracy of our formulation. Future work in both animals and humans

is also needed to explore the relationship between ventral tegmental activation and estimated probability rather than simple reward-prediction error.

The view of dopamine as a signal of incentive salience has a rich history in animal research, and also has been explored for clinical applications in understanding schizophrenia, ADHD and obsessive compulsive disorder (Kapur & Remington, 1996; Swanson et al., 2000; Denys et al., 2004). Reformulating the view of the DA signal as a measure of salience and surprise could potentially create closer integration between research on incentive salience, clinical psychology and the reinforcement learning literature. Previous attempts have been made to link the ideas of TD-learning algorithms based on reward-prediction error to an incentive salience model of the DA neurons. McClure et al. (2003) put forward the idea that the DA response to cues that predict reward could serve as the incentive salience signal. Tsai et al. (2009) built upon this work by adding physiological states as a motivator to change the incentive salience value of a CS on the fly. Both of these approaches used a cached value-based learning algorithm with the interpretation of the DA signal being reward-prediction error. This differs from our approach, where we take a model-based approach to learning, and the DA signal is a direct measure of reward probability that is inherently linked to the concept of incentive salience.

Limitations and future work

There are several limitations to our model. The greatest limitation of all model-based approaches to learning is that they become computationally intractable as the decision trees to be modeled increase in complexity. Planning many steps into the future is challenging, even for human experts. Chunking one's experiences into previously successful action sequences ("habits") or action schemas ("strategies") is one of the ways in which many organisms appear to deal with the formidable task of planning many steps into the future. Seamlessly integrating a means of chunking frequently used action sequences into a model-based system is an important direction for future work.

Cached value systems, once fully trained, have the advantage of being computationally very efficient. Within such a framework, habitual responses can be viewed as an effect of caching values and only looking ahead one step at a time. However, value is not the only thing that can be cached to save resources. Action policy can also be cached. Thus, both model-based and cached value systems could be extended to incorporate cached action sequences, or a hybrid of the two could be employed. At any given time it would be possible to store the current estimated optimal action policy and a choice could be made about whether to use the last known good policy or to create a new one by running through the internal model. Using a cached action policy is computationally cheap, and would behave very much like a habit system. This kind of action caching could occur while learning takes place, but would only be useful if the

confidence in reward predictions was very high, or if there were not enough resources to devote to processing the model. Thus cached action sequences would only be used after extensive exposure to the task, and in tasks with minimal variation in reward value or under significant computational load.

There is evidence that lesions to the pre-frontal cortex have dissociative effects on devaluation sensitivity regardless of the level of training animals receive. More specifically, lesions of the prelimbic prefrontal cortex appear to disrupt devaluation sensitivity in moderately trained animals, while lesions of the infralimbic prefrontal cortex disrupt devaluation resistance in highly over-trained animals (Killcross & Coutureau, 2003). Our model does not yet address the results of these lesion studies but the answers could be found in the cached action hypothesis. Lesions that disrupt the ability of the organism to actually perform model lookup for action policy could force the use of chunked action sequences, producing devaluation insensitivity regardless of training, while disruption of areas that store cached action plans could create the need to always use model lookups for action policy thus always creating devaluation sensitivity regardless of training.

Conclusions

We conclude that model-based reinforcement learning accounts can explain behavioral data that was previously considered to be inconsistent with such formulations.

We also demonstrate that conceptualizing the DA signal as a measure of probability, surprise and salience, rather than as a reward-prediction error signal, can account for both classic cell recording data also accounted for by the TD-error model as well as more recent data that the latter model does not match.

Chapter 3

Introduction

Dysregulation of the dopamine system has been linked to a wide range of disorders including schizophrenia, ADHD, Parkinson's disease and addiction (for review, see Iversen & Iversen, 2007). Understanding the actual role of dopamine is extremely important in elucidating the causal mechanisms underlying these disorders. OCD has historically been attributed to dysregulation of serotonin (Insel et al., 1985). However, mounting evidence indicates that dopamine (DA) dysregulation plays a major role, and may play the primary role in at least a subset of individuals with OCD (Goodman et al., 1990). Szechtman et al. (1998) developed an animal model of OCD in rats using a series of injections with Quinpirole, a dopamine D2/D3 receptor agonist. After the injection, rather than assessing performance on a specific task, rats are allowed to roam freely on a table with a few landmarks. Quinpirole-injected rats show a significant increase in activity, but more importantly this activity is coordinated and involves repeatedly returning to one or a few "home" zones and initiating "checking" behavior such as contact and ritual-like movements (Szechtman et al., 1998). These ritual-like repetitive behaviors have a similar spatio-temporal structure to human OCD compulsions.

The Quinpirole injections and free exploration data from these experiments offer a fascinating and unique way of testing the role of dopamine outside of classic

reinforcement learning paradigms. The early animal experiments looking at dopamine regulation focused on reinforcement learning models. Computational modeling of reinforcement learning has been strongly linked with the role of dopamine. Modelers have developed two basic approaches: some investigators view dopamine as conveying a reward-prediction error, while others have formulated the viewpoint of dopamine as conveying a salience signal (Dayan & Abbot, 2001; Berridge & Robinson, 1998; Gray, 1982; Smith et al., 2006). The former approach is referred to as temporal difference or "TD-learning," as the learning algorithm makes use of temporal differences between predicted and received rewards at successive time points ("TD-errors") to update predicted reward values. Both of these approaches have been shown to correctly model a wide range of behavioral data from reinforcement learning and classical conditioning experiments.

However, these two classes of models make very different predictions under conditions of dysregulated dopamine. Incentive salience models of dopamine postulate that DA dysregulation affects the animal's perception of salience and surprise, while temporal-difference (reinforcement) learning approaches predict that DA dysregulation affects the perception of estimated and received reward values. To test these predictions, dopamine can be manipulated via administration of DA agonists and antagonists. For example, Quinpirole injections modulate the firing rate of the dopaminergic neurons (Korotkova et al., 2003), and induce repeated revisiting of one or two landmarks in the

environment, a phenomenon that meets the criteria of compulsive checking behavior in rats (Szechtman et al., 1998).

The TD-learning model of dopamine and the incentive salience model of dopamine have very different explanations for this phenomenon. According to the TD-learning paradigm, increasing DA should heighten both the reward prediction and the perception of reward received. The increase in reward prediction would serve to motivate goal-directed behavior, thus increasing activity overall. More importantly, normal behavior, such as “checking” and “exploring,” are perceived as being unexpectedly rewarding. Thus the behavior is reinforced on a constant basis leading to significant increase in specific behaviors. Thus the TD-error model explains compulsive behavior in rats as being caused by increases in perceived reward of the activity being performed spontaneously. The incentive salience model, however, views increased dopamine-mediated activity as signaling surprise and salience. Accordingly, by mimicking the actions of dopamine, Quinpirole becomes an agent for creating aberrant salience.

To understand how aberrant salience might cause compulsive behavior it is worth looking more deeply into the symptoms of OCD and some of the proposed underlying causes of the disorder. While a significant part of OCD symptomatology in people involves disruption in cognition and thought processes, the main area of interest for our purposes of comparison to animal models is the external behavior. OCD is marked by repetitive, relatively stereotyped behaviors (compulsions) that the person feels compelled

to think or perform but recognizes as irrational or excessive (Goodman et al., 1990). The compulsive behaviors exhibited by those with OCD tend to fall into a few specific categories: (a) excessive checking activities, characterized by repeatedly performing actions supposedly related to security, orderliness, or accuracy; (b) avoidance behaviors, which are "activities engaged in to avoid feared objects, places, or situations"; and (c) compulsive washing and cleaning, generally of hands but sometimes also washing of clothes, teeth-cleaning, or the cleaning of possessions or parts of the home (Szechtman & Woody, 2004; Reed, 1985). One of the more seemingly contradictory aspects of OCD is that patients know intellectually that the compulsions are irrational and do not wish to perform them (Foa et al., 1995) but nevertheless do so. There seems to be some underlying mechanism unrelated to higher-order cognitive processing that prevents stopping the behavior. Szechtman and Woody (2004) developed a model of OCD based on disruption of a "security motivation system" that takes into account the nature of compulsions and the failure to stop the actions despite being aware of their irrationality.

The Szechtman-Woody model proposes an adaptive cognitive system selected by evolutionary pressure to help animals avoid potentially negative or harmful situations. They refer to this system as the "security motivation system" and identify four key properties of this system: 1) it is tuned to predicting potential danger using often subtle cues and incomplete information; 2) it is activated readily even for unclear or partial dangers, potentially even at the first signs of the slightest chance of danger; 3) it is

oriented towards action, inducing often automatic behaviors that are geared towards avoiding danger such as checking and surveillance; and 4) it is separate from other avoidance systems such as pain avoidance.

An important aspect of the security motivation system is that it induces the exact kinds of behaviors that are exhibited compulsively in OCD. If the security motivation system is indeed inducing these behaviors, then it follows that one of the core deficits underlying OCD is an inability to disengage the system once it has been activated. Szechtman and Woody propose that in normal, intact animals, the security motivation system is part of a feedback loop and is down-regulated by what they term "yedasentience," a "feeling of knowing." In the context of checking behavior, yedasentience is the sensation that everything is okay and that there is no surprise. This is the primary terminating signal for the security motivation system. In the case of OCD patients, they never experience the perception that everything is as it should be and therefore their security motivation continues to be active and drive security-related actions unabatedly, rendering behavior that is compulsive.

This conception of OCD fits well with the incentive salience model of dopamine. DA is conveying surprise and salience, the exact opposite of the perception that is required to terminate activation of the security motivation system. In fact, a signal of surprise and salience is likely to activate the system in and of itself. Therefore, when Quinpirole is injected the DA signal is conveying this surprise and salience and inducing

the compulsive checking behavior seen in the rats. Since this signal is always present while the drug is active the security motivation system never disengages, and the compulsive behavior continues.

While both the TD-learning and salience models of dopamine offer plausible explanations for the general pattern of behavior in QuinpiroleQuinpirole-injected rats, they make different predictions about specific behaviors under specific conditions. One of the more interesting contrary predictions involves what happens to behavior in the same environment if the Quinpirole injections are stopped and then started again later. In the animal model a saline injection is used for one of the injection periods in the QuinpiroleQuinpirole-treated rats. During the saline injection the rats appear to no longer demonstrate the increased activity or compulsive behavior. When Quinpirole injections are then resumed the behavior returns. While the reward-prediction model and the salience model can both explain this pattern of behavior, each model makes different and specific predictions as to the underlying mechanism that produces it.

In the TD-learning paradigm the compulsive behavior is learned through reinforcement based on a false unpredicted reward signal. During the saline trial the rat learns that the compulsive behavior is no longer rewarding and thus learns not to perform it anymore. However, when the Quinpirole injections resume the unpredicted reward is once more experienced and the compulsive behavior is once more learned through reinforcement. Under the security motivation system and aberrant salience model the

Quinpirole is preventing the termination of the compulsive checking because the rat is experiencing unexpected salience. During the saline injection this is not experienced and the animal learns that frequent checks are not needed and the compulsive behavior is terminated. When the Quinpirole is resumed the aberrant salience is once more experienced and the compulsive behavior cannot be terminated.

The difference between the predicted behaviors of the two models is most apparent after the saline injection when the Quinpirole is resumed. Under the TD-learning model, the animal has learned that the behavior is no longer rewarding and must relearn that it is rewarding. However, with the aberrant salience model the compulsive checking is initiated by the security motivation system and prevented from terminating; thus the salience model would predict the instantaneous emergence of compulsive behavior. In order to assess which model produces the better fit to the data, we require a measure of "learning" during the explorative behavior, so that learning can be assessed right before, during and after the saline injection. The TD-learning model of DA predicts that the introduction of the saline and resumption of the Quinpirole should produce similar but opposing effects, with unlearning following saline, and then relearning the perceived reward following resumption of Quinpirole. The aberrant salience model of DA predicts immediate effects of Quinpirole injections on behavior, and that only after the saline injection should there be any signs of learning. Another important difference between the two models' predictions is in terms of the path taken in between checking events.

Because the TD-learning model relies on cached values associated with each state, which must be learned, the choices made in between rewarding events in turn become predictive of reward, and hence acquire reward value in and of themselves. Therefore, we should see strong similarity between the paths taken when an animal revisits the same location during compulsive checking, and an increase in that similarity over time as the specific decision points become more and more reinforced by the Quinpirole-enhanced reward perception. The aberrant salience model does not rely on the checking behavior being derived from reinforcement learning and therefore the path taken between events should have no particular meaning or value. Even if there is some level of learning going on, the salience model of dopamine relies on a model-based approach rather than the cached values associated with each state along the path to the rewarded location, and would thus be insensitive to the specific paths taken between checking behavior. Therefore, the salience model predicts no strong similarity between paths or any increase in similarity over time.

Experiment 1

Methods

The data for this analysis was taken from Szechtman et al. (1998). In brief, the animal was injected with Quinpirole (0.5 mg/kg) and allowed to locomote on a table with four small objects on it, located at the same fixed locales throughout all trials. The table

was virtually subdivided into 25 numbered squares and each time an animal entered a square the square number and the arrival time was recorded. The departure time when the animal left each square was also recorded. The animal was allowed 60 minutes of free exploration. The location the animal went to the most frequently was referred to as the "home" square for that trial. This procedure was repeated 11 times. The 12th trial used a saline injection rather than Quinpirole. The same procedure was used for data recorded during the saline injection. Two more trials with Quinpirole were then completed after the saline injection. Figure 7 shows an example of the dataset showing the activity of the animals during both the Quinpirole injections and the saline injections.

