Computational Modelling of the Neural Systems Involved in Schizophrenia

A. J. Thurnham

A thesis submitted in partial fulfilment of the requirements of the University of Hertfordshire for the degree of Doctor of Philosophy

The programme of research was carried out as a joint venture between the Schools of Computer Science and Psychology at the University of Hertfordshire

November 2007

Abstract

The aim of this thesis is to improve our understanding of the neural systems involved in schizophrenia by suggesting possible avenues for future computational modelling in an attempt to make sense of the vast number of studies relating to the symptoms and cognitive deficits relating to the disorder. This multidisciplinary research has covered three different levels of analysis: abnormalities in the microscopic brain structure, dopamine dysfunction at a neurochemical level, and interactions between cortical and subcortical brain areas, connected by cortico-basal ganglia circuit loops; and has culminated in the production of five models that provide useful clarification in this difficult field.

My thesis comprises three major relevant modelling themes. Firstly, in Chapter 3 I looked at an existing neural network model addressing the Neurodevelopmental Hypothesis of Schizophrenia by Hoffman and McGlashan (1997). However, it soon became clear that such models were overly simplistic and brittle when it came to replication. While they focused on hallucinations and connectivity in the frontal lobes they ignored other symptoms and the evidence of reductions in volume of the temporal lobes in schizophrenia. No mention was made of the considerable evidence of dysfunction of the dopamine system and associated areas, such as the basal ganglia. This led to my second line of reasoning: dopamine dysfunction.

Initially I helped create a novel model of dopamine neuron firing based on the Computational Substrate for Incentive Salience by McClure, Daw and Montague (2003), incorporating temporal difference (TD) reward prediction errors (Chapter 5). I adapted this model in Chapter 6 to address the ongoing debate as to whether or not dopamine encodes uncertainty in the delay period between presentation of a conditioned stimulus and receipt of a reward, as demonstrated by sustained activation seen in single dopamine

neuron recordings (Fiorillo, Tobler & Schultz 2003). An answer to this question could result in a better understanding of the nature of dopamine signaling, with implications for the psychopathology of cognitive disorders, like schizophrenia, for which dopamine is commonly regarded as having a primary role. Computational modelling enabled me to suggest that while sustained activation is common in single trials, there is the possibility that it increases with increasing probability, in which case dopamine may not be encoding uncertainty in this manner. Importantly, these predictions can be tested and verified by experimental data.

My third modelling theme arose as a result of the limitations to using TD alone to account for a reinforcement learning account of action control in the brain. In Chapter 8 I introduce a dual weighted artificial neural network, originally designed by Hinton and Plaut (1987) to address the problem of catastrophic forgetting in multilayer artificial neural networks. I suggest an alternative use for a model with fast and slow weights to address the problem of arbitration between two systems of control. This novel approach is capable of combining the benefits of model free and model based learning in one simple model, without need for a homunculus and may have important implications in addressing how both goal directed and stimulus response learning may coexist. Modelling cortical-subcortical loops offers the potential of incorporating both the symptoms and cognitive deficits associated with schizophrenia by taking into account the interactions between midbrain/striatum and cortical areas.

Acknowledgements

Many thanks to the University of Hertfordshire for providing the facilities and academic support over the past six years as I progressed from undergraduate to postgraduate. In particular, I am most grateful for the funding I received during the course of my PhD, without which this would not have been possible. I would also like to thank the Experimental Psychology Society for a Grindley grant to attend CogSci 2005, Boston University for a travel fellowship to attend ICCNS in 2006, and Brain for a travel grant to attend CogSci 2006.

I thank my supervisors John Done, Neil Davey and Ray Frank for their enthusiasm, time, support and encouragement and will miss our weekly meetings, full of lively discussion and laughter. I have been continually inspired by my colleagues Avril, Alison, Kate, Fiona, Samar, Kevin, Jenny, Aruna, Avie, Sarah, Liz and Cathal, but special thanks must be given to Yi and Olivia, who supported me during those difficult times.

Thank you to my family Martin, Katy, Alec, Mum and Dad for their love and support, especially my husband for letting me realise my potential after all these years. The house was not always tidy and meals were often spasmodic, but no one seemed to complain too loudly!

Contents

Appendices

List of Figures

List of Tables

Publications and Presentations

Thurnham, A.J., Done, D.J., Davey, N., & Frank, R.J., A Model of Dopamine and uncertainty Using Temporal Difference. Abstract published in proceedings of *International Conference on Schizophrenia Research*, Colorado Springs, Colorado, March 28 – April 1 2007

Thurnham, A.J., Done, D.J., Davey, N., & Frank, R.J., How Do Computational Models of the Role of Dopamine as a Reward Prediction Error Map on to Current Dopamine Theories of Schizophrenia? In proceedings of *XXV111 Annual Conference of the Cognitive Science Society*, 2006a, p2263-2268, Lawrence Erlbaum Associates

Thurnham, A.J., Done, D.J., Davey, N., & Frank, R.J., A Model of Dopamine and Uncertainty Using Temporal Difference. In proceedings of *XXV111 Annual Conference of the Cognitive Science Society*, 2006b, p2257- 2262, Lawrence Erlbaum Associates

Thurnham, A.J., Done, D.J., Davey, N., & Frank, R.J., A Computational Model of Dopamine for Reward Prediction Error and Uncertainty. Poster presented at the *10th International Conference on Cognitive and Neural Systems*, Boston, Massachusetts, USA, May 17-20, 2006

Thurnham, A.J., Done, D.J., Davey, N., Frank, R.J. & Doughty, O.J., Schizophrenia, Dopamine and Temporal Difference Learning. Abstract published in proceedings of *The International Conference on Schizophrenia Research*, Davos, Switzerland, February 2006

Thurnham, A.J. & Pine, K.J. (2006) The effects of single and dual representations on children's gesture production. *Cognitive Development*. **21(1)**:46-59

Thurnham, A.J., Done, D.J., Davey, N., & Frank, R.J., A Connectionist Model of the Role of Dopamine in Incentive Salience and Temporal Difference Learning. Abstract published in the proceedings of the *XXV11 Annual Conference of the Cognitive Science Society*, Stresa, Italy, 21-23 July 2005

Chapter 1

Overview of Thesis

1.1 Motivation

The ultimate cause or causes of schizophrenia remain elusive in spite of the vast number of studies relating to the symptoms and cognitive deficits of the disorder. Much of the research into schizophrenia has been, and is still, centered on the robust finding that there is a remarkable correlation between the efficacy of antipsychotic drugs in treating psychosis and the ability of those drugs to block the dopamine D2 receptors. However, while it is posited that psychosis results from a dysregulation of the dopamine mesolimbic system (Weinberger 1987; Grace 1991; Kapur & Mamo 2003), there is still little evidence to support this hypothesis and no general consensus on why the medication is effective.

Symptoms and cognitive deficits can be seen as being separable, as various different symptoms can occur without any cognitive deficits, and vice versa. It is likely that symptoms and deficits arise from different brain areas, and this adds to the general difficulty of finding the cause or causes of schizophrenia. It is important to seek a more tractable model of the psychology of schizophrenia: a model of both symptoms and classical cognitive abnormalities incorporating cortical and subcortical systems which are connected by cortico-basal ganglia circuit loops (Alexander & Delong 1985). It may be possible to explain one in terms of the other, or it may be that the two cannot be equated, but computational modelling can help us to address these questions and to reach an answer to the longer term aims of this research, namely:
How can computational models of the neural systems involved in schizophrenia help to improve our understanding of the symptoms and cognitive deficits associated with the disorder?

This thesis aims to improve our understanding of the neural systems involved in schizophrenia by suggesting possible avenues for future computational modelling in an attempt to make sense of the vast number of studies relating to the symptoms and cognitive deficits of the disorder. Cognitive models of molecular theories are in a unique position to combine biological and psychological theories. They offer science through simulation and are a useful tool to help us to answer the challenging question of whether or not there is a simple mechanism on which higher level cognition can be built. However, it should be noted that the modelling in this thesis is a top down approach to providing a possible explanation of what *may* be occurring in the brain and cannot be predictive as there is no attempt to build a detailed replica of brain activity. Such computational insights can be used to generate quantitative findings, providing avenues for further empirical study or treatment strategies, and may contribute to biological theory.

1.2 Organisation of Thesis

This thesis consists of an exploration of some of the existing models relating to the symptoms of schizophrenia and the dopamine system, with a view to building upon them in order to make a contribution to the research area. Time constraints have limited the number of models investigated from this large field of research; however, I have covered three different levels of analysis:

- Abnormalities in the microscopic brain structure.
- Dopamine dysfunction and the neurochemistry of the brain.
- Interactions between cortical and subcortical brain areas, connected by cortico-basal ganglia circuit loops.

This is an account of my progression through the research field.

This multidisciplinary approach comprises of three major relevant modelling themes:

- **Modelling Theme 1:** (Chapter 3) An investigation into the connectionist approaches of:
    - Hoffman and McGlashan (1997) addressing the Neurodevelopmental Hypothesis of Schizophrenia.
    - Braver, Barch and Cohen (1999) addressing the learning and gating roles of dopamine.
- **Modelling Theme 2:** (Chapters 4 to 6) An investigation into the non connectionist Computational Substrate for Incentive Salience incorporating temporal difference (TD) by McClure, Daw and Montague (2003) addressing the dopamine system.
- **Modelling Theme 3:** (Chapters 7 to 8) A novel use of a connectionist model containing dual weights by Hinton and Plaut (1987) to look at the interactions between cortical and subcortical brain areas, connected by cortico-basal ganglia circuit loops.

This has culminated in the production of five models that provide useful clarification, given the complexity of the problem space:

- **Model 1:** (Chapter 3.1) An implementation of the speech perception network of Hoffman and McGlashan (1997) which aimed to test the hypothesis that schizophrenia was associated with reduced cortico-cortical connectivity and therefore may arise from excessive synaptic pruning during adolescence.
- **Model 2:** (Chapter 3.2) A simplified implementation of the feed forward connectionist implementation of Braver et al. (1999) that was able to learn the AX-Continuous Performance Test (CPT).
- **Model 3:** (Chapter 5) My version of the Computational Substrate for Incentive Salience by McClure et al. (2003), which analysed the differences in the patterns of behaviour seen in simulations of animal experiments between high and low dopamine receptor antagonism.
- **Model 4:** (Chapter 6) This was an extension of my third model, designed to address the ongoing debate as to whether or not dopamine encodes uncertainty in the delay period between presentation of a conditioned stimulus and receipt of a reward, as demonstrated by sustained activation seen in single dopamine neuron recordings (Fiorillo, Tobler & Schultz 2003; 2005).

- **Model 5:** (Chapter 8) This used the ideas behind the dual weights model of Hinton and Plaut to address the problem of arbitration between two systems of control, and describes how both inflexible stimulus-response actions and flexible goal-directed learning may operate together in one simple artificial neural network (ANN) containing dual weights.

The literature review is contained in chapters 2, 4 and 7:

- In Chapter 2 I give a detailed account of the motivation behind this thesis. I provide a basic description of schizophrenia, how it manifests itself in the human body and its biological underpinnings. I also outline the difficulties in finding the ultimate cause of the disorder and explain the advantages of using ANN models for this purpose.
- The theory described in Chapter 4 follows as a result of preliminary modelling of the Neurodevelopmental Hypothesis of Schizophrenia in Chapter 3. Here I explain the theory behind my TD model of phasic dopamine neuron firing, developed in Chapters 5 and 6.
- In Chapter 7 I identify the limitations of using TD alone to account for a reinforcement learning account of action control in the brain, and explain the reasoning behind the development of my dual weighted ANN in Chapter 8.

Modelling Theme 1

My initial instinct was to search for an existing connectionist model that could be used as a base for my future modelling and in Chapter 3 I describe my attempts at implementing two existing connectionist models. The first, a speech perception network by Hoffman and McGlashan (1997) designed to give rise to hallucinations, one of the major symptoms of schizophrenia (Section 3.1), and the second, a simulation of one of the cognitive deficits identified in schizophrenia relating to maintaining and updating working memory, demonstrated in the AX-CPT (Braver et al. 1999) (Section 3.2). While I did not pursue these two models further in this thesis, the modelling in this chapter made a valuable contribution to the direction of the remainder of this body of research. The flaws of these models, detailed in Chapter 3, pointed to a new line of reasoning to account for the symptoms and cognitive deficits of schizophrenia which included dopamine dysfunction and the neuroscience of the cortico-basal ganglia circuit loops, on which the remainder of this thesis rests.