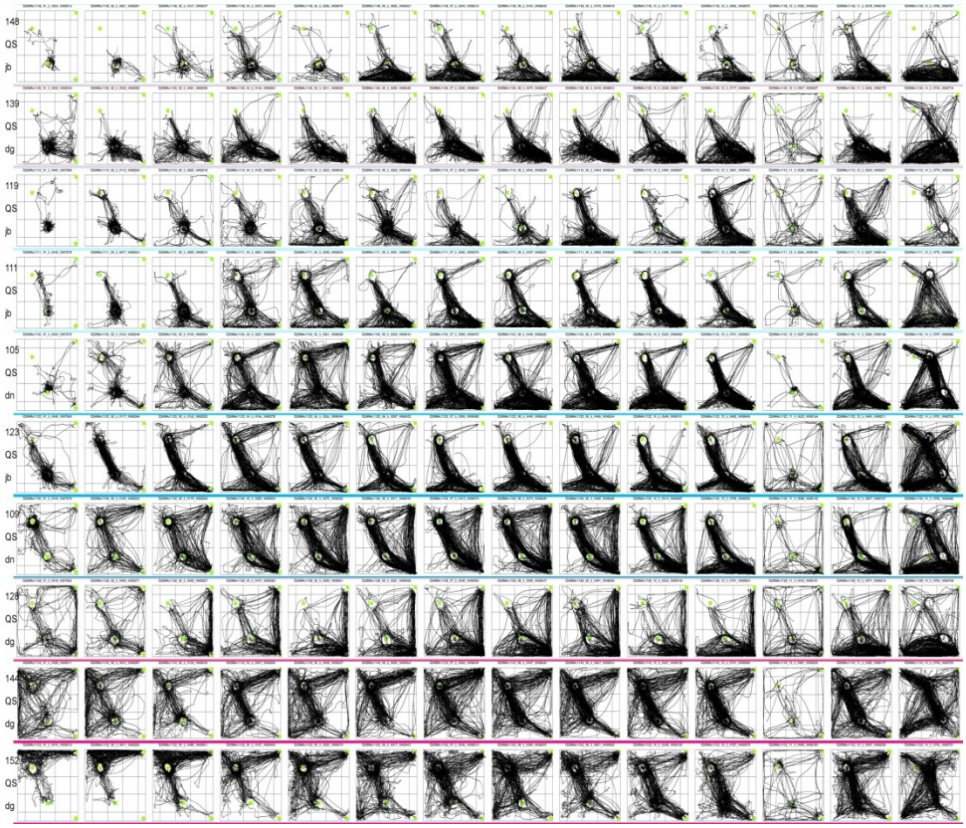


Figure 7: Position recordings for ten rats across Quinpirole and saline trials The figure above shows the path information recorded for ten different rats. Each rat is represented by one row. Each column represents one trial, with trial one commencing just after the first drug injection. The twelfth column is a saline injection trial showing more "normal" exploratory behavior. The dark lines densely connecting two to three locations show how the animal repeatedly revisited a small subset of locations.

The primary trials used for comparison to the computational models are the last Quinpirole trial before the saline injection (Q1), the trial with saline injection (S), and the subsequent trial with injection of Quinpirole (Q2). The compulsive behavior analyzed was checking behavior on the home square. Any time the rat returned to its home square

was counted as checking behavior. The percentage of times the animal returned to the home square compared to the overall number of squares visited was calculated each time it returned home. This helped normalize the count of checking behavior across conditions since the overall activity in the saline condition was significantly less than in either Quinpirole condition. Learning was defined as a change in the percentage of home square visits over time. This was measured by taking the root mean square error (RMS) of the slope, averaged across each point in the trial. This captures multiple learning styles including an exponential style with quick learning overall and a linear style varying slowly over time. A high RMS for a trial meant that there was significant change over time in the percentage of times the animal initiated checking behavior. If no learning took place that percentage stayed constant and the RMS would be close to zero.

The Model

The free exploration behavior of the animals was modeled with both TD-learning and our formulation of a model-based approach using the DA signal as a conveyor of salience and surprise. The table the rat explored was modeled as a rectangle made up of 25 smaller squares. A separate state in the model was associated with the rat being within each square. From any given state it is possible to move to any of the adjacent states (3-8). Action choices at each state are confined to either moving to another adjacent state or staying in the existing state. Each adjacent square has the same initial relative probability

of being chosen to move into. The probability of choosing to do nothing for that iteration can be adjusted to normalize the model's activity to that of control animals. Figure 8 shows a general schematic for the model of the free exploration behavior.

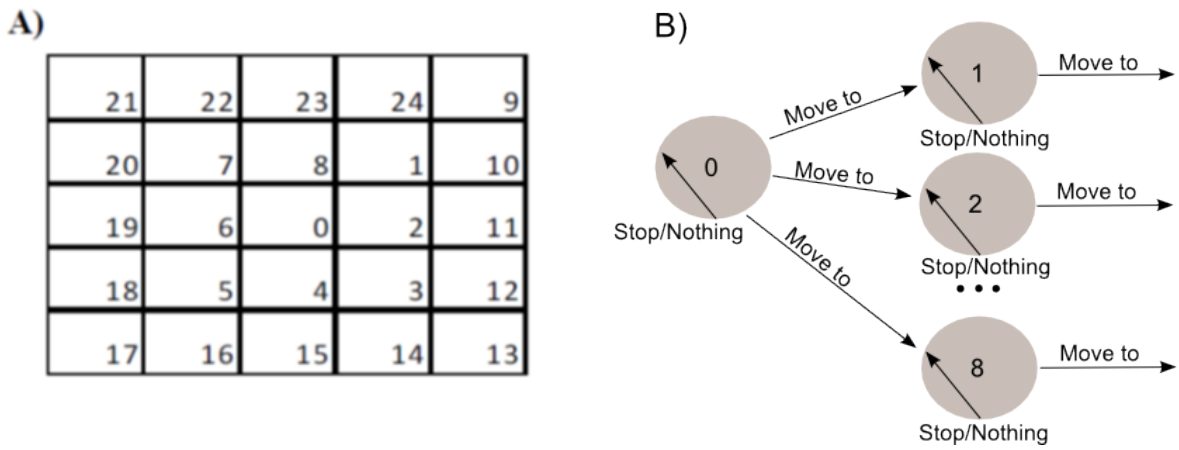


Figure 8: Schematics for how subsections in the simulation correspond to the data and Markov tree breakdown for states and action choices. *A* shows a schematic for how the table is divided into subsections; the labeling convention for the subsections corresponds to that used originally for the animal experiments (Szechtman et al., 1998). *B* shows a Markov tree breakdown for how the states and action choices are viewed by the algorithm. Each state corresponds to one of the subsections of the table.

The same basic procedure was used for both the TD-learning algorithm and the model-based approach. The DA signal is reset to a random value following a Gaussian distribution with a mean of zero and a standard deviation of five percent of the maximum firing rate allowed. The presence of Quinpirole is modeled by drastically increasing the DA signal by a factor of 10. Each time a decision is made it is considered one iteration

through the task, and the values associated with each of the states are updated. One iteration is treated as roughly equal to one second of real elapsed time for comparison purposes to the animal data.

Both approaches were able to model the compulsive checking behavior. In the TD-learning model the hyperactivation of the DA signal meant that the internal valuation of the reward received was significantly higher than anything that actually happened. The slight noise in the initial DA signal meant that a few areas ended up being more rewarding than others, and each subsequent visit to those areas was viewed as increasingly rewarding, under a continuously hyperactivated DA signal. Thus the system falls very quickly into a stable steady state in which 2-3 spots are treated as “highly rewarding” and the algorithm just moves between them.

For the aberrant salience hypothesis the same basic model was used. But rather than reinforcement learning driving the behavior, a security motivation system was added that initiated a return to the home state. This was accomplished by weighting the probability at any given decision point that the model would decide to head back to the home state by the DA signal it received the last time it was in that state. Therefore, when no hyperactivated DA signal was present the home state had an equal probability to that of any other state. In contrast, under conditions of DA hyperactivation, the home state had a significantly greater chance of being visited. In order to model the increase in overall activity level in Quinpirole-injected rats, the probability at any decision point to

move to another state was weighted by the DA signal convolved with a small amount of Gaussian noise. This meant that the amount of time the model spent “stopped” at any given state was significantly less under hyperactivated dopamine and thus the model showed significantly increased overall activity.

The model was run through a series of trials that mimicked the procedure used in the actual animal experiment. Ten trials under conditions of hyperactivated dopamine (Q1) were followed by one saline (S) trial with baseline levels of DA, followed by one more trial under hyperactivated DA (Q2).

Results

The learning measure, RMS percentage change in checking behavior, was calculated for the Q1 trial, the S trial and the Q2 trial for both the TD model and the aberrant salience model. Figure 9 shows the results of this measurement.

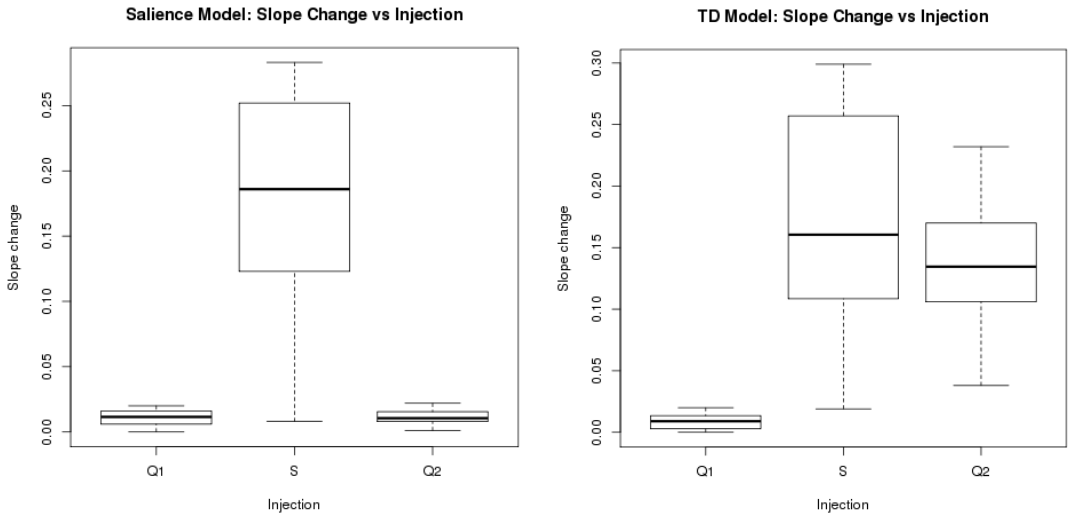


Figure 9: A boxplot comparison of the model-based and TD-algorithm approaches. Q1 is the last Quinpirole injection before S, the saline injection, followed by another Quinpirole injection, Q2. The slope change is a measure of learning, which shows that the primary difference in prediction is during Q2 where the model-based approach can immediately shift behavior without the need to relearn outcomes.

Overall both models showed similar behavior across trials and conditions and they matched the behavioral data available. The models also showed the predicted difference in the RMS between the Q1, S, and Q2 states. With the TD model the S and Q2 states were not significantly different from each other, but both were significantly different from Q1. Using the aberrant salience model, Q1 and Q2 were not significantly different, but S was significantly different from both Q1 and Q2.

The same results were calculated for the actual animal data. The RMS for the Q1 (right before saline injection), S (during saline injection), and Q2 (trial after saline) were

all calculated. The mean RMS for Q1 was 0.010, S was 0.038, and Q2 was 0.009. The standard deviation for Q1 was 0.007, S was 0.018, and Q2 was 0.009. Pair-wise comparisons were made between all three measures with Bonferroni corrected t-tests. Q1 and Q2 were not significantly different at all (p=0.90), Q1 and S were significantly different (p=0.003) and Q2 and S were significantly different (p=0.034). Figure 10 shows a boxplot of the average slope change across animals in all three conditions.

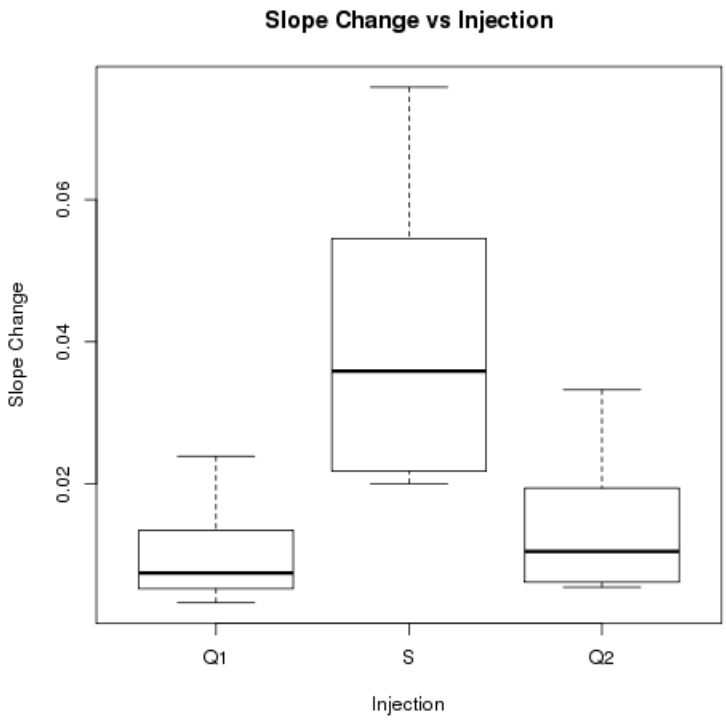


Figure 10: Learning rate as measured from animal behavioral data showing a close match to the model-based approach. Q1 is the last Quinpirole injection before S, the saline injection, followed by another Quinpirole injection, Q2. The slope change is a measure of learning. Behavior changed right away in during the Q2 injection as predicted by the model-based approach.

Experiment 2

Method

The same animal data used in the first experiment was analyzed in the second experiment in terms of similarity of locales visited between returns to the home square. The sequence of squares entered across a given trial was divided into trips by splitting the sequence at each home square arrival. A matrix was formed using each of these trips as one row in the matrix. Each trial then had its own matrix with rows corresponding to which squares the animal entered between home visits. To calculate a percentage similarity of trips the number of shared elements between each row of the matrix was calculated and then divided by the total number of elements in the matrix. This was done for the first ten sequential trials of Quinpirole injections for each subject.

For each injection point the mean percent similarity for each animal was calculated and evaluated for an increase in trip similarity during the course of the ten injections of Quinpirole. In addition, a mean split was used to place animals into a “high” similarity group or a “low” similarity group and the performance of the two groups compared.