Modelling Theme 2

Chapter 4 contains the second part of the literature review. Having justified switching to a new line of reasoning involving the dopamine system, I investigate a body of research inspired by the physiological recordings of dopamine neurons on alert monkeys, by Wolfram Schultz and colleagues (Hollerman & Schultz 1998; Ljunberg, Apicella & Schultz 1992; Romo & Schultz 1990; Schultz 1986; Schultz, Apicella & Ljunberg 1993; Schultz & Romo 1990) who showed that information about rewarding stimuli was encoded in dopaminergic activity. This includes:

- The idea of dopamine acting as a reward prediction error (Section 4.1).
- TD as an effective method of modelling the dopamine reward prediction error signal (Section 4.2).
- The role of the basal ganglia in the production of that signal (Section 4.3).
- The Incentive Salience Hypothesis (Section 4.4).

- Evidence of the dopamine reward prediction error in humans from fMRI studies (Section 4.5).

This theory is encompassed in the Computational Substrate for Incentive Salience by McClure et al. (2003), an interesting exception to the typical actor-critic in that it is capable of addressing free-operant behaviour (Niv, Daw, Joel & Dayan 2007). This model is described in detail in Section 4.2 and I implement a version of this in Chapter 5.

Chapter 5 contains a detailed description of the modelling of two sets of animal experiments by McClure et al. (2003) (Section 5.1). Firstly, they modelled the effects of a large concentration of dopamine receptor antagonist, in accordance with Ikemoto and Panksepp (1996) and secondly, they modelled the different effects of lower concentrations of dopamine receptor antagonism in a similar experiment by Wise et al. (1978). The difference in the pattern of the dopamine response between these two simulations provided evidence for the dual role of the dopamine reward prediction error as (i) a learning signal, and (ii) in action selection. In Section 5.2 I describe my own implementation based on the original model which replicates and explores the Computational Substrate for Incentive Salience by McClure et al. (2003). My simulations allow me to explore the changing parameters of the model, and to answer a number of interesting research questions that warrant further investigation.

In Chapter 6 I extend the model developed in Chapter 5 to address the ongoing debate as to whether or not dopamine encodes uncertainty in the delay period between presentation of a conditioned stimulus and receipt of a reward, as demonstrated by sustained activation seen in single dopamine neuron recordings (Fiorillo et al. 2003; 2005; Niv, Duff & Dayan 2005). The key to this debate appears to be how frequently sustained activation occurs in individual trials. Furthermore, if sustained activation is greatest with maximum uncertainty then it could be said that dopamine is encoding uncertainty in the delay period as a product of TD. This alternate model permits this valuable single trial analysis and allows me to make predictions that could, in theory, be tested and verified by experimental data. The simulations detailed in this chapter are an example of science through simulation and demonstrate how computational modelling can help to clarify a position by generating testable predictions. The arguments presented will strengthen the use of TD as a valid method of modelling and quantifying the dopamine reward prediction error.

Modelling Theme 3

Chapter 7 marks a change in the direction of modelling and contains the third part of the literature review. While the focus of this thesis in Chapters 4 to 6 has centred on using TD as a model of dopamine function in reinforcement learning, it is important to note that there are limitations to using a model free reinforcement learning paradigm, such as TD, as a complete account of action control in the brain. In this chapter I draw attention to those limitations and point to some alternate models that address these shortfalls. This will allow me to shift the focus of this thesis away from TD and back to my longer-term aim of improving the understanding of the neural systems involved in schizophrenia; taking into account a wider perspective of the interactions between cortical and subcortical brain areas, connected by cortico-basal ganglia circuit loops (Chapters 2.2 and 4.3).

I begin Chapter 7 by looking at the alternative model based account of dopamine function by Smith, Li, Becker and Kapur (2004; 2006) in Section 7.1. Their line of reasoning bears important similarities to mine as we both have the long-term goal of a better understanding of schizophrenia through the ideas of incentive salience and dopamine dysfunction. Section 7.2 contains a brief account of a new algorithm developed by O'Reilly, Frank, Hazy and Watz (2007) as an alternative to TD, which is not developed further in this thesis, and in Sections 7.3 and 7.4 I look at two models by Dayan and Balleine (2002) and Daw, Niv and Dayan (2005) that combine the benefits of both model free and model based accounts of reinforcement learning. Finally, in Section 7.5 I describe a framework by Yin and Knowlton (2006) for the role of the basal ganglia in habit formation that distinguishes between goal directed actions and stimulus response habits via separate networks that correspond to different, but interconnecting cortico-basal ganglia loops.

As well as offering alternatives to TD, Sections 7.1 to 7.5 all point to the concept of a dual system of control (Section 7.6): Smith et al. (2006) modelled both phasic dopamine in the learning process and tonic dopamine in the expression of previously acquired behaviour; the PVLV algorithm by O'Reilly et al. (2007) contains two different systems, where the PV system controls performance and learning during primary rewards and the LV system learns about conditioned stimuli; Dayan and Balleine (2002) distinguished between modelling Pavlovian and instrumental conditioning; and both Daw et al. (2005) and Yin and Knowlton (2006) distinguish between two different networks for goal directed actions and stimulus response habits. In particular, Daw et al. (2005) referred to the competition between multiple systems for behavioural choice in the brain, and the problem of arbitration between the systems when they disagreed. They suggested a model of dual action choice, where the systems operated separately and in parallel, governed by a Bayesian principal of arbitration. Their model based habit system of caching values was not immediately sensitive to the specific outcome information associated with animal devaluation experiments as it took time for a change in behaviour to occur following relearning of the values. Alternatively, their flexible model based system that was outcome sensitive and goal directed showed an immediate behavioural change.

In Chapter 8 I return to a connectionist approach and develop a model with dual weights which is capable potentially of implementing a dual system of control. The model has several advantages over the model by Daw and colleagues in that it is a biologically inspired connectionist application that will allow for interactions between the two controllers, without the need for an arbiter or homunculus. In this third line of reasoning I adapted an existing dual weighted ANN by Hinton and Plaut (1987), originally designed to address the problem of catastrophic forgetting in multilayer ANNs. I suggest an alternative use for a model with fast and slow weights to address the problem of arbitration between two systems of control through rapid learning in the fast weights, which will allow for more rapid changes in the environment than in a standard network with one set of weights. Furthermore, I describe how this novel approach is able to combine the benefits of model free and model based learning in one simple model.

In Section 8.1 I refer to the general problem of arbitration between two systems of control and in Section 8.2 I introduce the original dual weighted model by Hinton and Plaut, before developing my own constrained version of the model in Section 8.3. Here I conduct a series of experiments where I investigate the parameters of the model, the contribution of the fast weights and the interactions between fast and slow weights. This enables me to describe the mechanisms behind the dual system of control in Section 8.4 and its advantages over the model by Daw and colleagues in Section 8.5. Future improvements to the model, suggested in Section 8.6 may have important implications in addressing how both goal directed and stimulus response learning may coexist.

Finally, in Chapter 9 I summarise the contributions of this thesis (Section 9.1), draw my conclusions (Section 9.2), and explain how future improvements to those models could help to improve our understanding of the neural systems involved in schizophrenia (Section 9.3).

Chapter 2

Modelling Schizophrenia

In this Chapter I provide a basic description of schizophrenia, how it manifests itself in the human body and its biological underpinnings. I also outline the difficulties in finding the ultimate cause of the disorder and explain the advantages of using artificial neural network models for this purpose.

Schizophrenia is a disorder of the human brain with many different symptoms and cognitive deficits, which may or may not occur, in many different combinations. One interpretation of the symptoms is that they fall within three syndromes: (i) Reality Distortion (*positive* symptoms such as hallucinations and delusions), (ii) Psychomotor Poverty (*negative* symptoms such as flat effect or affective unresponsiveness) and (iii) Disorganisation (e.g., thought disorder, a disturbance in the form of thinking which manifests itself as a loss of intelligibility of speech) (Liddle 1996). It has also been claimed that patients exhibit a general cognitive deficit, with impairments in executive function, memory and attention, over and above this general level (Bilder et al. 2000; McKenna 1997). In particular, it is seen as a disorder in which patients fail to make appropriate use of context, due to a failure to internally represent, maintain and update task relevant information (Cohen & Servan-Schreiber 1992; Braver, Barch & Cohen 1999). Diagnosis generally involves the presence of at least two positive symptoms plus the absence of significant manic-depressive mood changes (McKenna 1997).

Although there are a vast number of studies relating to the symptoms and cognitive deficits of schizophrenia, the ultimate cause or causes of the disorder remain elusive. Historically, some have considered a psychodynamic or Freudian approach to understanding the aetiology of the disorder: a manifestation of the conflict in the inner self. However, current research seeks a biological approach and looks for a biological brain disorder. It is the biological approach that is developed in this thesis and Section 2.1 provides a brief account of some of the changes recorded in the schizophrenic brain. Section 2.2 describes the difficulties modelling schizophrenia and of finding a framework that incorporates both symptoms and cognitive deficits, and in Section 2.3 I explain how artificial neural network models can be used as a tool to help identify the causes, as they offer a noninvasive testing ground for theories by providing a link between the behaviour, and the biology of the schizophrenic. Such new models will generate quantitative findings, providing avenues for further empirical study or treatment strategies.

2.1 Schizophrenia: the Biological Approach

This thesis takes a biological approach towards explaining the aetiology of schizophrenia. Unlike neurodegenerative diseases, such as Alzheimer's and Parkinson's, there are few obvious biological disturbances to the brains of schizophrenic patients, and so it is hard to ascertain the underlying causes of these symptoms and deficits. Much of the research into schizophrenia has been, and is still, centered on the robust finding that there is a remarkable correlation between the efficacy of antipsychotic drugs in treating psychosis and the ability of those drugs to block the dopamine D2 receptors. So, while it is posited that psychosis results from a dysregulation of the dopamine mesolimbic system (Weinberger 1987; Grace 1991; Kapur & Mamo 2003), there is still little evidence to support this hypothesis.

In this section I discuss the limited evidence for brain disturbance in patients with schizophrenia. Research shows that the disorder has: (i) a known genetic component (Section 2.1.1); (ii) a few limited abnormalities in the macroscopic brain structure (Section 2.1.2) and (iii) changes in the neurochemistry of the brain (Section 2.1.3). These differences provide clues as to what processes underlie the disorder and offer explanations for the emergence of symptoms and cognitive deficits.

2.1.1 Genetic Component

With regard to the genetic component, Gottesman (1991) produced a table of risks, based on around forty studies for developing schizophrenia in first, second and third degree relatives of patients, suggesting a genetic role to the disorder. In particular, these results showed that if one of a twin with the same genes (monozygotic) developed schizophrenia there was a greater likelihood of the other twin developing the disorder than if the twins had different genes (dizygotic). However, while genetics has a role to play, this can only be a predisposition as there is no genetic certainty of developing the disorder and 63% of patients will have no family history at all (McKenna 1997).

2.1.2 Structural Changes: The Neurodevelopmental Hypothesis

There are some brain changes seen in schizophrenics, but compared to patients with Alzheimer's or Parkinson's disease, these changes are often inconsistent across studies and are not always obvious at a macroscopic level. For example, using computer assisted tomography (CT) Johnstone et al. (1976) identified lateral ventricular enlargement in schizophrenia, but the differences were small and found largely in male patients (McKenna 1997). Meta-analysis also points to reduced cerebral (cortical and hippocampal) volume (Egan & Weinberger 1997; Harrison 1999) that is present even in first-episode patients (Lim et al. 1996). Post-mortem studies have identified increased neuronal densities in schizophrenics in prefrontal and occipital cortex. This is not believed to be due to a loss of neurons but to a reduction in the interneuronal spaces, the neuropil, consisting of neuronal processes and synaptic contacts (Selemon, Rajkowski & Goldman-Rakic 1995).

Hoffman and McGlashan (2001) posited that schizophrenia may result from a pathological extension of normal reductions of neuropil and synaptic density during adolescent development, which also accounts for the characteristic age of onset of schizophrenia in the late teens and early twenties. This neurodevelopmental hypothesis of schizophrenia as a disorder arising from aberrant brain development prior to the emergence of symptoms is developed further in Chapter 3, where I reproduce a speech perception network by Hoffman & McGlashan (1997) designed to produce speech hallucinations by synaptic pruning.