The Model

The same basic model used in the first experiment was used in the second. To better illustrate how the TD model and the model-based salience system differ, consider the following reinforcement learning problem represented in Figure 11:

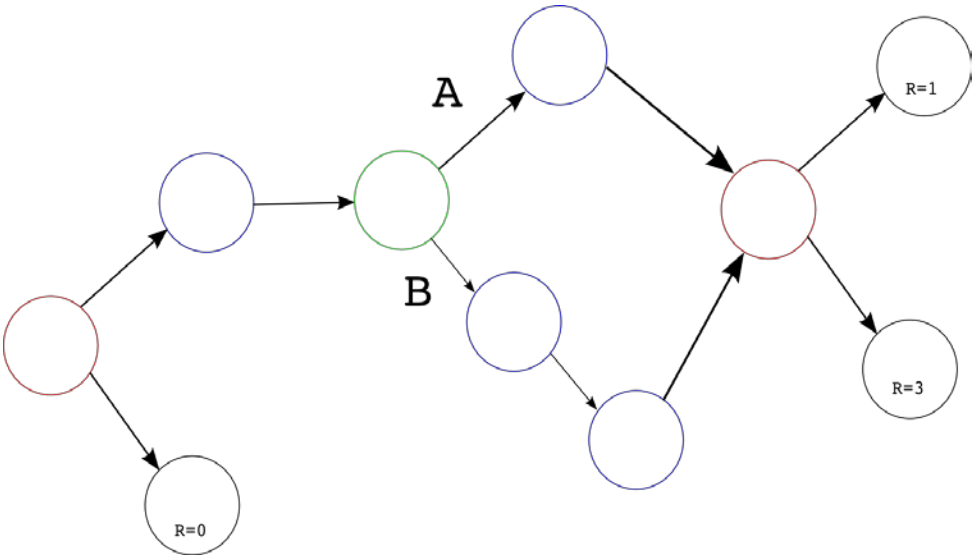


Figure 11: A schematic showing the Markov chain where the decision of action A or B does not alter reward beyond some base discount function. TD-algorithms would tend to favor the first choice picked while model-based approaches would show no preference.

Each circle represents a state, each arrow represents an action that, when preformed, moves the model into the state the arrow points at, and R represents a scalar value for a reward received at the marked state. The circles are color coded as follows: red represents states where an action choice has a direct effect on the reward received, blue represents transition states where there is no decision needed to advance, green is a

decision point but the decision has no effect on the potential reward received, and lastly black represents “terminal” states where rewards are delivered.

Both model-based and temporal-difference learning algorithms can easily solve this learning problem. In both models, actions at the red states that produce maximum reward are quickly learned to be valued more highly than those that produce less reward. However, the interesting action takes place on the green circles rather than the red circles. Neither action A nor B has any impact on the final scalar reward value received. With most algorithms a slight favoring of action A will develop as it is the “shortest” path to the goal. However, the way the model-based and temporal-difference-based learning algorithms “learn” which action to select differ significantly.

Temporal-difference-based algorithms cache the total expected reward from performing a given action into a “state value” stored one step ahead. The formula to update a state value is:

$$V(s_t) <- V(s_t) + \alpha\left[r_{(t+1)} + \lambda\, V(s_{(t+1)}) - V(s_t)\right]$$

where $V(s_t)$ is a value of a state at a certain time point, r is a reward received, α is a learning rate, and λ is a discount factor. The value of a terminal state is based on the reward received at that state, and that value is then back-propagated to update the value of preceding states, weighted by a discount factor and a learning rate. Because it is a one-

step lookup procedure, and the values are cached, in order for a state to update its value the state must be experienced before reaching the terminal state.

In contrast, the model-based approach calculates the value of a state based on a complete look-ahead simulation through the model of the learning task. A state value is calculated as:

$$Q(s,a)=E[R_{s,a}]+\gamma\sum_{s'} T(s,a,s')\,max_a Q(s',a')$$

where Q(s,a) is the value of a state and action pair, T(s,a,s') is the transition function between states and E[Rs,a] is a measure of total expected reward. Since the model-based approach has complete access to all the states, the transitions, and the estimated reward functions available at any point in the learning task, it can assign a value to a state, and update that value, regardless of whether or not that particular state was experienced paired with a reward.

Since the model-based approach does not need to experience a state paired with a given reward in order assign it a value and update that value, the probability of performing action A or B in the testing phase should be the same regardless of the frequency of exposure during the learning phase. This probability is determined by the weight of the discount function. In contrast, temporal-difference learning only assigns or updates a value when there has been a direct paring of a non-terminal state with a terminal state, so the frequency of experiencing that pairing during a learning phase

should have a profound effect on action probability during testing phases. This holds true only if several conditions are met: first, that the learning rate is not too high since that makes the system highly sensitive to single events, and second, that the animal has not had so much exposure during the learning phase that even for very high or low frequency choices complete learning has occurred.

This basic problem can serve as a framework for understanding the learning of visits in animals during the Quinpirole injections. The TD algorithm caches the perceived value of the checking behavior into the states it entered before returning to the home square. This should increase significantly the chances of entering those squares in the future. Thus over time the places visited between checking points should converge and increase in similarity.

The aberrant salience model makes the opposite prediction. First, learning may not be playing much, if any, role as the checking behavior is motivated by failure to terminate the security motivation system. However, even if the animal is learning or coding the behavior at some level, the model-based learning approach would not reinforce the specific trip the animal followed between checking events. Therefore, it is predicted that the previous trip should not correlate significantly with future trips.

Results

Figure 12 shows the percent similarity between multiple trips taken to the home point averaged for each injection time point for both the TD and salience models.

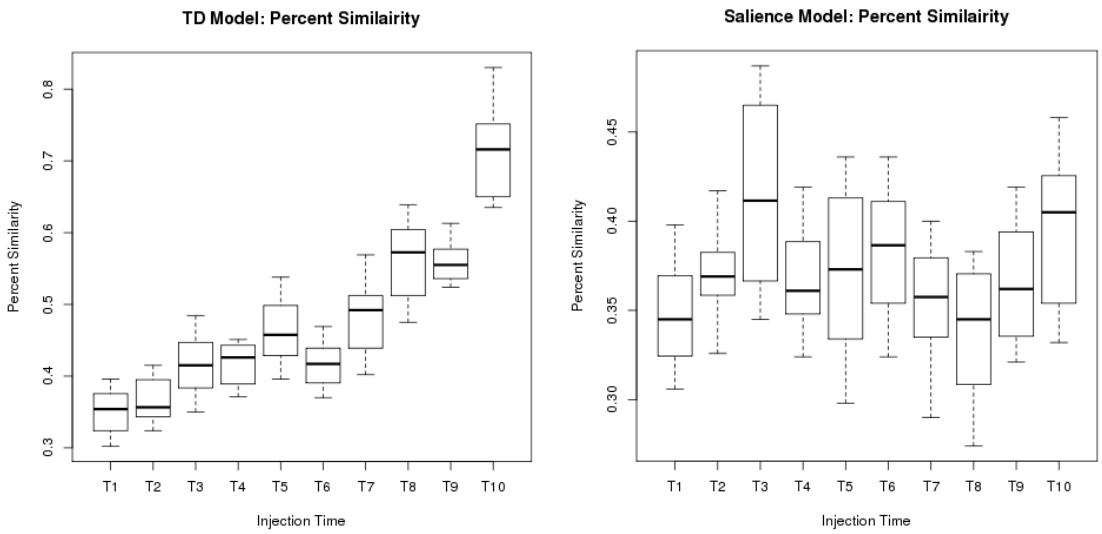


Figure 12: A comparison of the model-based and TD-algorithm approaches to path reinforcement. The TD approach tends to favor previous paths and so the percent similarity goes up over time. The model-based approach does not show this effect.

The TD model clearly shows a significant increase in similarity over time while the aberrant salience model shows no significant change over time.

Figure 13 shows percent similarity of trips across ten trials of Quinpirole injections in rats.

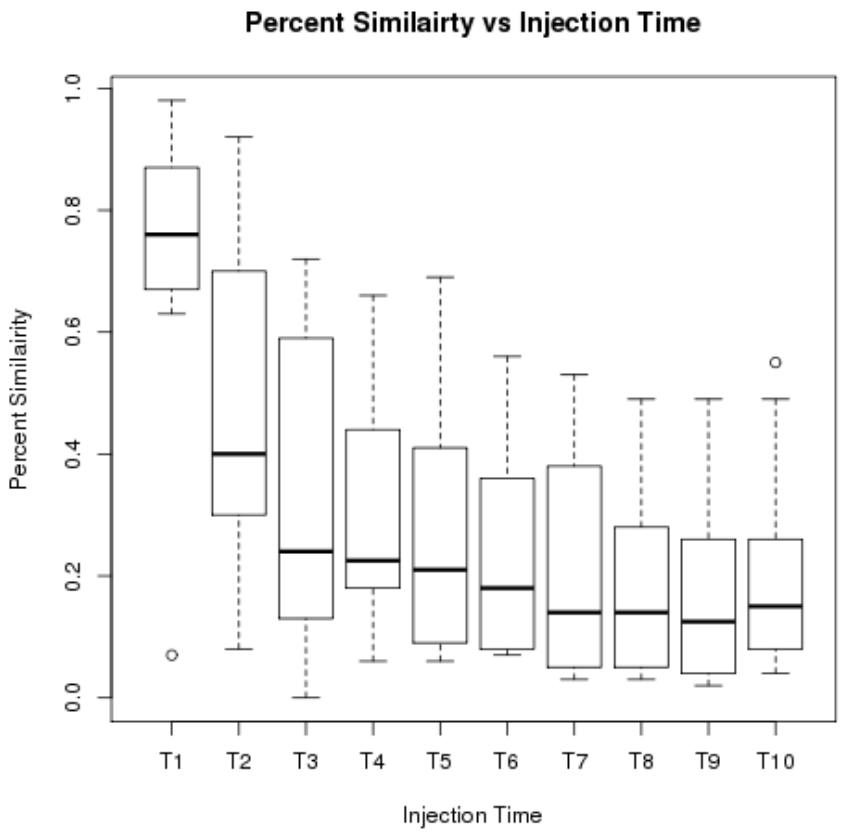


Figure 13: Animal data showing the percent similarity in trips over time. The data shows a significant decrease in similarity over time, though this is driven by a split in strategies between different animals.

While the percent similarity across all ten trials was significant, it was in the opposite direction predicted by the TD-learning model of dopamine. However, this masked a phenomenon where several animals showed significantly high similarity across their trips while others showed little to none. To further explore this, a mean split was

used and the high-similarity and low-similarity groups were compared. Figure 14 shows this comparison.

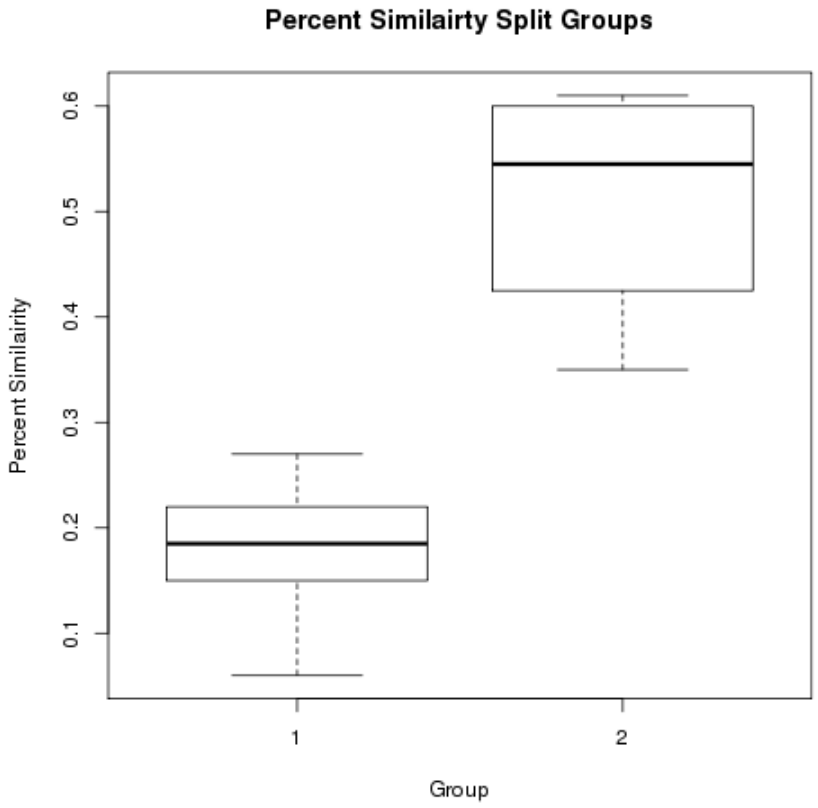


Figure 14: Animal data showing that there is a low and high group for percent similarity between trips. This may suggest that there are dual strategies with different animals using different strategies during free exploration.

The difference between the two groups was highly significant and might be suggestive of dual underlying causes dependent on the animal. Another area where the TD and aberrant salience models make slightly different predictions is on the increase in

activity over time. For the TD model the increased activity during the Quinpirole injection is due to an increase in perceived reward learned over time, while the aberrant salience model uses the DA salience cues to motivate increased exploratory behavior. Thus the TD model predicts that activity should increase over time while the aberrant salience model does not.

To see if there might be evidence of dual causes, we correlated the percent similarity in trips taken with the increase in activity over time. Figure 15 shows the correlation.