2.1.3 Neurochemical Changes: The Dopamine Hypothesis

It is not immediately apparent from a neurodevelopmental point of view how aberrant brain development transfers to the diverse and changeable symptoms and cognitive deficits associated with schizophrenia. Taking a different perspective and looking at the chemistry of the brain, neurochemical dysfunction could provide an answer. Drugs of abuse, such as amphetamine, LSD and PCP are known to induce abnormal mental states similar to some of the positive symptoms of schizophrenia (e.g., Ikemoto & Panksepp 1999; Carlsson et al. 2001; Smith, Becker & Kapur 2005). In addition, a biochemical process will account for the fact that symptoms both appear and disappear, and wax and wane in intensity.

The dopamine hypothesis of schizophrenia arose in the 1960's as a result of two different observations: (i) antipsychotic drugs prescribed to alleviate psychosis provide their effect by blocking dopamine receptors; and (ii) exposure to dopamine receptor agonists, such as amphetamine, induces psychosis (Abi-Dhargham 2004). In view of the remarkable correlation between the ability of antipsychotic drugs to block dopamine D2 receptors and the effectiveness of those drugs in the treatment of psychosis, it is posited that psychosis results from a dysregulation of the dopamine mesolimbic system (Weinberger 1987; Grace 1991; Kapur & Mamo 2003). A critical role for dopamine is its contribution to conferring reward during learning, or which choice gives the greatest pay back when making decisions (Chapter 4.1). More specifically it appears to be involved in attributing incentive salience to things we see or hear, or thoughts we generate ourselves (Chapter 4.4). A malfunction could then lead to us thinking that irrelevant ideas are suddenly really important, leading to feelings of persecution, or that a rustle of leaves may be a sign from God. Neurotransmitters such as dopamine are genetically controlled and could easily arise as expression of a genetic fault or disposition (McKenna 1997). The dopamine hypothesis is developed further in Chapter 4 and the remainder of this thesis relates to the function of the dopamine system.

While this thesis focuses largely on the dopamine system there are many other neurotransmitters systems operating in concert in the brain at any one time. I do not pursue these alternatives further but it is important to note that any one transmitter will not be working in isolation. For example, the NMDA hypothesis of schizophrenia arose from observations that drugs such as PCP and ketamine (NMDA antagonists) lead to schizophrenia-like effects, in particular negative symptoms as well as hallucinations, delusions, thought disorder (Stone, Morrison & Pilowsky 2007). This hypothesis seeks to explain some of the gaps in the dopamine hypothesis regarding the treatment-resistant negative symptoms and the onset of the disorder in late teens/early twenties. Carlsson et al. (2001) also take a wider view and see dopamine as one of many possible dysfunctional neurotransmitters affected in the brain in schizophrenia. Pharmacological evidence suggests that small differences in the fragile balance between multi-neurotransmitters at various points in local cortical microcircuits leads to many of both the positive and negative symptoms associated with the disorder. They posit that although there may be an elevated baseline release of dopamine in schizophrenia, it is possibly secondary to hypoglutamatergia.

Finally, while it is generally accepted that dopamine has an important role to play in the manifestation of schizophrenia, due to the robust finding that there is a striking correlation between the efficacy of antipsychotic drugs in treating psychosis and the ability of those drugs to block the dopamine D2 receptors, it should be noted that to date there is generally little evidence of dopamine receptor abnormality in schizophrenia (Stone et al. 2007). Although there is some evidence for a small elevation in D2 receptors in drug free patients (Laurelle 1998; Zakzanis & Hansen 1998).

2.2 A Model of Symptoms and Cognitive Deficits

It is known that patients with schizophrenia suffer from a wide-spread cognitive dysfunction that affects memory, executive functioning and attention and there seems to be a dissociation between these cognitive deficits and psychotic symptoms (delusions, hallucinations). The former occur well in advance of the onset of symptoms, and the trajectory of symptom recovery is not matched by cognitive recovery (Harvey, Koren, Reichenberg & Bowie 2005). Symptoms and cognitive deficits can be seen as being separable as various different symptoms can occur without any cognitive deficits and vice versa.

Some modellers have tried to explain symptoms in terms of neuropsychological impairment based on the neurodevelopmental hypothesis of schizophrenia (Section 2.1.2), for example, Hoffman and McGlashan (1987) developed a speech perception network that gave rise to hallucinations as a result of synaptic pruning (Chapter 3.1). However they did not address the issue of cognitive deficits. Others have focussed on the dopamine hypothesis of schizophrenia (Section 2.1.3) and modelled cognitive deficits, for example, Cohen and Servan-Schreiber (1992) looked at a dysfunction in working memory. They used artificial neural networks to simulate normal and schizophrenic performance in three tasks that relied on the correct use of context (Chapter 3.2). However, they did not address the various symptoms of the disorder. These connectivity and dopamine based accounts of impairment are two examples of connectionist modelling at different levels of analysis, which cannot easily be equated; although it is possible that dopamine, at a lower neurochemical level than the connectivity between neurons, could be accommodated within the neurodevelopmental hypothesis.

It is likely that symptoms and deficits arise from different brain areas and this adds to the general difficulty of finding the cause or causes of schizophrenia. It is posited that one of the symptoms, psychosis, is a state of aberrant salience, where excess levels of dopamine are no longer stimulus-linked and context-driven. Delusions (paranoia, aliens interfering with one's brain), and hallucinations (hearing voices), may arise then as a result of the individual attempting to provide their own explanations for experiences which come out of the blue and are imbued with high importance (Kapur 2003) (Chapter 4.4.1). This is in keeping with an earlier theory of schizophrenia by

Maher (1988) that patients make normal attributions, or reasoned normally to abnormal experiences. This would involve a subcortical abnormality, relating to the mesocorticolimbic dopamine pathway from the ventral tegmental area to the prefrontal cortex, hippocampus, amygdala and nucleus accumbens (Chapter 4.3), with normal cortical function.

On the other hand cognitive dysfunction would appear to be associated with cortical areas (Abi-Dhargham 2004; Winterer & Weinberger 2005). Traditional cognitive models of schizophrenia based on cognitive dysfunction in executive dysfunction/memory/attention have poor face validity when used to explain the spontaneous experiences (delusions/hallucinations) which are bizarre, or strange, since these are unrelated to past experience and stored memories (Simpson, Done, Valeé-Tourangeau 2002). It is important to seek a more tractable model of the psychology of schizophrenia: a model of both symptoms and classical cognitive abnormalities. Dopamine abnormalities in the cortex, in particular the dorsolateral prefrontal cortex, would not only account for the neuropsychological deficits found in schizophrenia but they could also integrate the abnormal symptomatic experiences into dysfunctional attributional, executive and memory systems. These dual roles pertaining to symptoms and cognitive deficits can be equated crudely as being due to dopamine abnormalities in the midbrain/striatum (Chapter 4.3) and cortex respectively. The interaction between these different levels via cortical-subcortical loops (Alexander & Delong 1985) means that they are able to operate in consort. This would permit a more tractable model of the psychology of schizophrenia: a model of both symptoms and classical cognitive abnormalities.

In conclusion, there are two highly complex aspects of schizophrenia that need to be addressed in the attempt to make sense of the disorder: symptoms and cognitive deficits. An important question arises of how both symptoms and cognitive deficits fit together in a framework of explanation. It may be possible to explain one in terms of the other, or it may be that the two cannot be equated. To address the problem it is necessary to reduce the complex to manageable portions, and Section 2.3 describes how neural network modelling can help.

2.3 Advantages of Neural Network Models of Schizophrenia

Multidisciplinary research can often overcome the limitations of a single discipline by introducing novel techniques, such as the use of artificial neural networks, a useful tool for exploring the relationship between neurobiology and computational performance. By uniting neurobiology, neuropsychology, cognitive and computational science, neural networks provide an important link between the physical brain (in terms of different brain regions through to individual neurons and the smaller chemical elements such as neurotransmitters) and behaviour (our thoughts, plans, decisions and actions). They offer a non-invasive testing ground for theories by modelling the nervous system effectively at many different structural levels, including the biophysical, the circuit and the systems levels. Such new models will generate quantitative findings, providing avenues for further empirical study or treatment strategies.

Neural network models are a valuable tool for exploring complex systems with large volumes of data, and enable the exploration of relationships in a way that may not be possible in vivo. Given the complicated systems involved in the brain, it becomes very difficult to control different factors and to pin down the deficit. The more complex the system, the more the situation is inherently uncontrollable and the less it is amenable to standard experimental method. It is often the case that manipulating one variable at a time will have a downstream knock-on effect and there becomes a danger of changing the nature of the whole system. Neural network models are inspired by biological brain systems and enable the simulation of complex and distributed systems in the brain by manipulating one or more parameters to seek an optimal combination. They let us explore changing systems by looking at the larger picture and generating data that can be tested empirically.

However, it is also necessary to understand the limitations of computer modelling. By its definition a model is an abstraction of the system or concept under scrutiny and, being underconstrained, it can never claim to be the real thing. Therefore, models are informative rather than definitive; qualitative rather than quantitative, and as such cannot be predictive. In order to combat the risks associated with speculation it is necessary for models to incorporate a wide range of empirical data spanning many different levels of analysis (O'Reilly 2006). In addition, the idea is to seek a broad qualitative correspondence between the model and experimental data that generalises to other experiments with minimal alteration to the model (Smith, Li, Becker & Kapur 2007). The modelling in this thesis involves top down modelling in an attempt to provide a possible explanation of what *may* be occurring in the brain and there is no attempt to build a detailed replica of brain activity.

The dopamine system is an example of such a complex system that is difficult to understand through conventional methods. The tools available to us, such as brain imaging and animal studies have their own limitations. Images of our own brains are useful when there is something wrong with one part of our brain, such as in a stroke, but they are less informative when there is a malfunctioning of a system distributed throughout the brain. While studies of animals whose brains are structurally different from our own can make a valuable contribution to the subject area, for example amphetamine induced psychosis (Section 2.1.3); there are limitations to the extent that they can be used as a full explanation of a human disorder, like schizophrenia. Modelling brain systems using computer based, or artificial neural network models offer a useful alternative providing the model is both biologically and psychologically plausible. Computer algorithms such as Temporal Difference (TD) operate very much like the dopamine system in our brain (Chapter 4.2). One great advantage of using these artificial models is that an unlimited number of experiments can be conducted, including lesioning the network, which cannot be done on human brains for ethical reasons. Although artificial neural networks will not provide answers to the causes of schizophrenia, their heuristic value permits us to generate new ways of thinking about the disorder, and also provides guidance on new experiments which can then be carried out with patients.

A review by Cohen, Braver and Brown (2002) gives a detailed account of the variety of computational models that exist at different levels of analysis pertaining to dopamine function in the prefrontal cortex. These include connectionist models of the neuromodulatory function of dopamine (Cohen & Servan-Schreiber 1992) (see Chapter 3.2); biophysically detailed models incorporating the electropyhsiology of a neuron simulating the effect of dopamine on performance in cognitive tasks that rely on prefrontal cortex function (e.g., Durstewitz, Kelc & Gunturkun 1999; Brunel & Wang 2001) and connectionist models of the role of dopamine in learning and updating working memory (e.g., Braver, Barch & Cohen 1999) (see Chapter 3.2). Cohen et al. stress the value of a multilevel approach to model building as the biophysical models are useful in assessing the abstractions used in the connectionist models, and the connectionist models may help guide future research on the biophysical processes that underlie the basic mechanisms.

Alternatively, Montague, Hyman and Cohen (2004) review the use of the Temporal Difference algorithm (Sutton 1988; Sutton & Barto 1998) (see Chapter 4.2) to model the role of dopamine as a reward prediction error (see Chapter 4.1) incorporating basal ganglia (see Chapter 4.3). Attempts have been made to incorporate both the cortical and subcortical effects of dopamine function in one model, for example, the Prefrontal Basal Ganglia Working Memory connectionist model of learning by O'Reilly & Frank (2006) and O'Reilly, Frank & Hazy (2007). However, much work remains to be done to apply this framework to various working memory tasks to test the cognitive neuroscience validity of the model. A neural network specifically designed to incorporate cortical and subcortical areas could, in theory, generate testable predictions concerning the symptomalogy and cognitive dysfunction seen in schizophrenia.

2.3.1 Connectionist Networks

Biologically inspired connectionist models are particularly suitable for modelling schizophrenia. They are a collection of artificial neurons (simplified versions of a biological neuron) with modifiable connections (representing synapses) between them. The brain consists of densely interconnected neurons in layers carrying signals in parallel and a connectionist architecture containing artificial neurons allows the integration and transfer of information from neuron to neuron in a similar manner. Examples of connectionist architectures can be found in Chapters 3 and 8, and an example of an artificial neuron is given in Figure 2.1., where the inputs and single output are analogous to the dendrites and axon in a biological neuron. Input is introduced to the neuron representing either external stimuli or input from other afferent neurons. This input is summed and if the total signal exceeds a predetermined threshold the signal will be passed as a single output to an efferent neuron(s). The strength of the connections (weights) between neurons is adjusted according to a learning rate permitting learning to occur from experience (a set of training patterns). Knowledge is held in the weights and is distributed across many neurons and many connections.

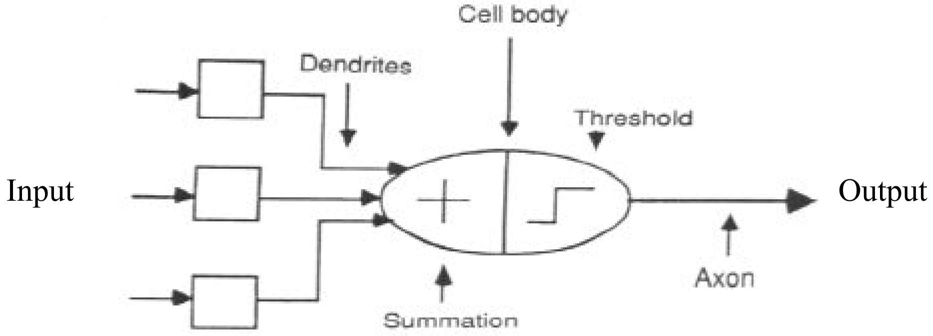


Figure 2.1. A schematic representation of an artificial neuron, where total input to the neuron is summed. If the total exceeds a predetermined threshold, the signal will be passed to an efferent neuron(s).

Connectionist models are metaphors of the brain because they are able to solve similar problems and because of their structural resemblance (Aakerlund & Hemmingsen 1998). They possess some of the advantages of the human brain, namely: (i) they are distributed in nature and process information in parallel, (ii) they have the ability to learn from experience, (iii) they are able to generalise to new situations by applying information from past experience, and (iv) they are fault tolerant and therefore resistant to damage (McLeod, Plunkett & Rolls 1998). By altering the network architecture, the numbers of neurons and learning rules to suit the occasion, connectionist models are flexible and capable of modelling a broad variety of human performance tasks. In particular, a connectionist system comprising of a collection of individual simplified artificial neurons brought together and working in parallel in a distributed manner to make a collective whole, is able to find patterns in otherwise streams of seemingly meaningless data. For example a newborn baby, incapable of speech, capturing exponentially over time the subtle patterns and nuances of a language to form words and sentences as a child, and even improving further into adulthood, with increasing vocabulary. In a similar manner a novice sports players actions will progress from random and awkward and improve over time to proficient and expertise, with practice.

2.4 Chapter Conclusions

This chapter contained a detailed account of the motivation behind this thesis:

- It provided a basic description of schizophrenia, how it manifests itself in the human body and its biological underpinnings.
- I outlined the difficulties in finding the ultimate cause of the disorder.
- I explained the advantages of using computational modelling for this purpose; in particular, a biologically inspired connectionist approach.

Cognitive models of molecular theories are in a unique position to combine biological and psychological theories. They offer science through simulation and are a useful tool to help us to answer the challenging question of whether or not there is a simple mechanism on which higher level cognition can be built.

Chapter 3

In Search of a Connectionist Model of Schizophrenia

The initial aim of this three year research period was to explore the application of connectionist models as a paradigm for schizophrenia, with a view to generating and testing theories of the disorder. There were few existing connectionist models on which to use as a base for my research, but I was drawn initially to an interesting application by Hoffman and McGlashan (1997) that claimed to simulate hallucinated voices (Section 3.1). This model was based on the neurodevelopmental hypothesis of schizophrenia and is described in detail in Section 3.1.1. My own simulations have allowed me to make a critical analysis of the model and have highlighted the limitations of a model of schizophrenia that does not incorporate dopamine dysfunction.

Following my decision to abandon further attempts to simulate the Hoffman and McGlashan speech perception network, I was actively seeking a new model as a base for my future work. Some of the early connectionist models addressing the dopamine hypothesis of schizophrenia are discussed in Section 3.2, but I was drawn to a model by Braver, Barch and Cohen (1999) which identified a new Learning and Gating theory of dopamine, incorporating phasic dopamine firing patterns. I implemented a simplified simulation of the model performing the AX-CPT in Section 3.2.1.

This chapter gives an indication of some of the early models of schizophrenia that were available, and the preliminary modelling I have performed in this chapter has provided the direction for the remainder of this thesis.

3.1 Modelling the Neurodevelopmental Hypothesis of Schizophrenia

In the late 1990's early 2000's neuropsychological models offered the potential to explore both symptoms and neuropsychological impairments as a window on the brain mechanisms of the schizophrenic. An interesting connectionist model came to my attention by Hoffman and McGlashan (1997) that claimed to simulate hallucinated voices. Their research focused on the Neurodevelopmental Hypothesis of schizophrenia, reflecting a popular approach at that time, temporarily dominating the dopamine hypothesis. While dopamine dysfunction was known to be associated with schizophrenia, it was not considered to be central to the etiology as: (i) Antipsychotic drugs targeting the dopamine system did not appear to be the perfect answer, as not all patients responded to drugs designed to reduce excess dopamine by blocking dopamine D2 receptors; (ii) Dopamine was unlikely to be the only neurotransmitter showing dysfunction in schizophrenia. For example, the use of phencyclidine (PCP, 'angel dust') gave rise to psychotic symptoms and could also be used as a model for schizophrenia. However, this was known to be a powerful antagonist of the glutamate receptor subtype NMDA and did not affect the dopamine system (Carlsson et al. 2001).

One of the strengths of the neurodevelopmental hypothesis was that it addressed the question of why the emergence of schizophrenia peaked in the late teens and early twenties, but did not show earlier in life. Using normal postmortem tissue from the middle frontal cortex, synaptic density was shown to peak during childhood, with a subsequent decline of between 30 to 40% during adolescence to reach adult levels (Huttenlocher 1979). It was posited that schizophrenia was associated with reduced cortico-cortical connectivity and therefore may arise from excessive synaptic pruning (Hoffman & McGlashan 1997). This would result in molecular and histogenic responses that would cumulatively lead to different developmental trajectories from those seen in a normal brain. In support of this hypothesis post-mortem studies have identified increased neuronal densities in schizophrenics in prefrontal and occipital cortex. This is not believed to be due to a loss of neurons but to a reduction in the interneuronal spaces, the neuropil, consisting of neuronal processes and synaptic contacts (Selemon, Rajkowski & Goldman-Rakic 1995). It is hypothesised that this reduction in neuropil could underlie abnormalities in information processing and cognitive dysfunction seen in schizophrenia (Selemon 2004). In particular, Hoffman and McGlashan (2001) posited that schizophrenia may result from a pathological extension of normal reductions of neuropil and synaptic density during adolescent development, which also accounts for the characteristic age of onset of schizophrenia in the late teens and early twenties.

An investigation of the Hoffman and McGlashan model was considered to be an ideal starting point for a body of research using artificial neural networks to look into schizophrenia. As an experimental method connectionist models are perfectly placed to investigate synaptic pruning as weight connections can be easily removed to simulate both pruning and cell death. It was hoped that building upon the work of Hoffman and McGlashan would provide a good opportunity to tell us something new about schizophrenia.

3.1.1. Simulation of a Speech Perception Neural Network (Hoffman & McGlashan 1997)

The Hoffman and McGlashan simulation was not a model of the neurodevelopmental hypothesis of schizophrenia, but it did focus on the product of that hypothesis. The emphasis was on neuropsychological dysfunction, specifically the changes to the dorsolateral prefrontal cortex reflected during working memory type tasks, rather than symptoms. They produced an artificial neural network trained to identify the semantics of words in sentences from a limited vocabulary of thirty words. The network was recurrent and so word order gave rise to a rudimentary working memory. By testing the effect of degraded and undegraded test sentences on both grammatical word sequences and randomised word sequences, Hoffman and McGlashan demonstrated dependence of the network on word order in decoding input information. They modelled both loss of synapses and cell death on the fully trained network. The former involved clamping to zero the absolute values of connection weights, linking the temporary and hidden layers, that fell below a threshold; and the latter where certain hidden layer neurons were eliminated by clamping their levels of activation to zero. They demonstrated that eliminating up to 65% of these working memory connections improved performance, but beyond that level of pruning, performance was adversely affected and, with synaptic elimination, speech hallucinations emerged at synaptic losses of around 80 to 95%. Their implied scaling claimed that an optimal 64% reduction in working memory connections corresponded to an overall synaptic reduction of 29% when all the connections in the network were taken into account, and that this figure approximated the 30-40% reduction of frontal area synapses from childhood to adult, identified from postmortem studies (Huttenlocher 1979). In the model cell death produced no hallucinations and they concluded that psychosis may arise from synaptic elimination.

In the first instance I decided to attempt to replicate the Hoffman and McGlashan model and pursue their argument of schizophrenia as a neurodevelopmental disorder resulting from irregularities in synaptic pruning during adolescence. It was my intention to use this information to implement my own model in order that I may model speech perception and carry out a series of experiments to explore the emergence of hallucinated voices from a connectionist model.

METHODS

In order to replicate Hoffman's results I needed answers to many questions that were not addressed in the original journal paper. For example, the criteria for assessing correct identification and misidentification of words by the network, how to separate the sentences in the input; feature coding; learning rate parameters etc. I contacted Hoffman who was good enough to send me details of the original program in Q BASIC, which was designed to generate the data and run the simulation. As we did not have the facilities to use Q BASIC it was necessary to convert the program into VISUAL BASIC, and some further modifications were necessary before it was possible to recreate the original data sets.

Preferring to use more sophisticated software packages, simulations were implemented in PDP++ (O'Reilly and Munakata, 2000). A four-layer speech perception neural network with 148 neurons was implemented, exactly the same as Hoffman and McGlashan (see Figure 3.1). The network had 25 inputs, 43 outputs and a temporary storage layer of 40 neurons (layer 3) that received a copy of the activation of the 40 hidden layer neurons (layer 2), with the aim of learning a collection of sentences over a period of time and to differentiate between grammatical word sequences and randomised word sequences.

The network was trained using the standard backpropagation algorithm with momentum (Rumelhart, Hinton & Williams (1986) and was a typical Elman recurrent network architecture (Elman 1990) containing a context layer (the temporary storage layer in Figure 3.1). Elman used a recurrent link, where a copy of the hidden layer activations from the previous timestep was passed to the context layer. This had the effect of providing a dynamic memory where internal representations were created reflecting task demands in the context of prior internal states.