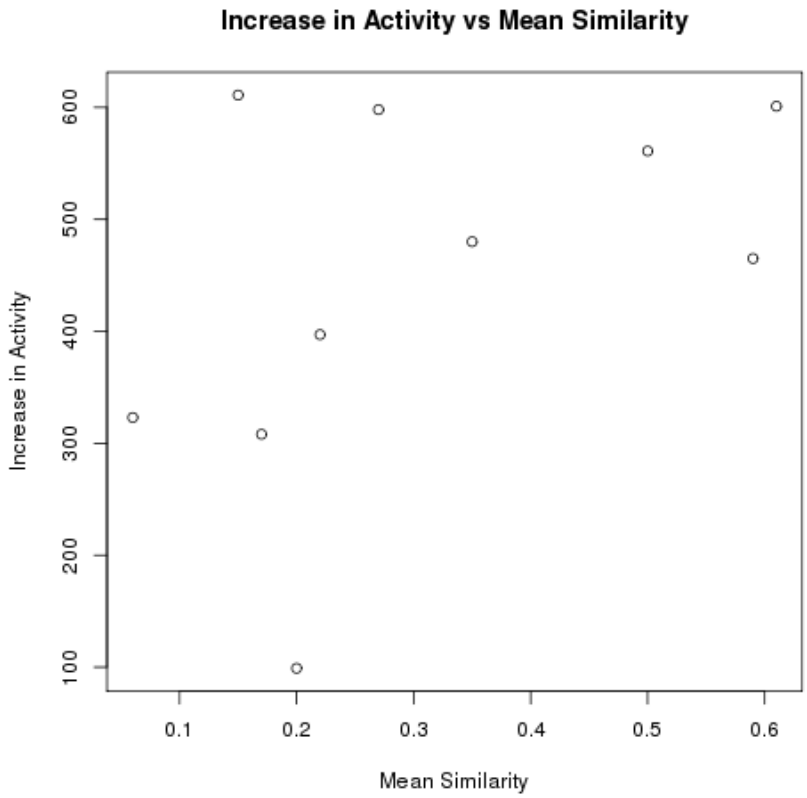


Figure 15: Animal data looking at increased activity over time (a prediction of the TD-approach) versus percent similarity between trips. The slight trend seen in the data might be evidence for the dual strategy hypothesis.

There was a trend toward a positive correlation, as predicted, although it was not statistically significant (r=0.47, p=0.08).

Discussion

While both models were able to account for the majority of the behavioral data from the OCD animal model, there were several distinct and important differing predictions. The clearest difference was in the predicted changes in behavior after the saline injection trial. Both the TD and aberrant salience model predicted a higher root mean square (RMS) of the percentage of visits to the home square during the saline injection trial but only the TD model predicted an increased RMS when the Quinpirole injections resumed after the saline trial, a prediction not borne out by the animal data. However, the aberrant salience model simulated a clear increase in RMS during the saline phase but no increase in RMS during the two Quinpirole phases, which were not significantly different from each other.

The trip analysis was a bit less conclusive. When averaged across all animals there was no evidence of trip similarity in between checking events. Moreover, the level of similarity showed a decrease over time rather than an increase. However, the animals appear to be split into two distinct groups. The first group, made up of the majority of the animals, showed no similarity across trips at any point, while the second, smaller group showed relatively high similarity in trip choice.

This was examined further using activity increase as another potential predictor of TD or aberrant salience models of the behavior. The mean similarity for each subject, while not significant, was suggestive of an increase in path similarity over time in a

subset of animals. One potential explanation for this is that the security motivation system and reinforcement learning both play some role in explaining the behavior of the animals during Quinpirole injections. The total contribution of each system to the animal's behavior might differ across subjects.

If this is the case, then teasing out the contributions of the two systems could be an interesting avenue for further research. One approach to potentially increasing the contribution of the reinforcement learning system could be to add an actual reward to the home state such as a small amount of food. To increase the contribution of the security motivation system a small fear stimulus could be introduced if the animal stays away from the home square for any length of time. The same trip analysis and activity information could be compared to see if introducing either of these increases the number of animals showing reinforcement motivated behavior vs. danger avoidance behavior.

While the trip information is somewhat suggestive of a split, the overall evidence most strongly supports the Szechtman-Woody model and our model of dopamine as a form of aberrant salience. This suggests that, to thoroughly understand the role of dopamine in learning and behavior, we need to go beyond the limited view of DA as signaling a reward-prediction error. Viewing the DA signal as conveying salience and surprise can contribute to our understanding of cognitive disorders that are linked to dysregulation of dopamine. It is this connection of DA as a salience signal elucidating causal elements of psychological disorders that we explore more fully in the next chapter.

Chapter 4

Introduction

The ventral striatal dopamine system is widely believed to play a pivotal role in reward learning and motivated behavior. Moreover, substantial empirical evidence suggests that a disruption of the dopamine system underlies the perceptual and cognitive deficits observed in a variety of psychiatric disorders including the positive symptoms of schizophrenia (Iversen & Iversen, 2007). The temporal difference (TD) model of reinforcement learning has gained substantial interest among neuroscientists and theorists as an explanation for the role of dopamine in learning, and has led to computational accounts of how abnormal reward learning might cause some of the positive symptoms of schizophrenia (Menon et al., 2007). Another influential theory of the role of dopamine in motivated behavior is based on its role in signaling incentive salience (Kapur, 2003). This view has argued that a dysregulated dopamine system leads to aberrant perceptions of salience, and can provide an alternative account for many of the positive symptoms in schizophrenia. The models described in Chapter 2, based on the incentive salience account of dopamine, were shown to account for a range of seemingly contradictory behavioral data in animals with dopaminergic disruptions. Moreover, it was shown that these models, which conceptualize dopamine

as a signal of surprise and salience, account for some electrophysiological data that the standard TD-learning approach cannot.

A novel aspect of the incentive salience-based models described in this thesis is the incorporation of context-specific learning. Mediated by the dopamine signal, shifts in context allow rapid adjustment of salience attribution and behavior. Thus, inherent differences in the strength of the DA signal dictate differentially the effects of contextual changes. When an unexpected reward is encountered, the DA signal is used as a measure of its significance relative to expectations of the model-based predictions for that reward. If a large DA signal is detected, our model starts searching for contextual cues that might have changed, for example, whether the internal state of a rat is hungry or satiated. In the absence of contextual changes, the model simply treats the unexpected reward as another data point and updates incrementally a new estimated value (assuming the change in reward value is not a one-shot trial). However, if there is a detectable context change coupled with an unexpected reward, then the model behaves as if something has changed, instantly drops previous expectations, and learns the new reward value within the altered context. Dopamine mediates this contextualized response in two ways, first signaling that a salient surprising event has occurred, and second, via the size of the DA signal, determining how heavily the model weights the hypothesis that there is a contextual change. The larger the DA signal, the more likely a contextual shift will be detected.

According to the model proposed here, potential differences in the endogenous levels of dopamine affect how individuals respond to contextual changes. If the baseline DA signal is larger than average, then individuals will be more sensitive to contextual shifts. They should then be more likely to immediately respond to changes in reward expectations. Lower levels of endogenous dopamine should have the reverse effect, making individuals more prone to overlooking a context change and attributing a change in reward value to a poor initial estimate. In the latter case, a more gradual, incremental updating of reward value estimates and decision-making policies takes place. It is important to note that the likelihood of quickly adjusting behavior to a context change is not dependent solely on the strength of the DA signal, as the context shift itself has salience. Highly salient context shifts can cause immediate updating even with low levels of dopamine, while undetectable context shifts should not cause immediate updating even in individuals with highly elevated levels of endogenous dopamine. These predictions can best be tested empirically in people who have naturally varying endogenous dopamine levels. Those individuals with larger endogenous levels should be far more likely to quickly update decision-making strategies when faced with subtle or low salient context shifts, while highly salient context shifts should not result in substantial differences in decision making between those with high and low endogenous dopamine levels.

Several behavioral indicators of endogenous dopamine levels have been investigated in the literature. These include cognitive tests (e.g. McNab et al., 2009; Croply et al., 2006) and neuropsychological assays of certain personality and behavioral traits. We used a battery of both types of tests: Schizotypal Personality Questionnaire, Magical Ideation Scale, and Impulsivity Scale, a reversal learning task and a working memory task to divide our participants into groups based on presumed differences in endogenous dopamine levels. To assess sensitivity to changes in context and reward value, we used a novel two-choice reward-based learning task where the reward values reversed during either a high salience or low salience contextual shift. The main measure of interest is accuracy of choice after a return to the original reward contingency. We predicted that those whose scores on the questionnaires and other tasks suggest higher endogenous dopamine levels would be significantly faster in returning to the original choice, compared to those with normal or lower levels, but only in the low salience context change condition. There should be no difference in the high salience context change condition.

One simulation and two experiments were run. The simulation looked at the predictions of the two models of the DA signal context. The first experiment examined the appropriateness of using certain questionnaires as measures of endogenous dopamine by correlating scores with performance on a reversal learning task known to predict endogenous dopamine as measured by positron emission tomography. The

second experiment expanded considerably on the first. We added two questionnaires, dropped the Cools et al. (2009) reversal learning task, added a working memory test and introduced our context-dependent reinforcement learning task.

Simulation

Methods

The same model described in Chapter 1 was used to examine more closely how decision making is updated over time in the presence of high-context or low-context cues. The model was presented with two decisions, one leading to a large reward magnitude with a μ=4, σ=0.2.The other was a reward magnitude using μ=2, σ=0.2. After 30 trials the reward distributions for the two choices were reversed, and after another 30 trials the reward distributions were switched back to the original reward scheme. The model has a built-in probability measure for detecting contextual changes. For this experiment the context change probability was decreased by 30 percent for the low-context shift and increased by 30 percent for the high-context shift.

The main value of interest measured was the change in the frequency of the correct choice by the model, after reward contingency reversals. This change is shown in Figure 16 below.

Results

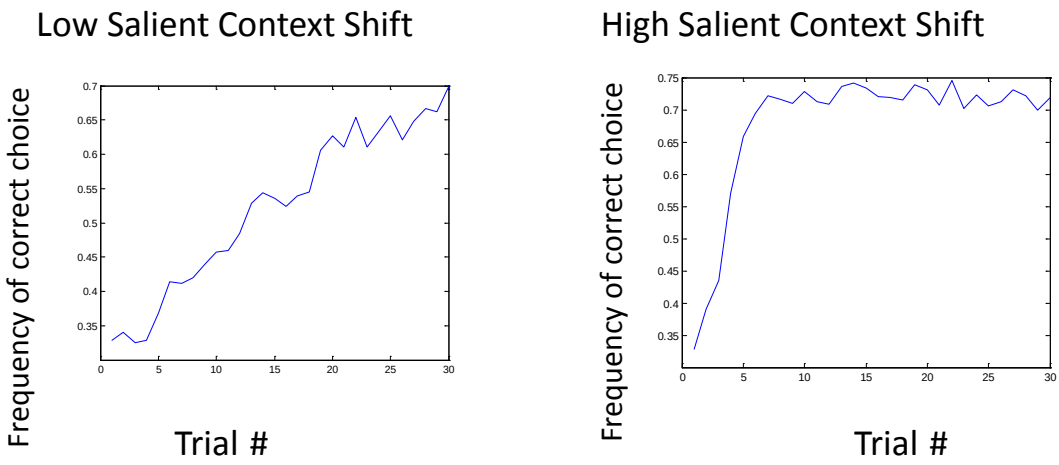


Figure 16: Model data showing a comparison between low and high salience context shifts with normal DA levels. The low salience context shift is not registered by the algorithm as a true context shift; thus no context effect on reward is learned and change in reward estimate must be relearned after each reversal.

With a standard DA signal modified by a high salience context shift the model detects that the different choices have different reward distributions and quickly changes its decision making between context shifts. With a low salient context shift the model attributes the change to a poor prediction and uses the new values to slowly update its decision policy over time.

Figure 17 shows the results of the same simulation under conditions of elevated baseline DA levels. Under these conditions the model now treats the low salient context shift the same as a high salient context shift and updates its action policy quickly. This result is the heart of our prediction for the experiment. High endogenous

dopamine levels should lead people to perform better on this type of tasks but only for the low salient context shift.

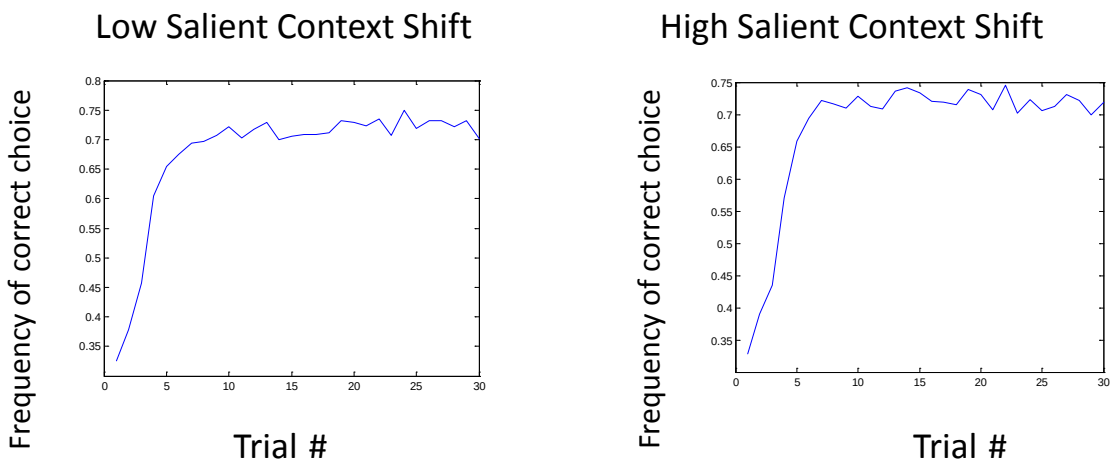


Figure 17: Model data showing a comparison between low and high salience context shifts with elevated DA levels. The low salience context shift is registered by the algorithm as a true context shift. The context effect on reward is learned and the reward expectation can be shifted instantly.

Experiment 1

The purpose of this experiment was to take several questionnaires and correlate the scores with performance on a reversal learning task known to predict endogenous dopamine as measured by positron emission tomography. We also developed a control for context-effects not related to reward learning.

Method

We used the Raine (1991) Schizotypal Personality Questionnaire (SPQ). A substantial body of evidence links Schizotypal Personality Disorder (SPD) to aberrant dopaminergic levels (e.g. Myin-Germeys et al., 2005; Levitt et al., 2002). Many of the positive symptoms of schizophrenia are shared in individuals showing SPD. Schizotype symptoms are of particular interest here because of their close relationship with salience and aberrant salience. The SPQ is a set of 72 true-false questions that has shown validity as a diagnostic tool in identifying schizotypal individuals. The SPQ was one of the main measures we used in this experiment as a baseline indicator of potential differences in endogenous dopamine.

In order to validate the SPQ as a potential predictor of endogenous dopamine levels, we compared SPQ scores to performance on a reversal learning task that is known to predict endogenous dopamine based on PET scans (Cools et al., 2009). The reversal learning task involved participants picking between two cards with either a face or a mountain printed on them. After selecting a card the participant sees a display of either a reward (adding points to their total) or a punishment (subtracting points from their total). The mountain or face would always lead to the same result, either a reward or punishment. After a set number of trials the reward or punishment was switched. Cools et al. (2009) found that individuals with higher endogenous dopamine

levels, as measured by a PET scan, learned more quickly in response to unexpected reward than unexpected punishment.