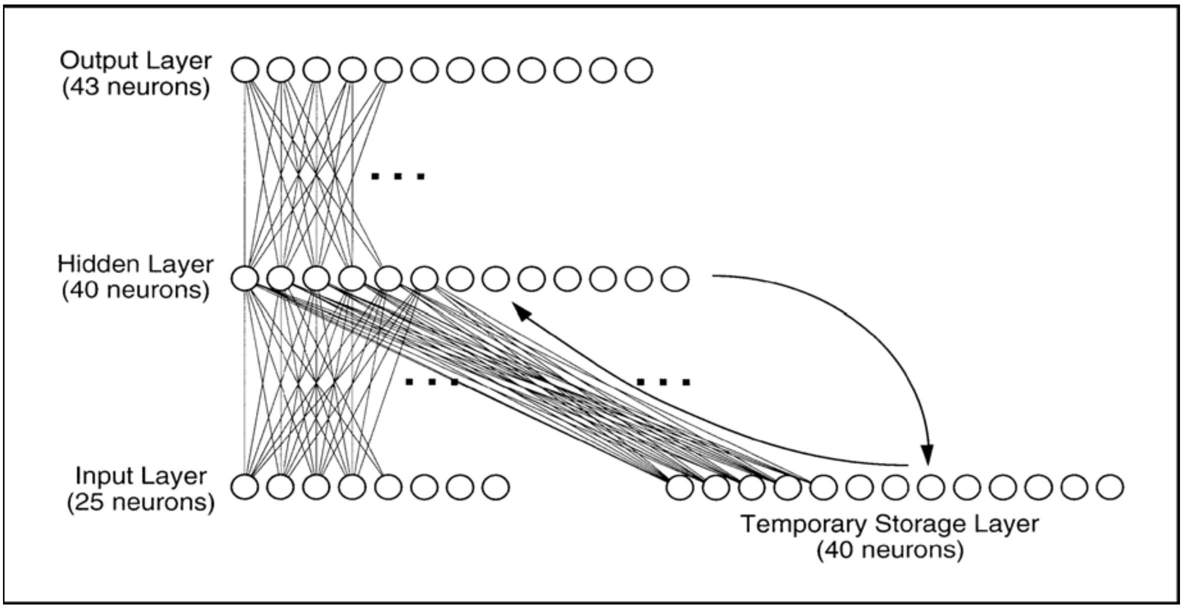


Figure 3.1. Architecture of speech perception network. Taken from Hoffman & McGlashan (1997)

A. Training on grammatical word sequences

Training Inputs - Phonetic

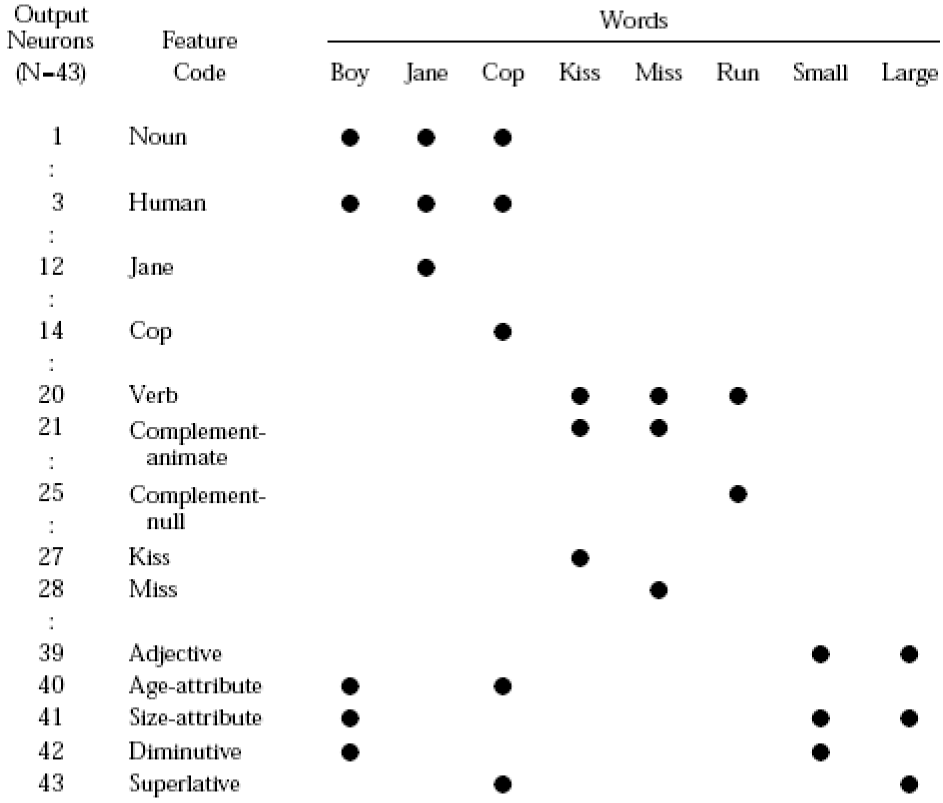
I used exactly the same training set as Hoffman, where two hundred and fifty six grammatically correct sentences of three to six words were constructed from a vocabulary of 30 words (14 nouns, 11 verbs, 4 adjectives, plus the negative - ‘won’t). An example of typical sentences used include: *large boy tell Jane story; cop give Sam warning; bill love girl*. Each word was given a unique phonetic representation consisting of a random 25 bit binary code, where approximately 25% of the inputs were turned on (set to plus one). This random allocation of the input vector simulates the random correspondence between phonetic input and semantic output, where similar sounding words do not have a similar meaning.

Although it was not clear from the journal paper, it was evident from Hoffman’s program that in the original model the sentences in the training set were separated by a *nil* input, where all 25 input neurons were turned off (set to zero), in order to signify the end of a sentence. I was using PDP++ which requires input sequencing in groups and has no need for further separation of the input sentences. However, I found that it was still necessary to include the *nil* input at the end of each sentence in the training set; otherwise the errors on the *nil* inputs presented in the test set were very high, as the network had not yet encountered a *nil* input. Therefore, in my simulations the sentences were presented to the network as 256 groups, each containing 3-6 words from the word set plus the additional *nil* input, in order to replicate Hoffman exactly.

Training Outputs – semantic and syntactic

Each of the 43 neurons in the output layer represented a feature in exactly the same way as Hoffman. Each word input was allocated three to six of these features, providing semantic and syntactic information (Figure 3.2). By training the network to detect feature codes of the sentence words a word can be classified as detected or misidentified when testing with novel data. For example, when *Jane* is input, represented by a unique 25 bit binary code, the corresponding semantic output will contain three of the 43 neurons in the output layer turned on: *noun, human* and *Jane*. The remaining 40 neurons will be turned off.

It is important to note that it is intended for the network to learn the simple association between a word and its output features (syntax and semantic meaning). It is not intended that the network should learn to predict the next word, as in Elman (1990).



Output Neurons (N-43)	Feature Code	Words							
		Boy	Jane	Cop	Kiss	Miss	Run	Small	Large
1	Noun	●	●	●					
:									
3	Human	●	●	●					
:									
12	Jane		●						
:									
14	Cop			●					
:									
20	Verb				●	●	●		
21	Complement-animate				●	●			
:									
25	Complement-null						●		
:									
27	Kiss				●				
28	Miss					●			
:									
39	Adjective							●	●
40	Age-attribute	●		●					
41	Size-attribute	●						●	●
42	Diminutive	●						●	
43	Superlative			●					●

Figure 3.2 giving examples of output neural codes. Taken from Hoffman & McGlashan (1997).

Training

Hoffman designed a special program with an 'online' variant of backpropagation learning, where training consisted of 60 repetitions of a set of 256 different grammatical sentences, separated by one *nil* (equivalent to 1200 input vectors). With no clear indications from the paper as to the values for learning rate and momentum, I ran a number of pilot studies with starting weights at various initialisation points to find an optimal combination of learning rate and momentum which would reduce the network error to a minimum. With an optimal combination of a learning rate of 0.18 and momentum of 0.81 the network was trained ten times, using ten different initialisations for starting weights, on a total sequence of 1200 vectors (words and *nils*), each represented by 25 bits. The optimal initialisation point trained to a minimal sum squared error of 0.94 for the entire training set, over 60 epochs, at which point no further reduction in error was seen. All results shown are for this optimal initialisation point.

Testing

In exactly the same manner as Hoffman, four test sets were constructed containing 23 novel sentences using the same vocabulary and presented in 23 groups. Each group contained a sentence of 3-6 words, from the word set, followed by a pause consisting of 5 *nil* inputs, giving a total of 5250 inputs (210 words and *nils* x 25 bit binary code), detailed below:

(i) **Grammatically correct and undegraded:** Grammatically correct sentences containing words with the same inputs and targets as those words learnt in the training set (undegraded).

(ii) **Grammatically correct and degraded:** In order to test the network on phonetically degraded information, the same test set as in (i) above was used, but two of the inputs for each word that were previously on, i.e., set to 1, were turned off, by setting them to 0 (degraded).

(iii) R**andomly constructed (grammatically incorrect) and undegraded:** In order to demonstrate that the network had taken account of the order of the words within the sentences in the training set, and was learning syntax, the same words in the test set were presented in undegraded form, but in random order within their individual groups.

(iv) **Randomly constructed (grammatically incorrect) degraded:** A test set with a combination of random words and phonetically degraded information.

The total sum squared error across all 43 output neurons for each test set were compared to determine the effect on the network of both phonetically degraded information and randomly constructed test sentences.

B. Training on randomised (grammatically incorrect) word sequences

This was an important stage in the simulation as Hoffman used it to demonstrate dependence of the network on word order and that it was acting effectively as a working memory. To see the effect of training on grammatical word sequences versus randomised word (grammatically incorrect) sequences, the network was also trained with 256 randomised word sequences and tested with the four test sets in A above. Reductions in word detection rate and increased word misidentifications were expected (Hoffman & McGlashan 1997).

The effect of training on grammatical word sequences (in A. above) and randomised (grammatically incorrect) word sequences (in B. above) was investigated by comparing the effects of the four test sets on each training set, as demonstrated in Table 3.1. If there is a difference between testing on grammatical versus randomised word sequences, and the network is able to detect more output features correctly associated with the words when trained on grammatical word sequences, then the network is acting as an effective working memory and neuroanatomic alterations can be applied.

Table 3.1 summarising training and test sets

TRAINING SETS **256 sentences** **5250 inputs**	**TEST SETS** **23 Novel Sentences** **215 inputs**			
Grammatically Correct Sentences	Grammatically Correct and Undegraded	Grammatically Correct and Degraded	Grammatically Incorrect and Undegraded	Grammatically Incorrect and Degraded
Grammatically Incorrect (Randomised) Sentences				

C. Word detection and misidentification rates

According to Hoffman & McGlashan (1997) an algorithm was used to decide which word was the best fit for a particular semantic output pattern. The best fit became the detected word and when there was no clear best fit the word was classed as not perceived. Hoffman clarified the position further in an e-mail and explained that their criterion for word detection was built into their program. In effect, each output produced a distance calculation for each word (that was adjusted for the number of output neurons that originally were coded as ‘on’ so as not to penalize words associated with a larger number of output neurons turned ‘on’). The algorithm calculated an adjusted mean square between the observed output and the target output for all output neurons that were turned on. It then ranked each word according to its distance from the actual output, and if one word outranked another by 0.12 then that word was scored as detected. If no word outranked any other by 0.12 or more then the network was said to produce a null output. A misidentification was when the network confused one word with another. The figure of 0.12 was derived empirically to produce a very high rate of word detection success without spurious word outputs produced by *nil* inputs once the network was well trained.

It was difficult to incorporate Hoffman’s criterion for word detection into my simulation as I was using a standard simulation package and he had written his own program from scratch. However, after various pilot studies to determine a criterion in line with that used by Hoffman, I decided on two criteria in order to give the network as great a chance as possible to produce a very high rate of word detection. I used a liberal criterion of a total sum squared error of 3.9 and below, and a stricter criterion of a total sum squared error of 0.9 and below (the error, being the squared difference between the desired output and the actual output for each output neuron, summed over all output neurons). For sentences trained with grammatical word sequences the model achieved 100% word detection success, according to both criteria, when tested with grammatically correct, undegraded sentences; while models trained with randomised word sequences achieved word detection successes of 95% and 92% with liberal and strict criteria respectively, using the same test set (see Figure 3.6 in Results Section).

I was able to look at word detection rates for each training set, but my simulations did not allow for misidentification rates at the present time. A hallucination would be recorded if output layer activations gave rise to words during the 5 nil pauses with no phonetic inputs.

RESULTS

A. Training on grammatically correct sentences

It can be seen from figure 3.3 that the network trained with grammatical word sequences gave rise to a minimal sum squared error of 0.18, being the difference between the desired output and the actual output seen over all output neurons when tested with grammatically correct and undegraded test sentences, but the error rose to 182.32 when tested with grammatically correct, but degraded sentences. A similar pattern was seen with randomly constructed test sentences as the network gave a minimal sum squared error of 9.54 with randomly constructed and undegraded test sentences, which rose to 187.77 when the randomly constructed test sentences were degraded. Degradation would appear to have an effect on the total error produced by the network as turning off (setting to zero) two of the input neurons increased the error on both grammatically correct and randomly constructed undegraded test sentences by factors of approximately 90 and 18 respectively.

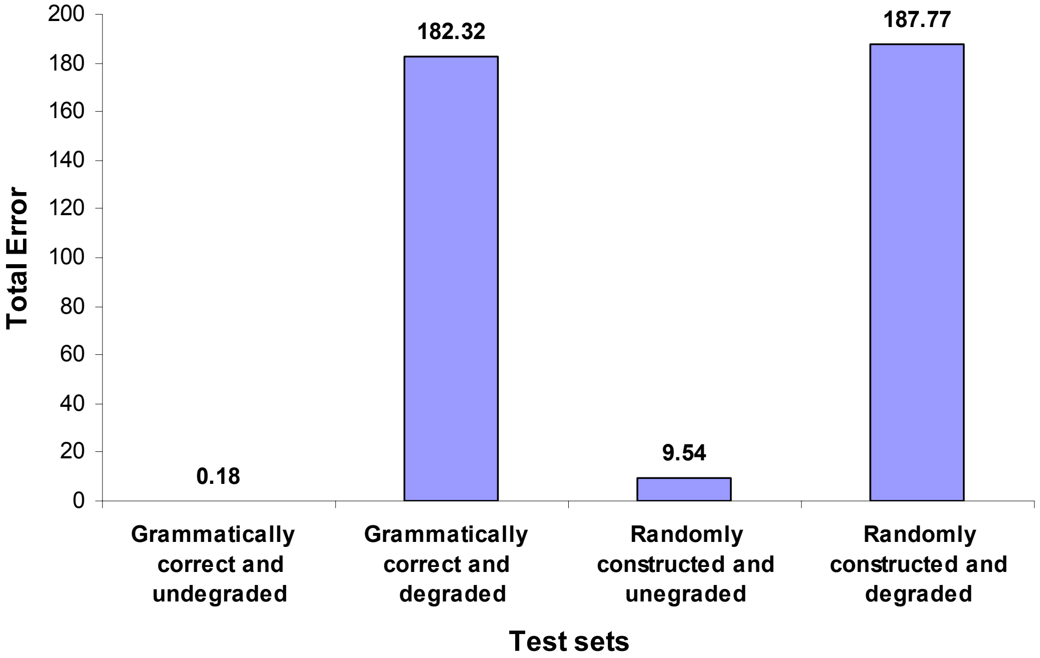


Figure 3.3 results from A. showing the effect of testing a network trained with grammatically correct sentences, with grammatically correct undegraded, grammatically correct degraded, randomly constructed undegraded and randomly constructed degraded test sentences

A small effect was seen also when testing with grammatically correct, undegraded sentences as opposed to randomly constructed, undegraded sentences (Figure 3.3) where sum squared errors are 0.18 and 9.54 respectively. An analysis of multiple networks was undertaken: while a paired samples t test showed that there was no significant difference between the two means of 0.0005 and 0.0440 respectively, t (214) = -1.59 (p > 0.05), there was a small to modest effect size, with a Partial Eta squared of 0.02. The increased error for the randomly constructed undegraded test sentences shows that the network was relying upon word order, to a limited extent, in the 256 grammatical word training sequences.

B. Training on grammatically incorrect (randomly constructed) sentences

Similar error patterns were recorded for the four test sets when randomised word sequences were trained (Figure 3.4, purple bars) to those when grammatical word sequences were trained in A above (Figure 3.4 blue bars, taken from Figure 3.3). Minimal sum squared errors were seen of 4.22 and 4.49 when tested with grammatically correct undegraded and randomly constructed undegraded test sentences respectively, rising to 171.56 and 184.09 when tested with grammatically correct degraded and randomly constructed degraded test sentences respectively. However, contrary to the findings in A. above, there was negligible difference between the errors from testing with grammatically undegraded and randomly constructed undegraded sentences, of 4.22 and 4.49 respectively. This reflects the lack of reliance of word order in the randomised word sequences training set.

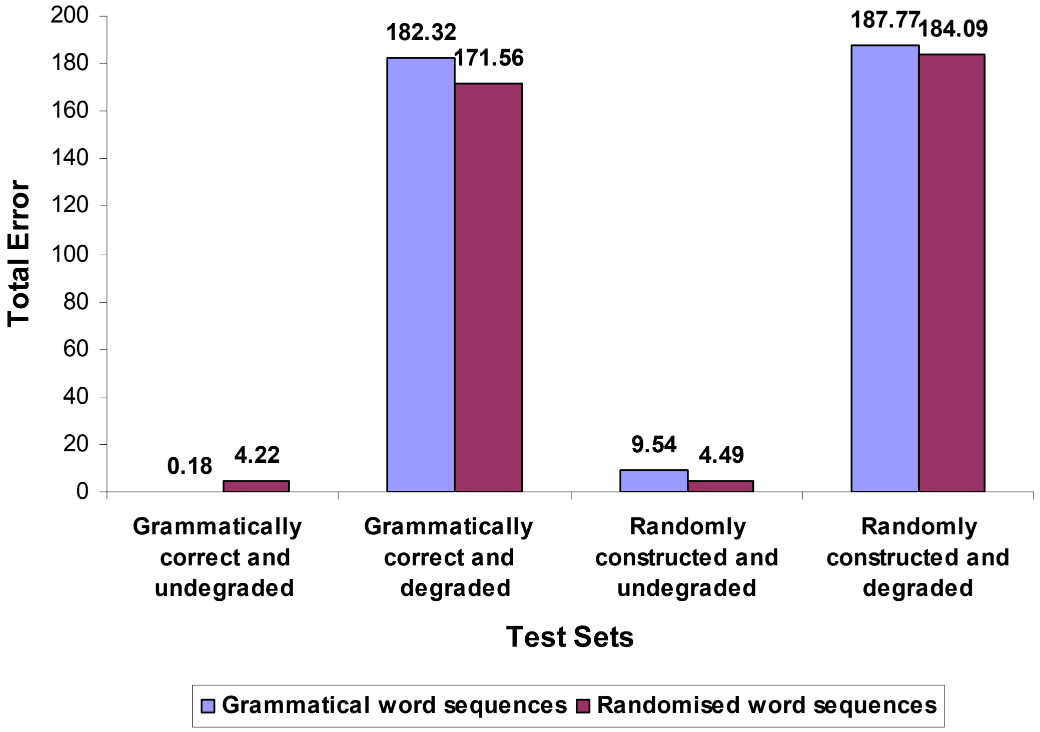


Figure 3.4 results from A and B showing the effect of testing networks, trained with both grammatical and randomised word sequences, with grammatically incorrect undegraded, grammatically correct degraded, randomly constructed undegraded and randomly constructed degraded test sentences.

C. Word detection and misidentification rates

Figures 3.3 and 3.4 above provide detail not included in the original paper. Hoffman and McGlashan recorded the percentage of words successfully detected by the network together with the percentage of misidentifications (Figure 3.5). I was able to look at word detection rates for each training set, but my simulations did not allow for misidentification rates at the present time.

My results for word detection can be seen in Figure 3.6. With both liberal and strict criteria, word detection rate for randomized word sequences was particularly high at 95% falling to only 93%, or 82% with no reductions due to degradation, respectively, in comparison to Hoffman's 70% falling to about 40%. In accordance with Hoffman & McGlashan's results, no hallucinations were recorded at this stage.

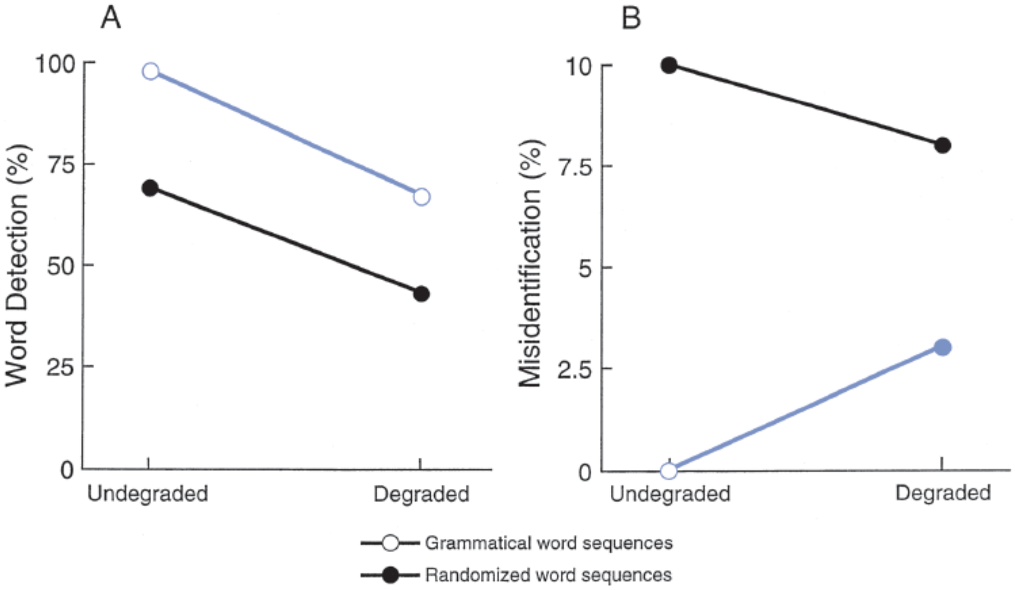


Figure 3.5 A. Word detection and B. misidentification rates. Effects of reducing phonetic information and randomization of input word sequences are represented. These data demonstrate that the network utilizes meaning intrinsic to grammatical sequences of words to facilitate translating phonetic inputs into percepts. Taken from Hoffman & McGlashan (1997).

Contrary to Hoffman and McGlashan, there appeared to be a good correspondence of output features to words with both grammatical and randomised word sequences. In this simulation there were no dramatic reductions in word detection rate as a result of using randomised word sequences. Therefore, unlike the Hoffman and McGlashan network, the present model did not appear to demonstrate dependence on word order in decoding input information and thus, I was not able to demonstrate the specific contribution of verbal working memory and grammatical word order shown by Hoffman and McGlashan.

While in Section A. above, there was a small effect seen for word order for grammatical word sequences, which was reflected in the total errors for the test sets, this was not reflected in word detection rates. Degradation of both grammatical and randomised word sequences appeared to be the only major variable to cause an effect. Without dramatic reductions in word detection rate or the corresponding increased word misidentifications seen as a result of using randomised word sequences, I was not in a position to proceed with neuroanatomic manipulations involving synaptic pruning and cell death that may obtain the purported hallucinations.

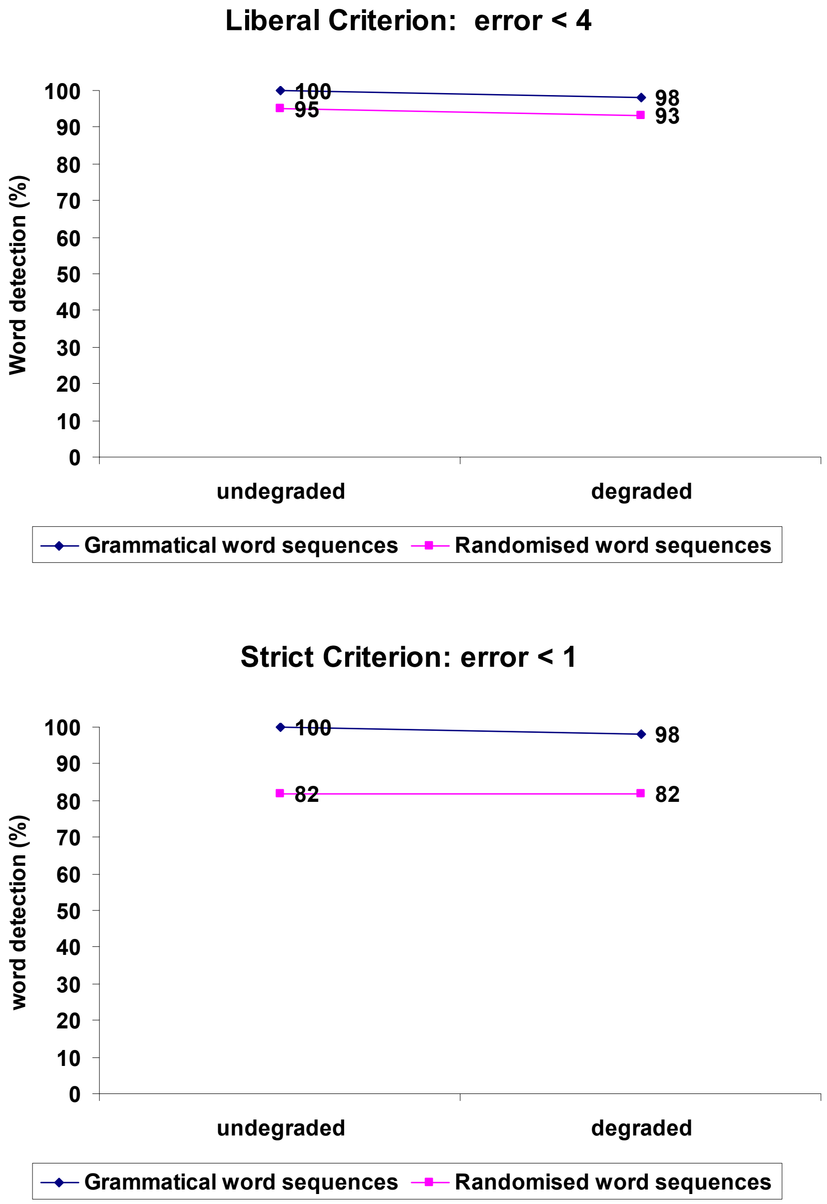


Figure 3.6 Word detection rates taken from my simulation, using both strict and liberal criteria, showing little effect of reducing phonetic information and randomisation of input word sequences.