This experiment also developed our control for context effects in general relative to context effects specifically on reinforcement learning. To do this we used a word recognition task based on one used by Hockley (2008). A list of words was presented against various backgrounds and followed by a recognition test of those words against backgrounds where that specific word was originally learned, or a background used in the study session but with a different word, or a completely novel background. Performance on this task was compared to SPQ scores to see if there was some aberrant effect of context in general, not just linked to reinforcement learning. Participants were 23 psychology undergraduate students recruited from an online application. Students were given course credit for participation.

Results

Of key interest was whether or not SPQ was a good predictor of other behavioral correlates of endogenous dopamine levels. To test this we compared SPQ scores to performance on the Cools et al. (2009) reversal learning task. Cools et al. previously established using PET assays that the difference in performance on trials where unexpected reward and unexpected punishment were received correlates with endogenous dopamine levels. Figure 18 shows a plot of this difference score relative to

the SPQ score. One major problem to be noted is that this task was very complex, and many participants were unable to perform the task well enough to have usable data. Despite adding a tutorial and a detailed walkthrough, most of the students performed worse than chance. We used a cut off of 75 percent accuracy in our comparison, which resulted in 17 out of 27 participants being excluded from this analysis.

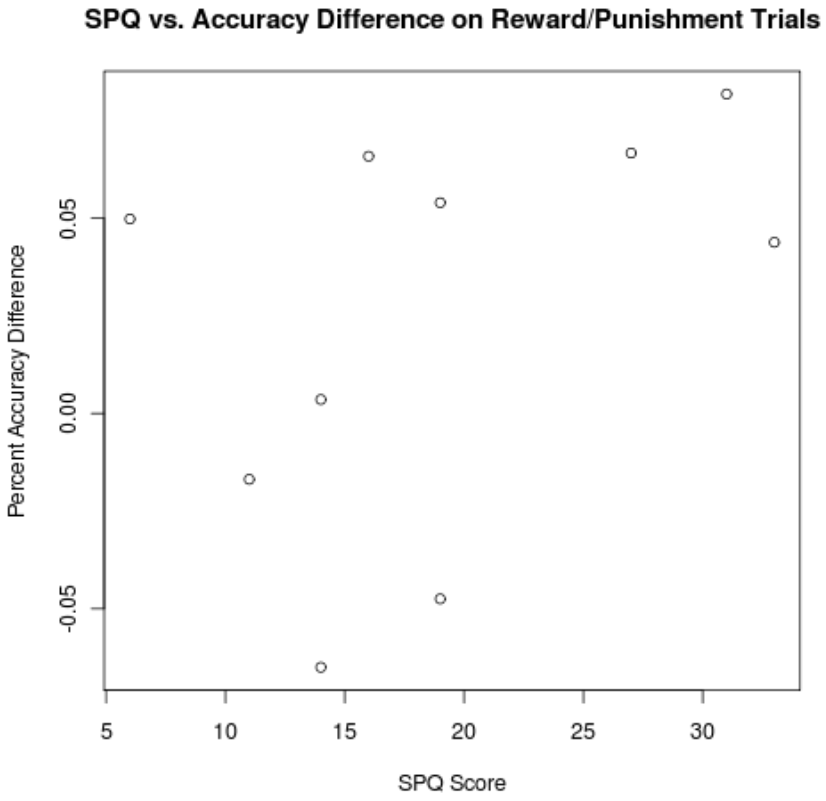


Figure 18: Experimental data showing the correlation between SPQ scores and the accuracy difference between unexpected punishment versus unexpected reward. Several of the perceptual SPQ subtypes were significantly correlated.

Although the overall correlation was not quite significant (p=0.09), the sub-score of perceptual and cognitive schizotypy traits did yield a significant correlation (r=0.57, p=0.043; r=0.58, p=0.039). The apparent difficulty of the reversal learning task reduced the power of the analysis. Nonetheless, the data support our hypothesis that the self-reported SPQ scores in normal populations can be indicative of differences in endogenous dopamine levels.

Another issue of interest was whether SPQ scores might predict aberrant performance on any context-dependent task. Since we specifically predicted an effect on reinforcement learning, if subjects showed differential performance on any context-dependent task it would be difficult to reconcile with the model. We tested using the Hockley (2008) word recognition task. Figure 19 shows the results of this task, divided into high and low SPQ scores. The top 30 percent of respondents to the SPQ were the high group with the remaining subjects in the low group.

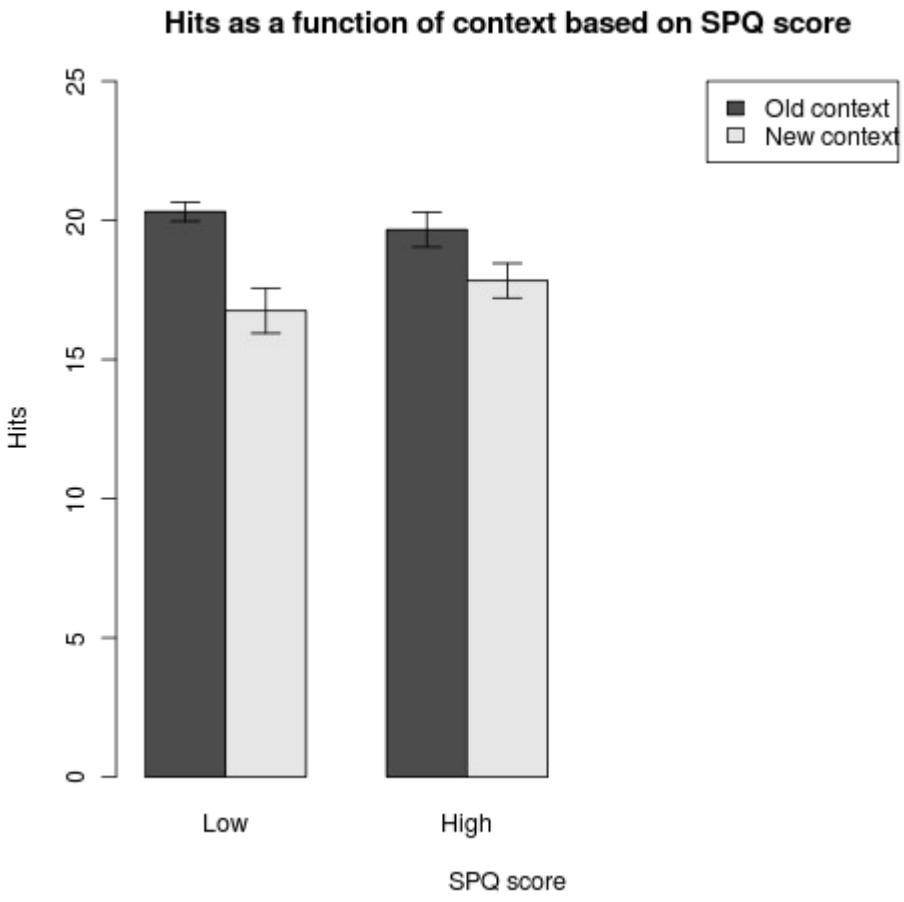


Figure 19: Experimental data showing the relationship between SPQ score on non-reward-based contextual learning. While context had a significant effect on recall, SPQ scores did not show a significant relationship.

A two-way ANOVA showed a significant effect of context but not of group, and no significant interaction effect (context: f= 3.8569, p=0.056; group: f= 0.0469, p= 0.82960; interaction: f= 0.4469, p= 0.50765). Main effect of context was significant in both groups (high: t= 2.06, p= 0.052; low: t=2.01, p=0.054). The results indicate that

high SPQ scores do not effect context-dependent processing in a general way, and show normal effects in tasks that do not use reinforcement learning.

Experiment 2

The second experiment built on the first but added two new questionnaires, a magical ideation scale and an impulsivity scale, both of which likely correlate with endogenous dopamine levels. Understanding how magical thinking could be related to dysregulation of the dopamine system is understudied and could offer significant societal benefits, while impulsivity has been shown to correlate strongly with PET measures of endogenous dopamine and cognitive measures known to be affected by dopamine.

In addition to the new questionnaires, we added a working memory task. Performance on this task as a measure of working memory has been shown to correlate with endogenous dopamine levels measured directly by a PET scan (Aalto et al., 2005). Finally, we developed a new learning task to examine the interaction of context and salience on reward-based learning. The task was based on common reversal learning tasks but the reward reversals were accompanied by either a low salience context shift or a high salience context shift.

We predicted that participants with higher endogenous dopamine levels (as measured by SPQ scores) will have greater scores on the impulsivity and magical

ideation scales as well as better working memory. We also predicted that they will have greater sensitivity to context shift during reward learning than participants with lower endogenous dopamine levels.

Methods

This experiment used 27 undergraduate psychology students recruited online and given course credit for participation. Developed from the first experiment, Experiment 2 added several new questionnaires and introduced our context-dependent reinforcement learning task.

The first questionnaire added was the Eckblad and Chapman (1983) Magical Ideation Scale. This set of 30 true-false questions focuses on a participant's tendency to believe in supernatural events and powers. It asks about UFOs, psychic powers, predicting the future, mind reading and many other similar topics. The final questionnaire we used was the Barratt (1994) Impulsivity Scale. This set of 30 questions asks about focus, planning, motor activity and impulsiveness. To assess working memory, we also included the Salthouse and Babcock Listening Span task (Salthouse & Babcock, 1991). This task involves listening to a series of sentences and having to remember the last word in each sentence. The working memory score is the largest number of words the participant can remember without error, and ranges from one to seven.

During each trial participants were shown two symbols of a star. One contained no pattern; the other contained red polka dots inside the star. The subject was then told that each star led to a different points reward and asked to choose between the two symbols. Subjects were asked to maximize the total points earned over all of the trials. After the choice was made, the points rewarded were flashed on the middle of the screen before the next choice was made available. One choice had a reward distribution with a mean (μ) of 5, and a standard deviation (σ) of 1.5, while the other choice had a reward distribution of μ=2, and σ=0.6. Which symbol started with the higher reward distribution was counter-balanced across subjects. After 30 trials, the reward distributions were reversed and associated with one of two context shifts. The high salient context shift involved drastically changing the entire background color where the symbols appeared. The low salient context shift involved slightly altering the hue of the polka dots. Figure 20 shows examples of how the different contexts appeared with the stimuli presented. The order of the context shifts (high vs. low) was counter-balanced across subjects. After another 30 trials the reward distributions returned to baseline. The baseline reward values were used for 30 trials then a reversal happened once again using whichever context shift was not used the first time. Again, after 30 trials the reward values were returned to baseline and the subject completed one last set of 30 trials under that baseline.

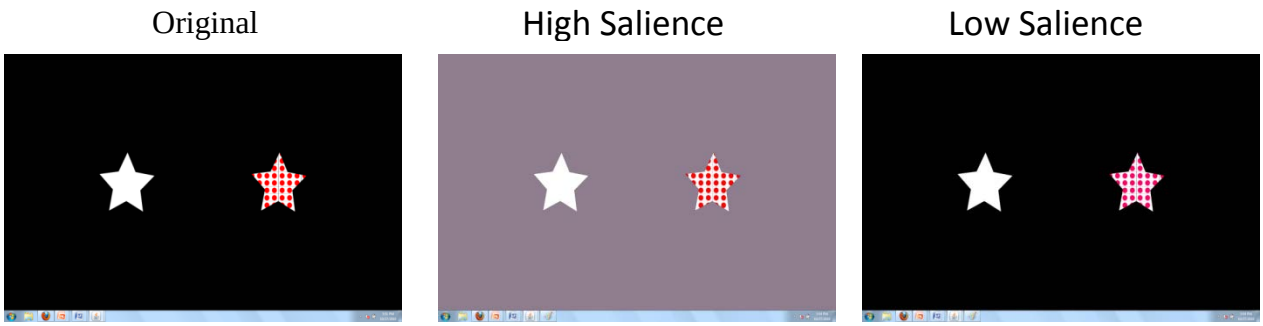


Figure 20: Screen shots from the context salience-based reversal-learning task. When the stimuli reverse their reward distribution is linked with either a low-context shift in the color of the dots, or a high-context shift change in background color.

Results

For the second experiment we looked at how well magical ideation and impulsivity scores correlated with the SPQ measure. Magical ideation and SPQ strongly correlated with each other (r=0.70, p=0.0007). Impulsivity and SPQ also correlated with each other (r=0.6, p=0.007). Figure 21 shows the plots of each of these correlations.

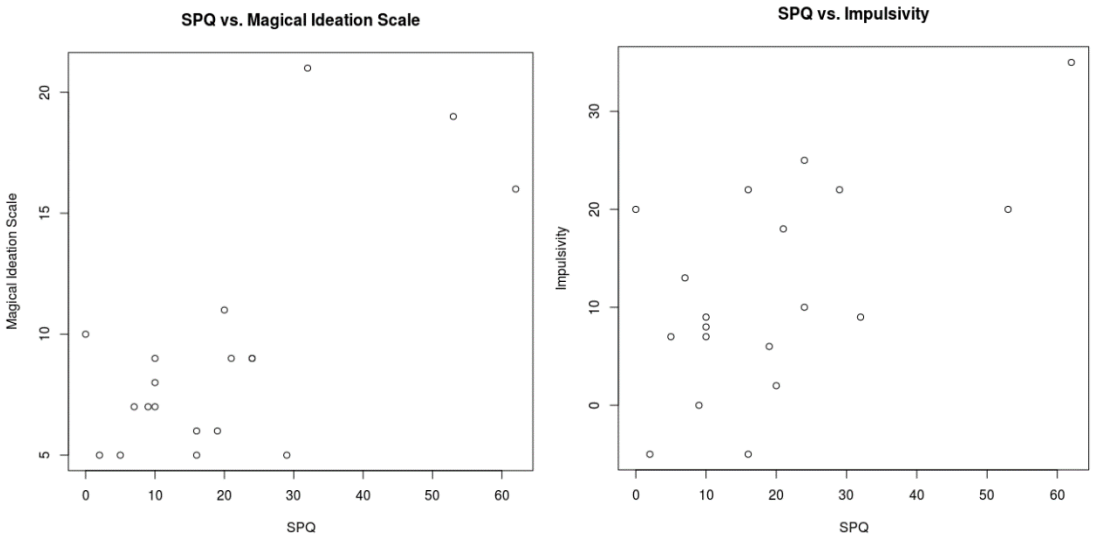


Figure 21: Empirical data showing the relationship between SPQ and two other potential correlates of increased endogenous dopamine. Both the Magical Ideation scale and the Impulsivity scale highly correlated with SPQ.