Training on 5 *nil* inputs

As the network was trained with 1 *nil* but tested with 5 *nils*, I wanted to see if any findings were an artifact of the difference between the training and test sets. Two grammatical word sequences training sets, containing one and five *nil* inputs between sentences, both gave rise to the same minimal error of 0.18 (Figure -) when tested with grammatically correct, undegraded sentences. It was therefore concluded that the results did not appear to be due to an artifact of training with one *nil* input between each sentence and testing with five *nil* inputs between each sentence.

DISCUSSION

As there were no dramatic reductions in word detection rate or increased word misidentifications seen as a result of using random sentences, the present model did not appear to demonstrate the dependence on word order in decoding input information exhibited by the Hoffman and McGlashan model. While there was a small effect seen for word order for grammatical word sequences, which was reflected in the total errors for the test sets, this was not reflected in word detection rates. Degradation of both grammatical and randomised word sequences appeared to be the only major variable to cause an effect on word detection rate and I was, therefore, unable to proceed with neuroanatomic manipulations involving synaptic pruning and cell death that may obtain the purported hallucinations.

While my criterion for correct word detection was not exactly the same as Hoffman and McGlashan's and I did not look at misidentification rates, my results did provide a comparison between networks trained on both grammatically correct and randomly constructed sentences. However, early indications suggested that my simulations were not demonstrating the required dependence on word order in decoding input information. It is possible that using exactly the same criteria may have made a difference, but, in view of the considerable amount of work that was still needed to be done in order to improve the model, and advances in other directions that appeared to be more fruitful, I decided to abandon the model at this stage.

For a sizeable network of 148 neurons and 5920 connections the training set of 1200 items used by Hoffman was very small. Elman (1990) used 3500 connections and worked with a training set of nearly 30,000 items. It was nearly half the size and used twenty-five times as many items in the training set as Hoffman and McGlashan. It is generally acceptable that for every free variable (or connection) there should be at least one training item. A training set that is too small allows the network to find a solution to the problem in many ways and it is probably the case that Hoffman's original experiment had some flaws and was under-constrained. In effect it is possible that Hoffman had a non repeatable experiment which may explain why in our experiment behaviour is no different between syntactic and non-syntactic strings. No mention was made in the original paper of replication in order to seek a number of solutions to the problem. This practice was typical of early connectionist models, where they tended to be over resourced with too many free parameters, and I encounter a similar problem with the model of Hinton and Plaut (1987) which is described in Chapter 8.3. A statistical analysis would have strengthened their argument, providing robustness and validity to their findings. In addition, it would have been interesting to see the contribution of the dynamic memory in the recurrent link by removing the temporary storage layer. It may well have been the case that the network could learn the associations using a simple feedforward network architecture.

Regrettably the program Hoffman sent to me was complicated and not easy to replicate and there were still too many unanswered questions. Hoffman's training set contained one *nil* between each sentence, but the testing sets used five *nils*. This should have been mentioned in his discussion as a possible artefact of the model and once again, the issue of replication is very important. In addition, the purported hallucinations that emerged from the Hoffman and McGlashan model was actually only a single word hallucination ‘won’t’ that occurred during the pauses between sentences. While it is feasible that a larger network with a more complex vocabulary may generate hallucinations of other words and possibly sentences, it is also possible that this particular one-word hallucination was an anomaly of either the data set or the nature of the word and the function of the negative in the sentences.

In the Hoffman model speech hallucinations emerged with synaptic losses of around 80 to 95%, which they claimed corresponded to an additional 20% loss of synapses in the frontal cortex compared to normal adult levels. However, later evidence (Shenton 2001) suggests that there is no significant volume loss in frontal cortex in schizophrenia.

As I began to look at other research in this area, other questions began to arise that were not addressed by the Hoffman and McGlashan model. Their model focuses on connectivity in the frontal lobes, but subsequent connectivity models pointed to reduced parahippocampal connectivity in the temporal lobes as an explanation of schizophrenia-like episodic memory deficits (Talamini, Meeter, Elvevag, Murre & Goldberg 2005). Furthermore, Hoffman and McGlashan only modelled the end point of the neurodevelopmental process and not the formation and progression, which is thought to be pre-programmed and to begin in early life.

In addition, I found other models that focused more on the neural basis of schizophrenia. In particular, dopamine was seen to play an important role in both the symptoms and cognitive deficits associated with the disorder. Hoffman and McGlashan did not focus on dopamine aberration, for accepted reasons at the time, with possible dopamine dysfunction as an exacerbating factor, but not central to the issue. No mention was made of the considerable evidence of dysfunction of the dopamine system and associated areas, such as the basal ganglia. While Hoffman and McGlashan addressed one of the symptoms, hallucinations, it ignored other symptoms and cognitive deficits and therefore a degree of biological plausibility. Their mistake, and the community at that time, was to downgrade the importance of DA dysfunction.

To conclude: i) my model was a valid first attempt to implement an early connectionist model given the zeitgeist in 1997, i.e. the focus on connectivity in the dorsolateral prefrontal cortex; ii) I learnt a lot about how to replicate and critically evaluate an existing model. My critique of the limitations of this model, given the change of emphasis toward dopamine dysfunction, provides a plausible reason for a major switch in my modelling approach. In view of other models which came to light I decided not to pursue the neuropsychological approach any longer as it was difficult to instigate and would yield a limited set of results. If the modelling had been successful, further work could have been undertaken to strengthen the model in these areas to make it more biologically plausible. However, the above forms a valuable part of the argument for the exploration, and future choice, of a valid model that would permit the change of theoretical emphasis on the causes of schizophrenia. This 'change' should embrace the shift of emphasis from a brain dysfunction based on a neuropsychological profile, to one based upon the mechanism responsible for the positive symptoms and the dopamine hypothesis of schizophrenia.

3.2 Modelling and the Dopamine Hypothesis of Schizophrenia

Very early on in my research I noticed that much of the literature on the computational modelling of dopamine and schizophrenia referred to an early connectionist model by Cohen and Servan-Schreiber (1992). This seminal work used artificial neural networks to simulate normal and schizophrenic performance in three tasks that relied on the correct use of context: the Stroop task, the continuous performance test and a lexical disambiguation task. It was posited that damage to the dopamine system could account for the inability to use context appropriately. Dopamine was modelled as a change in the slope (or gain *G*) in the activation function of processing units and was shown to modulate the responsivity of neurons (corresponding to the prefrontal cortex) to external input, thus increasing the signal to noise ratio (Servan-Schreiber, Printz & Cohen 1990). Essentially, dopamine was believed to be crucial for optimising the signal-to-noise ratio thought to enhance working memory by reducing interference or noise. This theory related to the slow-acting (tonic), diffuse, non-specific effects of dopamine in the system.

However, since that time, the discovery of rapid (phasic), behaviour-specific bursts of dopamine (Schultz 1992; Schultz et al. 1993), discussed further in Chapter 4, led Cohen and colleagues to expand and update their existing ideas. They identified a transient gating role for dopamine where dopamine was able to modulate both afferent input and local inhibition in the prefrontal cortex (Braver 1997). This phasic gating system was very different to the earlier ideas of tonic changes in dopamine activity aiding the signal to noise ratio in working memory mentioned above, and posited a separate mechanism for salience. Instead, dopamine was able to control what was to be retained, by actively gating salient information into the prefrontal cortex. Salient information causes positive dopamergic activity, which opens the gate to update working memory. When there is no salient information, the gate remains closed, nothing is able to enter working memory and the representations currently active are maintained (Figure 3.7).

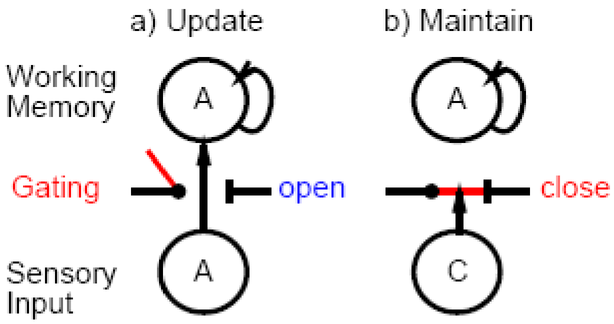


Figure 3.7 Illustration of active gating. When the gate is open, sensory input can rapidly update WM, but when it is closed, it cannot, thereby preventing other distracting information (C) from interfering with the maintenance of previously stored information. Taken from O'Reilly & Frank (2006).

Following recognition of the shorter, phasic effects of dopamine in the Reward Prediction Error Hypothesis (Chapter 4.1) and the subsequent discovery that Temporal Difference Learning was a good way of modelling this phenomenon (Chapter 4.2), a new Learning and Gating Theory of Dopamine was developed (Braver, Barch & Cohen 1999). This new theory conformed more closely to accumulating neurobiological evidence and also claimed to be a more powerful and complete theory of the mechanism of cognitive control.

The new theory updated previous ideas of dopamine as a neuromodulator and its role in active memory (Cohen & Servan-Schreiber 1992; Braver 1997), by combining it with the work of Montague, Dayan and Sejnowski (1996) on reward prediction error and learning. They hypothesised that schizophrenia resulted from increased noise in the dopamine system, leading to abnormal updating and maintenance of context information in working memory. Specifically, dopamine was seen as a unitary function which enabled an organism to predict and respond appropriately to events that led to reward.

It was believed that dopamine provided flexible access of task relevant information to active memory in the prefrontal cortex, but at the same time protected against interference from competing irrelevant information. The phasic dopamine prediction error mediated both learning and gating effects where learning was driven by prediction errors that affected synaptic strength and biased on-going processing, while the same responsivity effectively gated access to active memory, via the effects of excitatory afferent and local inhibitory input.

A connectionist model was produced in support of the new theory based on the AX continuous performance task (CPT) (Braver et al. 1999), which suggested that reduced phasic activity, i.e., reduced update to active memory, led to perservatory behaviour; while increased phasic activity, i.e., increased update, led to poor interference control, and therefore distractibility. Additionally, increased tonic (or longer-term background) activity led to delay related decay of active memory, and therefore maintenance deficits.

Both perseverations and distractibility are known disturbances to the prefrontal cortex and are typical symptoms of schizophrenia, along with poor maintenance control. Perseveratory behaviour occurs when a patient becomes preoccupied with a task and is unable to change strategy or appropriately update goal representations, while distractibility is the inability to concentrate or focus on the task at hand. This model posited that both perseverations and distractibility were due to impairments in phasic dopaminergic activity which affect working memory. However, the model was of two very different systems in the brain doing different jobs and possibly coding for two different things; salience in the midbrain and working memory in the prefrontal cortex. It is important, therefore, to investigate how these behaviours relate to each other and it is this interaction that will be explored in the current research.

3.2.1 Simulation of the AX-Continuous Performance Task (CPT) (Braver, Barch & Cohen 1999)

Following my decision to abandon further attempts to simulate the Hoffman and McGlashan speech perception network (3.1.1), I was actively seeking a new model as a base for my future work. Any model incorporating the cognitive deficits of schizophrenia should be able to demonstrate the impairments in both the maintenance and updating of context seen in the AX-CPT. To understand the task in more detail, I implemented an extremely simple version of the original model using a supervised simple feed forward architecture implemented in T Learn (Plunkett & Elman, 1997).

Prior to the addition of the new gating mechanism Braver and colleagues were able to demonstrate both normal and defective schizophrenic performance in the AX-CPT using a simple feed forward connectionist network. The gating mechanism was added later to see whether it could also capture both performances and provide a more refined model of dopamine activity. I did not add the gating mechanism at this stage as I did not intend to use this particular mechanism in my model.