Scores on the working memory task were also analyzed for correlations with the SPQ, magical ideation and impulsivity measures. The working memory score correlated strongly with the magical ideation scale (r=0.70, p=0.001), and was marginally, though not significantly, correlated with the SPQ score (r=0.41, p=0.062). The working memory score, however, did not correlate at all with the impulsivity scale. Figure 22 shows the plots for each of these correlations.

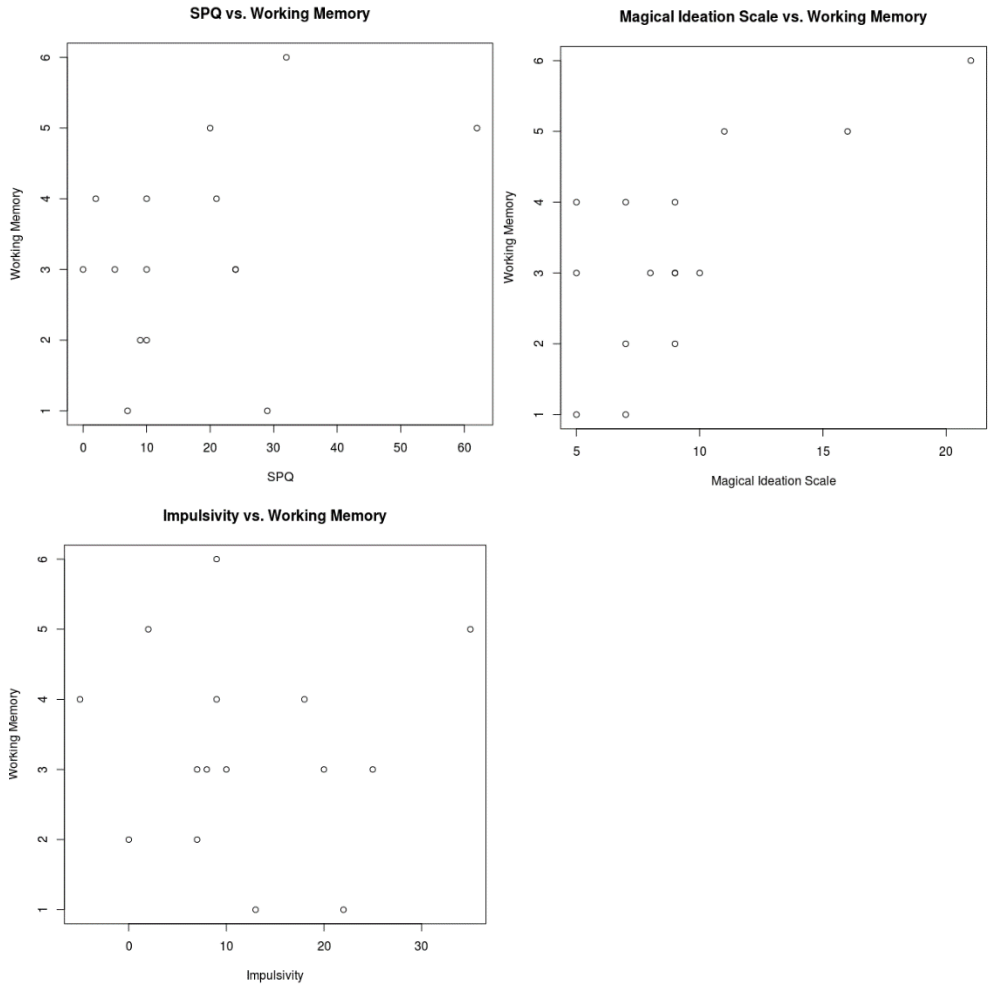
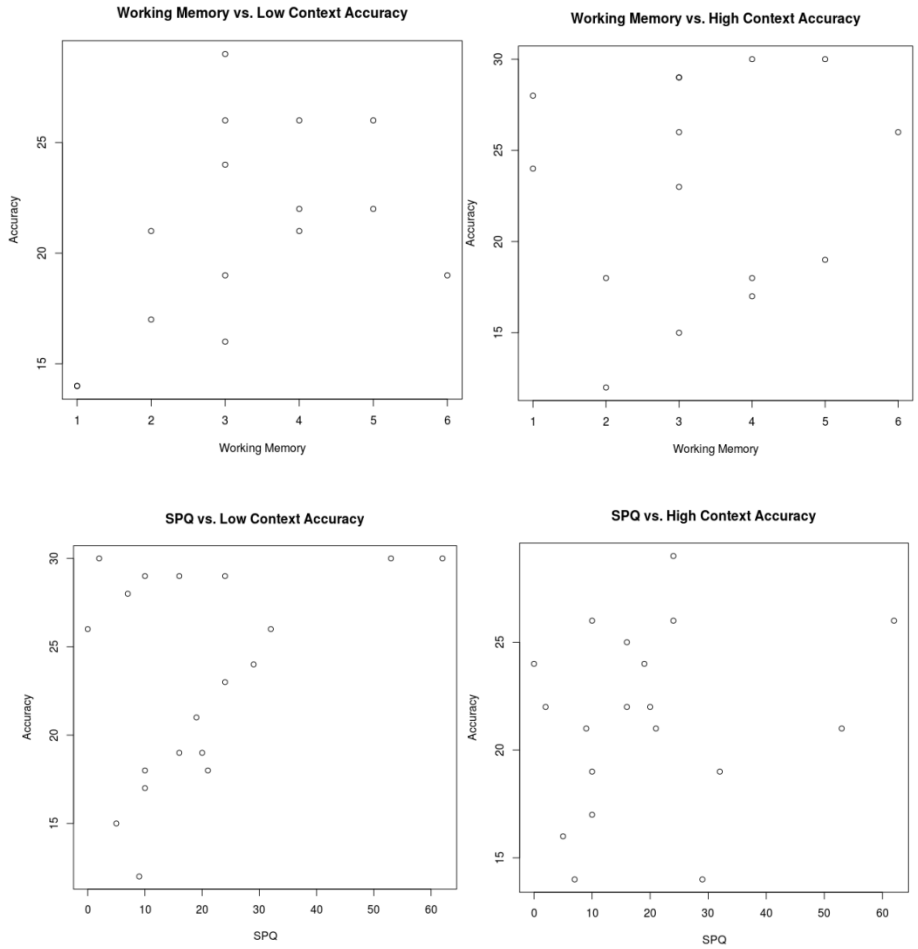


Figure 22: Empirical data showing the relationship between working memory and other potential correlates of endogenous dopamine levels. Working memory correlated with both SPQ and Magical Ideation but not with Impulsivity.

The main score of interest in the context-dependent reinforcement learning task was the accuracy when returning to the baseline reward after a reversal from one reward state to another reward state. This measure was calculated for both the high

salience context and the low salience context and compared to the scores on each questionnaire and the working memory task. SPQ correlated significantly with accuracy after the low salience context switch (r=0.37, p=0.05), but did not correlate with accuracy after the high salience context shift (r=0.2, p=0.19). Impulsivity also correlated significantly with accuracy after the low salience context shift (r=0.41, p=0.03) but not after the high salience context shift (r=0.22, p=0.17). The Magical Ideation Scale correlated marginally, but not significantly, with accuracy after the low salience context shift (r=0.33, p=0.08) and still more weakly after the high-context shift (r=0.1, p=0.33). Finally, scores on the working memory task correlated significantly with accuracy on the low salience context shift (r=0.47, p=0.03) but not the high salience context shift (r=0.12, p=0.32). Figure 23 shows the correlation graphs for each of these measures.


Working Memory vs. Low Context Accuracy
Accuracy
Working Memory
Working Memory vs. High Context Accuracy
Accuracy
Working Memory
SPQ vs. Low Context Accuracy
Accuracy
SPQ
SPQ vs. High Context Accuracy
Accuracy
SPQ

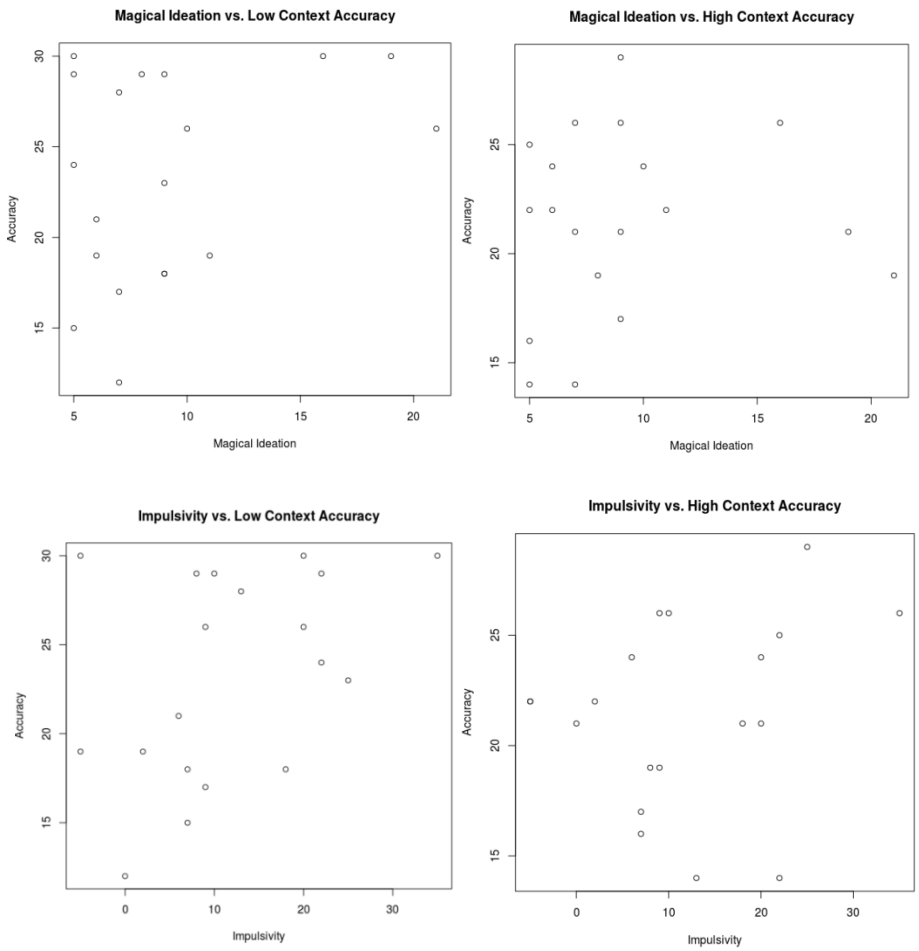


Figure 23: Empirical data showing the relationship between accuracy on the low- and high-context tasks and the set of potential endogenous dopamine level correlates. As predicted by the model, these correlates predicted low-context accuracy significantly more than high-context accuracy.

Discussion

The model described here differs significantly from the widely-used temporal difference learning algorithm, in modeling the DA signal as conveying surprise and

salience, and providing a probabilistic framework that accounts for performance in the face of contextual shifts during reinforcement learning. The somewhat counter-intuitive prediction that increased levels of endogenous dopamine could actually increase accuracy in a reversal learning task with low salience comes directly from this formulation of the DA signal, and accords with the empirical evidence presented here.

To validate our model empirically, we employed two behavioral tasks that are well established to correlate with endogenous levels of dopamine based on PET imaging: the Cools et al. (2009) reversal learning task and the Salthouse and Babcock (1991) working memory task. We also employed several neuropsychological scales, the SPQ, magical ideation and impulsivity scales, all of which assess traits that are associated with dopamine dysregulation in human studies, including sensitivity to dopamine agonist drug manipulations (for review, see Thompson, 2009). These experiments showed that SPQ and Magical Ideation scales strongly correlated with the behavioral tasks known to predict endogenous dopamine levels. They also strongly correlate with each other. Impulsivity, however, did not correlate directly with any behavioral task but did show correlations with the other measures. The preponderance of evidence is strongly suggestive that these measures, both individually and especially when combined, can offer a strong baseline for predicting difference in endogenous dopamine levels in normal subjects.

All of these measures were also strongly predictive of increased accuracy during the low salience context shift in the reversal learning task but not predictive of the high-context shift. This is exactly as predicted by the model. While this evidence does not directly address whether the DA signal is conveying a “salience” signal, it does suggest strongly that it mediates sensitivity to detecting context shifts, and that context shifts alter the way decision-making policy is updated. The reversal learning task is strongly analogous to the animal devaluation tasks where insensitivity to reward devaluation is dependent on both significant training as well as exposure to the devalued reward in a new context. Existing cached value algorithms view the DA signal as conveying merely a reward-prediction error and not being dependent on a context shift. It is important to note that the reward reversal is the same in both the high-context and low-context shifts so the reward-prediction error is exactly the same. If the DA signal was conveying only reward-prediction error then any changes in behavior due to differences in endogenous dopamine should be consistent across both context types.

Understanding what the DA signal is actually doing and conveying is important in understanding mental disorders that arise from dysregulation of dopamine. The Kapur (2003) model of schizophrenia being a form of aberrant salience relies on the DA signal actually conveying salience information. Our model directly attributes this role to dopamine. The experimental work done here is directly predicted from our

model of the DA signal and not from reward-prediction error models. All of the subjects used were from a normal pool of psychology undergraduates. While clearly there was some differentiation within the scores, these are all normally functioning individuals. The next step would be to try the experiment with individuals with strongly dysregulated dopamine systems such as Parkinson's patients, schizophrenics, or those suffering from obsessive compulsive disorder. If the same improved function in low-context shifts could be demonstrated it would add significant weight to our interpretation of the results. Another possible extension of the work would be to adjust the reward distributions during reversals to create more salient reversals. Our model does predict that, with a large enough reversal and even with a small level of endogenous dopamine, subjects should become sensitive to low salience context shifts.

One interesting element of this experiment is the Magical Ideation Scale. This scale asks people about a lot of popular topics such as UFOs, psychics, mind reading and predicting the future. These topics also correlate with beliefs in conspiracies that affect the choices people make in everything from paying taxes to seeking medical treatment. There is a strong disconnect between many "believers" and "skeptics" as to what constitutes evidence or reasonable beliefs. This experiment is suggestive that the relative strength of magical thinking and skepticism may actually have its root in the DA signal. This makes sense if the DA signal is conveying a salience signal. People with higher endogenous levels of dopamine might be more easily convinced by single,

dramatic events or more likely to attribute underlying importance to coincidental events. Further exploration of skepticism and belief as it relates to the DA signal would be fascinating and aid in educating or even persuading people traditionally resistant to important topics in science and medicine.

The results of this experiment directly match the predictions of our model and support the view that the DA signal is conveying a wide range of information beyond prediction error. It offers strong support for the aberrant salience model of schizophrenia.

Chapter 5

Overview

The VTA dopamine (DA) system is extensively involved in a wide range of cognitive and behavioral functions. It has reciprocal connections throughout the brain from the tip of the prefrontal cortex to the bottom of the brain stem. Understanding exactly what the DA signal is conveying has important implications for how animals learn, how they make decisions, and how they motivate behavior. It is strongly implicated in addictive behavior and a multitude of mental disorders associated with a dysregulated DA system. Even the small individual differences of endogenous DA levels in the normal population appear to have measurable effects on cognition and behavior. The major contribution of this thesis is to link neuropsychiatric conceptions of the DA system as a signal of incentive salience with computational modeling of reward-based learning and to apply these concepts to understanding and predicting mental disorders such as OCD, as well as individual differences in contextual learning based on differences in endogenous dopamine levels.

Contributions from machine learning to understanding the DA signal in the VTA have been immense. The correlation of the reward-prediction error term in the temporal-difference algorithm of Sutton and Barto (1998) with the cell recording data out of the

Schultz lab has solidified the role that computational modeling plays in understanding the DA system. In addition to predicting some of the cell recording data, reinforcement learning models have been able to explain a wide range of animal behavioral data involving reward-based learning. Alternate accounts of the function of dopamine such as the incentive salience account of Kapur (2003) have received considerable theoretical interest but are less well grounded in a comprehensive computational theory.

To date most of the focus in computational modeling has been on the DA signal as a reward-prediction error term. While this correlates well with the Schultz et al. (1997) cell recording data there are other data, such as from Tobler et al.'s (2005) experiments with varying reward probabilities, which suggest that the DA signal is conveying something more akin to reward probability than a direct error signal. Moreover, the TD-learning model does not readily explain data on reward devaluation and researchers have resorted to complex dual-system accounts combining model-based and model-free components (Daw et al., 2005) to explain these data. The TD-error account also fails to explain the sensitivity of the DA system to novelty and general uncertainty, and is not readily compatible with the neuropsychiatric explanations of the DA signal as conveying an aberrant sense of incentive salience in schizophrenia (Kapur, 2003). Building on the earlier modeling work of Smith et al. (2006), in this thesis we provide a solid computational basis for the incentive salience account of dopamine that can explain all of the above data.

Contributions to reinforcement learning models

Smith et al. (2006) formulated the DA signal as conveying salience but their formulation of the DA signal does not match all the observed cell recording data, nor does it predict the pattern of behavior seen in the reward devaluation experiments on animals. The model presented in this thesis uses a single system model-based approach built on the one presented by Smith et al. The most significant contribution is directly modeling a context-specific memory system, incorporated within a probabilistic framework that can generalize new information across contexts when appropriate or develop specific context-dependent memories when information should not be generalized.

The most likely neural locus for this context-specific learning and memory is the hippocampus. As discussed in the introduction, there is evidence of a reciprocal relationship between the DA cells in the VTA and the hippocampus, with the hippocampus able to modulate the firing rate of DA cells in the VTA and the VTA being important in the development and memory of context-specific cues in the hippocampus. A large body of literature supports a key role for the hippocampus in context-specific memory formation, for example, in contextual fear conditioning (e.g. Maren et al., 1998; Phillips & LeDoux, 1992). Dopamine agonists increase the encoding of recently acquired information, as well as its retention in the hippocampus, while antagonists inhibit

encoding and retention (for a review see Lisman & Otmakhova, 2001). These data are consistent with the role played by dopamine in our model. While a general estimate of reward value can just be cached and applied across many contexts, context-specific value estimates (e.g. food reward delivered in a satiated state) have to be encoded as separate memories with contextual cues retained over considerable time periods. The increase in DA activity in response to reward expectation violations could modulate the encoding of such context-specific memories in the hippocampus. The inclusion of this context-dependent system allows the model to explain all the seemingly contradictory data observed in the devaluation literature within a single system.

In addition to including a context-sensitive memory system in the model-based learning approach described in this thesis, the DA signal was modeled as conveying probability information about the observed outcome. It was shown that the DA signal of probability not only matches classic cell recording data from Schultz et al. (1997) but also explains the Tobler et al. (2005) data on the effect of varying reward probabilities on DA firing. Moreover, our view of the DA firing as a signal of reward probability is compatible with the theoretical view of DA as an incentive salience signal as proposed by Berridge and Robinson (1998), and by Kapur (2003) in his aberrant salience model of schizophrenia. This is a major contribution, allowing the insights of neuropsychiatric data and theories to be unified with the ideas of machine learning and computational modeling.

There are several limitations to our model. The greatest limitation, shared by all model-based approaches to learning, is that they become computationally intractable as the decision trees to be modeled increase in complexity. On the other hand, it is well established that even for human experts, planning many steps into the future is challenging. Chunking one's experiences into previously successful action sequences ("habits") or action schemas ("strategies") are some of the ways in which many organisms appear to deal with the formidable task of planning many steps into the future. Seamlessly integrating a means of chunking frequently used action sequences into a model-based system is an important direction for future work.

There is evidence from lesion data that lends support to Daw et al.'s (2005) dual systems account and is more difficult to reconcile with our model. Lesions to the pre-frontal cortex have dissociative effects on devaluation sensitivity regardless of the level of training animals receive. More specifically, lesions of the prelimbic prefrontal cortex appear to disrupt devaluation sensitivity in moderately trained animals, while lesions of the infralimbic prefrontal cortex disrupt devaluation resistance in highly over-trained animals (Killcross & Coutureau, 2003). The current version of our model does not address the results of these lesion studies but the answers could be found in the cached action hypothesis. Lesions that disrupt the ability of the organism to actually perform model lookup for action policy could force the use of chunked action sequences, producing devaluation insensitivity regardless of training, while disruption of areas that

store cached action plans could create the need to always use model lookups for action policy thus always creating devaluation sensitivity regardless of training.

Contributions to obsessive compulsive disorder and animal models

The Szechtman Quinpirole animal model of obsessive compulsive disorder presented an opportunity to compare the reward-prediction error (TD-error) explanation of the DA signal to that of the incentive salience account proposed in this thesis. A major contribution of this work was to simulate both models in terms of their ability to explain data from an animal model of OCD based on the dysregulation of the DA system. The reward-prediction error model showed how the behavior could manifest due to some states being perceived as increasingly rewarding even though no external reward was delivered. Our model, extended to incorporate the Szechtman-Woody account of OCD based on the disruption of the security-motivation system, relies on the DA signal conveying incentive salience and fits with other neuropsychiatric explanations for the role of DA.

Both models made similar predictions data during the administration of the DA agonist drug Quinpirole, in general agreement with observed animal data. The major difference in the two models was at the point of re-administration of the drug after a saline trial. In this case, the animals exhibit an immediate adjustment in behavior and re-emergence of obsessive behavior, a pattern that was only predicted by the salience model.

This is consistent with the aberrant salience model of Kapur, where the initial development of the psychosis can take time but that, after remission, administration of a dopamine agonist will almost immediately cause the re-emergence of the psychotic symptoms. In contrast, the TD-learning account incorrectly predicts that a relearning period will be required before drug re-administration has an effect. Thus, by this account, the DA signal is used to adjust behavior gradually over time, while the Szechtman-Woody security motivation model of OCD and the Kapur aberrant salience model of schizophrenia view the DA signal as conveying information that causes an immediate adjustment in expectation, behavior and thought processes. The DA signal is directly driving the conceptual framework being used to understand the world. Dysregulation in this system doesn't just cause aberrant learning but also results in an immediate adjustment in the way the environment is perceived and understood. Showing that the salience model better predicts data from the animal model of OCD is a major contribution of this thesis, and contributes to unifying the neuropsychiatric and computational accounts of mental disorders. It also agrees with self-reports from patients about their perception and cognitive processes that drive the psychotic or obsessive thoughts and behaviors.

The greatest limitation of this work is that, while it can accurately predict the observed data once the symptoms have developed, it is murkier during the early stages. Attempts to look at whether specific paths taken by animals through the open field were

reinforced and repeated over time, as predicted by temporal-difference learning, were inconclusive. The data showed animals actually splitting into two groups, with one (the minority of animals) showing frequent repetitions of the reinforced paths, while the other (the majority of animals) showing no path similarities across trials. This relates to the Kapur model as well, with the development of delusions occurring slowly over time. There clearly seems to be some role for learning both in the self-reports and observations of schizophrenics and in the animal model of OCD.

Neither the reward-prediction model nor the salience model of dopamine fully predicted the observed behavior during this learning phase. Under the Kapur model of schizophrenia there is a role for learning with increased salience. While the perception of salience is increased right away, psychosis doesn't set in until that perception can be conceptually linked to some cue or thought process. A similar formulation of the emergence of OCD involving both early learning as well as disruption of the security-motivation system could elucidate the differences we saw between animals that seemed to spend more time “learning” the obsessive behavior than others.

In addition to expanded modeling, further animal behavioral work could test the difference by looking at behavior not under drug administration but in a paradigm that encourages “checking” behavior through either reward delivery or changes in the “home” location of the animal. If the Quinpirole-injected behavior is based on reward learning, animals not receiving the injection but instead actually having food placed on the home

square during exploration should show similar behavioral development. Likewise if the behavior is driven by disruption of the security-motivation system changes in the home space during exploration should show development of a similar behavioral pattern.

Contributions to understanding individual differences and endogenous dopamine

While large-scale disruption of the DA signal has been linked to multiple mental disorders, there are much smaller differences in endogenous dopamine levels across all individuals. Cools et al. (2009) showed that these slight differences, measured using PET imaging, have quantifiable effects on behavior and cognition during reward learning tasks. These slight differences have also been proposed to explain development of certain personality traits and disorders. The Raine (1991) Schizotypal Personality Questionnaire and the Eckblad and Chapman (1983) Magical Ideation Scale both measure positive symptoms found in schizophrenia, particularly those linked to an increased sense of salience in one's environment.

Our experiments, reported in Chapter 4, showed that the SPQ and Magical Ideation Scales are both strongly correlated with performance on behavioral tasks known to predict endogenous dopamine levels. They also strongly correlate with each other. This evidence strongly suggests that these two scales, both individually and especially when combined, can offer a strong baseline for predicting individual differences in endogenous

dopamine levels in normal subjects. This allows for a non-invasive means of estimating endogenous dopamine levels and linking these to novel behavioral tasks.

A major contribution of this thesis was to show that the context-dependent learning developed in our reinforcement learning algorithm was able to directly predict the relationship between salience and context on reward-based learning. All of our behavioral correlates of endogenous dopamine levels were strongly predictive of increased accuracy during the low salience context shift in the reversal learning task but not predictive of accuracy during the high-context shift. This effect also fits in perfectly with the Kapur model of schizophrenia. The slight increase in endogenous dopamine causes these people to be far more attuned to various changes in their environment particularly when there are changes to reward structures.

An important next step for future work would be to replicate the experiment in individuals with strongly dysregulated dopamine systems such as Parkinson's patients, schizophrenics, or those suffering from obsessive compulsive disorder. If the same improved function in low-context shifts could be demonstrated in those with elevated dopamine, it would add significant weight to our interpretation of the results. Another possible extension of the work is to adjust the reward distributions during reversals to create more salient reversals. The model predicts that with a large enough reversal, even subjects with low levels of endogenous dopamine should become sensitive to low salience context shifts.

Another very interesting extension of this work would be to examine more fully the role of differences in endogenous dopamine and magical thinking. Strong tendencies toward magical thinking are directly correlated to a wide range of important social and political attitudes, from avoiding medical care to falling for massive Ponzi schemes. There is clearly a continuum from extreme skepticism to complete credulity. The data is suggestive that the relative strength of magical thinking and skepticism may actually have its root in the DA signal. This makes sense if the DA signal is conveying a salience signal. People with higher endogenous levels of dopamine might be more easily convinced by single, dramatic events or more likely to attribute underlying importance to coincidental events. Demonstrating an effect of dopamine on this belief continuum would shed light on an understudied field with important implications for education and public policy decisions and, in the more extreme end, could explain how these thought processes can become so blown out of proportion as to enter into the territory of psychosis.

Conclusions

This thesis has shown strong evidence that the VTA DA signal can be viewed as conveying a signal of incentive salience. This formulation can be well modeled using reinforcement learning algorithms linking context and memory. It can explain a wealth of cell recording data from both traditional conditioning tasks as well as more complex probability-based conditioning tasks and reward devaluation paradigms. By

understanding the DA signal as conveying salience we can unify machine learning with neuropsychiatry. It combines explanations of the role of dopamine in both schizophrenia and obsessive compulsive disorder.

This unification of ideas between machine learning and neuropsychiatric explanations of mental disorders is an important step forward in understanding both these disorders as well as learning and memory in normally-functioning individuals. The work here was able to show that sensitivity to salience of context cues during reinforcement learning is well explained by behavioral correlates of endogenous dopamine levels within a normal population.

Much work remains to be done to fully understand the brain's reward and motivational circuits. The VTA has reciprocal connections throughout the whole brain, and cells that are sensitive to other neurotransmitters such as serotonin, gamma-aminobutyric acid, and glutamic acid. There is evidence of differential responses to reward, punishment, novelty and uncertainty. Further exploration at a more micro-circuit level, looking at specific drives in burst firing rate, will further elucidate the role of the VTA in learning, mental disorders, cognition and behavior.

References

Aalto, S., Brück, A., Laine, M., Någren, K., & Rinne, J. O. (2005). Frontal and temporal dopamine release during working memory and attention tasks in healthy humans: A positron emission tomography study using the high-affinity dopamine D2 receptor ligand [11C] FLB 457. *The Journal of Neuroscience*, 25(10), 2471-2477.

Abi-Dargham, A., Gil, R., Krystal, J., Baldwin, R. M., Seibyl, J. P., Bowers, M., & Laruelle, M. (1998). Increased striatal dopamine transmission in schizophrenia: Confirmation in a second cohort. *American Journal of Psychiatry*, 155(6), 761-767.

Angrist, B., & Kammen, D. V. (1984). CNS stimulants as tools in the study of schizophrenia. *Trends in Neurosciences*, 7(10), 388-390.

Barratt, E. S. (1994). Impulsiveness and aggression. *Violence and mental disorder: Developments in risk assessment*, 10, 61-79.

Bernardo, J. M. 1979. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society*, Series B (Methodological), 41, 113–147.

Berridge, K, & Robinson, T. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience. *Brain Research Reviews*, 28, 309–69.

Carlson, J. M., Foti, D., Mujica-Parodi, L. R., Harmon-Jones, E., & Hajcak, G. (2011). Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: A combined ERP and fMRI study. *NeuroImage*, 57(4), 1608-1616.

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, 482(7383), 85-88.

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2009). Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to

dopaminergic drug administration. *The Journal of Neuroscience*, 29(5), 1538-1543.

Courville, A., Daw, N., & Touretzky, D. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10, 294-300.

Cropley, V. L., Fujita, M., Innis, R. B., & Nathan, P. J. (2006). Molecular imaging of the dopaminergic system and its association with human cognitive function. *Biological Psychiatry*, 59(10), 898-907.

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

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. *Association for the Advancement of Artificial Intelligence/????*, 761–768.

Denys, D., Zohar, J., & Westenberg, H. G. (2004). The role of dopamine in obsessive-compulsive disorder: Preclinical and clinical evidence. *The Journal of Clinical Psychiatry*, 65, 11.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Learning & Behavior*, 22, 1–18.

Dickinson, A., & Balleine, B. (2004). The role of learning in motivation. *Stevens Handbook of Experimental Psychology*. 3rd ed., Vol. 3.

Eckblad, M., & Chapman, L. J. (1983). Magical ideation as an indicator of schizotypy. *Journal of Consulting and Clinical Psychology*, 51(2), 215.

Foa, E. B., Kozak, M. J., Goodman, W. K., & Hollander, E. (1995). DSM-IV field trial: Obsessive-compulsive disorder: Correction.

Geisler, S., Derst, C., Veh, R. W., & Zahm, D. S. (2007). Glutamatergic afferents of the ventral tegmental area in the rat. *The Journal of Neuroscience*, 27(21), 5730-5743.

Goodman, W. K., McDougle, C. J., Price, L. H., & Riddle, M. A. (1990). Beyond the serotonin hypothesis: A role for dopamine in some forms of obsessive compulsive disorder? *Journal of Clinical Psychiatry*.

Gray, J. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septohippocampal system*. Oxford: Oxford University Press.

Harris, D., & Batki, S. L. (2010). Stimulant psychosis: Symptom profile and acute clinical course. *The American Journal on Addictions*, 9(1), 28-37.

Hockley, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6), 1412.

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

Hull, C. (1943). *Principles of Behavior*. New York:Appleton-Century-Crofts.

Insel, T. R., Mueller, E. A., Alterman, I., Linnoila, M., & Murphy, D. L. (1985). Obsessive-compulsive disorder and serotonin: iIs there a connection? *Biological Psychiatry*, 20(11), 1174-1188.

Iversen, S. D., & Iversen, L. L. (2007). Dopamine: 50 years in perspective. *Trends in Neurosciences*, 30(5), 188-193.

Kapur, S., & Remington, G. (1996). Serotonin-dopamine interaction and its relevance to schizophrenia. *American Journal of Psychiatry*, 153, 466-476.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, 13, 400-408.

Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27, 712–719.

Korotkova, T. M., Sergeeva, O. A., Eriksson, K. S., Haas, H. L., & Brown, R. E. (2003). Excitation of ventral tegmental area dopaminergic and nondopaminergic neurons by orexins/hypocretins. *The Journal of Neuroscience*, 23(1), 7-11.

Levitt, J. J., McCarley, R. W., Dickey, C. C., Voglmaier, M. M., Niznikiewicz, M. A., Seidman, L. J., ... & Shenton, M. E. (2002). MRI study of caudate nucleus volume and its cognitive correlates in neuroleptic-naive patients with schizotypal personality disorder. *The American Journal of Psychiatry*, 159(7), 1190.

Lindström, L. H., Gefvert, O., Hagberg, G., Lundberg, T., Bergström, M., Hartvig, P., & Långström, B. (1999). Increased dopamine synthesis rate in medial prefrontal cortex and striatum in schizophrenia indicated by L-(β-11 C) DOPA and PET. *Biological Psychiatry*, 46(5), 681-688.

Lisman, J., & Otmakhova, N. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: Elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus*, 5, 551-568.

Luo, A. H., Tahsili-Fahadan, P., Wise, R. A., Lupica, C. R., & Aston-Jones, G. (2011). Linking context with reward: A functional circuit from hippocampal CA3 to ventral tegmental area. *Science*, 333(6040), 353-357.

Mannor, S., Simester, D., Sun, P., & Tsitsiklis, J. N. (2004). Bias and variance in value function estimation. *ACM International Conference Proceeding Series*.

Maren, S., Anagnostaras, S. G., & Fanselow, M. S. (1998). The startled seahorse: Is the hippocampus necessary for contextual fear conditioning? *Trends in Cognitive Sciences*, 2, 39–42.

Martig, A. K., & Mizumori, S. J. (2011). Ventral tegmental area disruption selectively affects CA1/CA2 but not CA3 place fields during a differential reward working memory task. *Hippocampus*, 21(2), 172-184.

McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P., Forssberg, H., & Klingberg, T. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science Signalling*, 323(5915), 800.

Menon, M., Jensen, J., Vitcu, I., Graff-Guerrero, A., Crawley, A., Smith, M. A., & Kapur, S. (2007). Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: Effects of dopaminergic modulation. *Biological Psychiatry*, 62(7), 765.

Myin-Germeys, I., Marcelis, M., Krabbendam, L., Delespaul, P., & van Os, J. (2005). Subtle fluctuations in psychotic phenomena as functional states of abnormal dopamine reactivity in individuals at risk. *Biological Psychiatry*, 58(2), 105-110.

Oades, R., & Halliday, G. (1987). Ventral tegmental system: Neurobiology. 1. Anatomy and connectivity. *Brain Research Reviews*, 12, 117-165.

O'Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, 33(5), 815-826.

Omelchenko, N., & Sesack, S. R. (2005). Laterodorsal tegmental projections to identified cell populations in the rat ventral tegmental area. *The Journal of Comparative Neurology*, 483(2), 217-235.

Phillips, A., Vacca, G., & Ahn, S. (2008). A top-down perspective on dopamine, motivation and memory. *Pharmacology, Biochemistry and Behavior*. 90, 236-249.

Phillips, R. G., & LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience*, 106, 274 –285.

Raine, A. (1991). The SPQ: A scale for the assessment of schizotypal personality based on DSM-III-R criteria. *Schizophrenia Bulletin*, 17(4), 555.

Redgrave, P., Prescott, T., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in Neuroscience*, 22, 146-51.

Reed, G. F. (1985). *Obsessional experience and compulsive behavior: A cognitive-structural approach*. Orlando, FL: Academic Press.

Sabatinelli, D., Bradley, M. M., Lang, P. J., Costa, V. D., & Versace, F. (2007). Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex. *Journal of Neurophysiology*, 98(3), 1374-1379.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, 27(5), 763.

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

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

Senior, C. (2003). Beauty in the brain of the beholder. *Neuron*, 38(4), 525-528.

Sesack, S. R., Aoki, C., & Pickel, V. M. (1994). Ultrastructural localization of D2 receptor-like immunoreactivity in midbrain dopamine neurons and their striatal targets. *The Journal of Neuroscience*, 14(1), 88-106.

Smith, A., Li, M., Becker, S., & Kapur, S. (2006). Dopamine, prediction error and associative learning: A model-based account. *Network: Computation in Neural Systems*, 17, 61–84.

Strens, M. (2000). A Bayesian framework for reinforcement learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, 943–950.

Swanson, J., Flodman, P., Kennedy, J., Spence, M., Moyzis, R., Schuck, S., & Murias, M. (2000). Dopamine genes and ADHD. *Neuroscience & Biobehavioral Reviews*, 24, 21-25.

Symmonds, M., & Dolan, R. J. (2011). The neurobiology of preferences. *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms*.

Szechtman, H., Sulis, W., & Eilam, D. (1998). Quinpirole induces compulsive checking behavior in rats: A potential animal model of obsessive-compulsive disorder (OCD). *Behavioral Neuroscience*, 112(6), 1475.

Szechtman, H., & Woody, E. (2004). Obsessive-compulsive disorder as a disturbance of security motivation. *Psychological Review*, 111(1), 111.

Thompson, L. W. (2009). The dopamine hypothesis of schizophrenia. *Perspectives in Psychiatric Care*, 26(3), 18-23.

Thorndike, E. (1932). *The Fundamentals of Learning*. New York: Teachers College Press.

Tsai, H. C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, 324(5930), 1080-1084.

Walter, M., Bermpohl, F., Mouras, H., Schiltz, K., Tempelmann, C., Rotte, M., & Northoff, G. (2008). Distinguishing specific sexual and general emotional effects in fMRI--Subcortical and cortical arousal during erotic picture viewing. *NeuroImage*, 40(4), 1482-1494.

Yui, K., Goto, K., Ikemoto, S., Ishiguro, T., Angrist, B., Duncan, G. E., & Ali, S. F. (1999). Neurobiological basis of relapse prediction in stimulant-induced psychosis and schizophrenia: The role of sensitization. *Molecular Psychiatry*, 4(6), 512.

Appendix

Conditioning tasks can be described as spaces of states and actions governed by Markov Decision Processes (MDPs). An MDP is defined by a quadruple of parameters $(S, A, p_t(s \rightarrow s'), R_{s,a})$, where *S* is the set of states characterizing the task, *A* is the set of actions that the agent can perform, $p_t(s, a \rightarrow s')$ is a transition function that describes the probability of reaching new state *s'* after performing a given action *a* in state s, and $R_{s,a}$ is the function that describes the distribution of possible reward values when action *a* is performed in a particular state *s*. The reward distribution is normal and described by the mean and standard deviation (*μ, σ)*. To select the best action at a given state it is only necessary to compute the Q-values for each state-action pair. The Q-value is defined as:

$$Q(s,a) = E[R_{s,a}] + \gamma \sum_{s'} T(s,a,s') max_a Q(s',a') \quad (1)$$

The reward and transition functions are not known when first encountering the problem and so must be estimated. In order to learn the estimates for these functions, the Bayesian Parameter Estimation (Dearden et al., 1998; Mannor et al., 2004; Strens, 2000) is used for the transition probability function, and the mean and standard deviation of the reward probability function. Each iteration through the MDP provides a new data point for updating the posterior probability density function for each parameter. A uniform prior is used for the single parameter transition function. For the reward function, the

normal distribution reference prior $P(p_r)_i = \frac{1}{\sigma_i}$ is used in order to minimize the effects of any a priori effects on reward estimation (Bernardo, 1979).

The easiest way to conceptualize the learning procedure is to define an agent's action policy in terms of its beliefs, where the beliefs can be updated based on experience using Bayes' equation. There are two important beliefs for the agent in the learning task. The first is the belief that a given action performed in a specific state will transition to a predicted future state. This is the transition model and is denoted as θ_a^s . The probability $p(\theta_a^s)$ is then the belief in the transition model. If the agent performs a given action in a given state it receives feedback *t* which is a binary variable that equals 1 if the expected transition takes place and 0 if it does not. The likelihood function for the transition model given *t* is $p(t \mid p(\theta_a^s)) = p(\theta_a^s)$ and $p(t' \mid p(\theta_a^s)) = 1 - p(\theta_a^s)$. Finally, $p(\theta_a^s)$can then be updated based on *t* using Bayes' equation:

$$P(\theta_a^s \mid t) = \frac{P(t \mid \theta_a^s) P(\theta_a^s)}{\sum_{i=1}^{n} P(t \mid \theta_a^s)_i P(\theta_a^s)_i} \quad (2)$$

The second important belief is the model of the expected reward given a state and action pair. With $R_{s,a}$ being the expected reward, the probability $p(R_{s,a})$ becomes the belief in the modeled reward. With *r* being the actual experienced reward the likelihood function can be calculated as:

$$p(r \mid R_{s,a})_i = \frac{1}{\sigma_i \sqrt{2\pi}} e^{\frac{-(r-\mu_i)^2}{2\sigma_i^2}} \quad (3)$$

where μ and σ are the hypothesized mean and standard deviation of the estimated reward probability function. Once more Bayes' Equation can then be used to update $p(R_{s,a})$:

$$P(R_{s,a} \mid r) = \frac{P(r \mid R_{s,a}) P(R_{s,a})}{\sum_{i=1}^{n} P(r \mid R_{s,a})_i P(R_{s,a})_i} \quad (4)$$

The posteriors are then used as the priors for the next iterations of training for the MDP.

For action policy selection the model uses a probability function of the likelihood for a given action *a* at state *s*. This probability is calculated using the q-value and a softmax function:

$$p(a \mid s)_i = \frac{e^{\frac{Q_i}{\tau}}}{\sum_{j=1}^{n} e^{\frac{Q_i}{\tau}}} \text{ where } \tau = 1 - p(\mu) \quad (5)$$

μ is the expected reward and *p(μ)* is the average probability for μ across all action choices. This means that, as the probability that the expected μ is the correct μ approaches 1 for each state action pair, the policy becomes completely fixed based on the relative Q-values. However, when uncertainty is high about the estimates for μ, then the

policy is more variable. This is a solution to the exploitation/exploration problem that falls naturally out of the state of knowledge about the MDP.