The AX-CPT provides a measure of cognitive control function, involving both the maintenance and updating of context (Chapman & Chapman 1978; Braver et al. 1999). Specifically, during the sequential visual presentation of single letters, it is necessary to identify a target letter, but only when it is preceded by a specific cue. For example, if the cue is an A and the target is an X, it is necessary to signal when an X is shown, but only when it is preceded by an A.

METHODS

The simple feed forward network architecture (Figure 3.8), containing a hidden layer was implemented in T Learn.

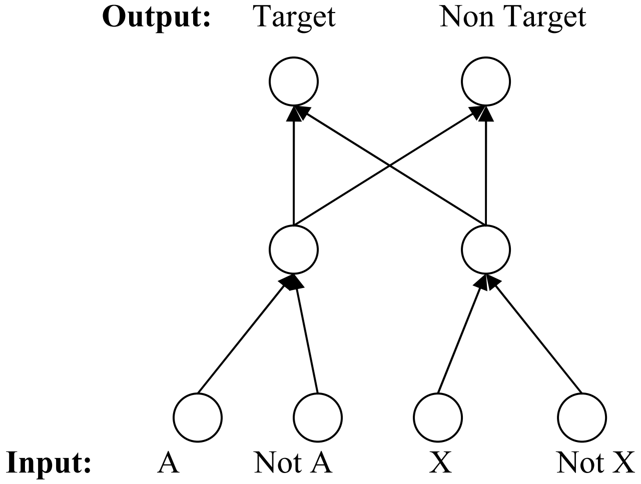


Figure 3.8 Simple feed forward architecture to demonstrate the AX-CPT

Training

Letters were presented to the network in bipolar form, where the presence of an A or an X was represented by plus one, and their absence by minus one. The network was trained to output a value of plus one when an A was followed by an X, but to output a minus one when it was followed by a non X, or when no A was presented. During training, in accordance with Braver et al., cue A followed by target X was presented to the network on 70% of occasions, while the other permutations: A followed by non-X; non-A followed by X and non-A followed by non-X, were each presented 10% of the time (see Table 3.2 for the four training set representations).

The network was trained ten times at different initialisation points.

Table 3.2 Showing input and output representations. Positive inputs/outputs are in bold.

	Input				Output	
	A	non A	X	non X	Target	Non-Target
A-X (70%)	**1**	-1	**1**	-1	**1**	-1
A-non X (10%)	**1**	-1	-1	**1**	-1	**1**
non A-X (10%	-1	**1**	**1**	-1	-1	**1**
non A-non X (10%)	-1	**1**	-1	**1**	-1	**1**

Testing

In order to test the robustness of the network to the task, the fully trained network was tested with noisy data, where a value of 0.5 was substituted for all values of plus one, for the four representations in the training set. The noisy data represents a confound in the experiment where, for example, the X may be on screen for an insufficient length of time to be attended to, it may be blurred, or it may represent a deliberate distraction to the participant.

RESULTS

Training

Extensive training over 20,000,000 iterations was undertaken to produce a minimal sum squared error of less than 0.004 for maximum learning. The network was trained ten times, with ten different initialisations. An example of one of the actual outputs can be seen in Table 3.3.

Table 3.3 showing actual outputs and sum squared error per pattern following training compared to expected outputs, in brackets.

	Input				Actual Output (expected)	
	A	non-A	X	non-X	Target	Non-Target
A-X	**1**	-1	**1**	-1	**1.00 (1)**	-1.00 (-1)
non A-non X	-1	**1**	-1	**1**	-1.00 (-1)	**1.00 (1)**
non A-X	-1	**1**	**1**	-1	-0.99 (-1)	**0.99 (1)**
A-non X	**1**	-1	-1	**1**	-1.00 (-1)	**0.99 (1)**

Testing

The fully trained network responded correctly to the noisy test data by producing the correct output, to a total sum squared error of 3.64 across all input patterns (Table 3.4), and the model successfully simulated the AX-CPT.

Table 3.4 showing actual output and sum squared error per pattern when test input data given to trained network. Positive inputs/responses are in bold.

	Input				Actual Output (expected)	
	A	non-A	X	non-X	Target	Non-Target
Noisy A-X	**0.5**	-1	**0.5**	-1	**0.98 (1)**	-0.98 (-1)
Noisy Non A-non X	-1	**0.5**	-1	**0.5**	-1.00 (-1)	**1.00 (1)**
Noisy non A-X	-1	**0.5**	**0.5**	-1	-0.99 (-1)	**0.99 (1)**
noisy A-non X	**0.5**	-1	-1	**0.5**	-1.00 (-1)	**1.00 (1)**

DISCUSSION

A simple feed forward network was able to learn the CPT. The four test cases show that the model responds correctly, in a categorical manner, to all the input tests even when the input is noisy. This model could have been developed further to include the gating system and the dopamine reward prediction error (Braver, Barch & Cohen 1999), and been subjected to lesioning techniques, but as with Hoffman & McGlashan (1997), the model appeared to have major shortcomings for use in schizophrenia research.

In particular, the model focussed on the direct dopamine pathway from the ventral tegmental area, which delivers a homogeneous signal to prefrontal cortex, and does not include the basal ganglia and the cortico-basal ganglia circuit loops (see Chapter 4.3). The indirect pathway, via the basal ganglia, allows for hierarchical updating of the prefrontal cortex, where differentiated inputs are received by prefrontal sub regions. This indirect pathway is better equipped to address the important fundamental issue of selective updating, where higher order goals are actively maintained, while updating lower order sub-goals (Cohen, Braver and Brown 2002).

In addition, the model focused on the dysfunctional processing of context in schizophrenia using the AX-CPT. This task is a device often used as a measure of executive dysfunction, with emphasis on both the inhibition and correct use of context. However, there are many instances of damaged contextual processing in patients with amnesia, executive dysfunction and frontal lobe damage that do not suffer from psychotic episodes. While the model made a valuable contribution to the understanding of the neural systems underlying schizophrenia at that time, as the task used does not apply uniquely to schizophrenia, I question the validity of such models as an explanation of schizophrenia.

While I chose not to pursue this particular model and task, the literature referred to the work of Montague et al. (1996) and their theoretical framework for mesencephalic dopamine systems using temporal differences (Sutton 1988). In addition, a later review on computational perspectives on dopamine function in prefrontal cortex by Cohen et al. (2002) pointed to other research by Houk, Adams and Barto (1995) and Schultz, Dayan and Montague (1997) involving phasic dopamine as a reinforcement learning signal and the basal ganglia. This opened up a whole new avenue of research of dopamine as reward prediction error and is the foundation on which the remainder of my thesis rests.

3.3 Chapter Conclusions

The principal objective of this chapter was to implement two important, fundamental models of aspects of schizophrenia; the speech perception network of Hoffman and McGlashan (1997) and the learning and gating model of Braver et al. (1999). I have attempted to describe and explore both models by implementing some aspects of them in two different neural network packages, PDP++ and T Learn. While I do not intend to progress further with these models, the modelling already undertaken has made a valuable contribution to the direction of this body of research. Subsequent research and the flaws of the models have pointed to a new line of reasoning to account for the symptoms and cognitive deficits of schizophrenia which include dopamine dysfunction and the neuroscience of the cortico-basal ganglia circuit loops. Dysfunction of the dopamine system should be a ‘cornerstone’ of any neuroscientific model of schizophrenia and this is the focus of Chapter 4.

Chapter 4

Dopamine, the Basal Ganglia and Temporal Difference: The Prediction Error Hypothesis

Preliminary models described in Chapter 3 provided the direction for future modelling work and the remainder of this thesis. Braver et al. (1999), and a subsequent review on the computational perspectives of dopamine function in prefrontal cortex by Cohen, Braver and Brown (2002), pointed to other modelling work regarding the role of dopamine in learning, by Houk, Adams and Barto (1995), Montague, Dayan and Sejnowski (1996) and Schultz, Dayan and Montague (1997).

In the late 1980's and early 1990's Wolfram Schultz and colleagues conducted a series of experiments where they recorded the activity of single midbrain dopamine neurons in alert monkeys while they performed behavioural acts, such as reaching for food or pressing a lever, for a juice reward (Hollerman & Schultz 1998; Ljunberg, Apicella & Schultz 1992; Romo & Schultz 1990; Schultz 1986; Schultz, Apicella & Ljunberg 1993; Schultz & Romo 1990). The major finding that the fluctuating outputs of dopamine neurons signals changes in reward prediction errors had an enormous impact on the direction of research at that time. Houk, Davis & Beiser (1995) gathered together multidisciplinary contributions from researchers, including Schultz, linking theoretical studies with experimental approaches at various different structural levels, with the aim of modelling and understanding the nervous system. The basal ganglia networks, including dopamine and their linkages with the cerebral cortex, played a central role in reward processing; and the book culminated in a model by Houk, Adams and Barto (1995) of how dopamine neurons in the basal ganglia predict reinforcement, and how outputs from those neurons could be used to reinforce behaviours leading to reward, using an actor-critic architecture (Barto 1995).

Shortly after, Montague et al. (1996) and Schultz et al. (1997) suggested that the activity of dopamine neurons could be represented by Temporal Difference (TD) (Sutton 1988; Sutton & Barto 1998) errors in the predictions of future reward. The correspondence between dopamine reward prediction error and TD error was striking, and they both developed theoretical frameworks to explain the physiological and behavioural data by Schultz and colleagues. Montague et al. explained how fluctuations in the firing of dopamine neurons above and below baseline could deliver reward prediction errors to both cortical and subcortical targets, (1) during learning, and (2) during ongoing behavioural choice. Furthermore they used TD to model those errors and were able to make testable predictions about human choice behaviour. Schultz et al. also went on to explain the functional role of the dopamine signal through the TD algorithm and developed a novel way of representing a stimulus through time, where a representation of each sensory cue had more than one associated adaptable weight.

The important messages from this early modelling work were: (i) the idea of phasic dopamine acting as a reinforcement learning signal, or reward prediction error, indicating the difference between actual and expected reward, with a view to reducing subsequent prediction errors; and (ii) the dopamine reward prediction error could be modelled effectively using TD; specifically using an actor-critic architecture, which can be related to basal ganglia circuits and dopamine neurons. These two ideas are discussed in Sections 4.1 and 4.2 respectively.

In addition to reward processing, the basal ganglia have an important role to play in the contextual analysis of the environment by receiving input from diverse areas of the cerebral cortex, generating the reinforcement signal and relaying this information for use in planning and behaviour (Houk, Davis & Beiser 1995). The functional suitability of the basal ganglia for this role is discussed in Section 4.3.

It has been suggested that the dopamine system mediates the incentive salience of rewards (Section 4.4), modulating their motivational value, which is dissociable from hedonia and reward learning (Berridge & Robinson 1998). In order to relate the dopamine prediction error hypothesis to schizophrenia, a framework is described in Section 4.4.1 that describes psychosis in terms of *aberrant* saliences (Kapur 2003). Ideally, the ideas of incentive salience should be incorporated into future neurocomputational models of the role of dopamine in psychosis and the positive symptoms of schizophrenia.

Finally, I include evidence from functional magnetic resonance imaging studies that the reward prediction error model of dopamine activity applies to human reward learning and not just non-human primates (Section 4.5). This is a theoretical chapter and it is the work detailed in the next five sections that has inspired the modelling in the remainder of this thesis.

4.1 Dopamine as a Reward prediction Error Signal

Current theories of the effects of dopamine on behaviour focus on the role of dopamine as a neuromodulator, where organisms learn to organise their behaviour under the influence of goals, and expected future reward is believed to drive action selection, as seen during conditioning (Montague et al. 1996; Schultz et al. 1997). Specifically, it is dopamine that signals a reward prediction error of the difference between the current and expected future reward.

In this section I give a basic account of conditioning (Section 4.1.1), which puts into context the work of Wolfram Schultz and colleagues, who demonstrated the role of dopamine as a reward prediction error (Section 4.1.2).

4.1.1 Conditioning

Classical and instrumental conditioning are examples of Associative Learning, i.e., making a new association or connection between two events. The classic Pavlovian conditioning paradigm (Pavlov 1927) looked at dogs salivating, the unconditioned response (UR), when provided with food, the unconditioned stimulus (US). The dogs were trained to associate light, the conditioned stimulus (CS), with food, resulting in salivation, the conditioned response (CR), at the earlier stage when the light came on. Here, the CS (the light) takes on the specific motivational properties of the US and elicits the response previously given to the US.

Before Conditioning: CS (light) → No response
US (food) → UR (salivation)
After Conditioning: CS (light) → CR learned (salivation)

Should a new stimulus occur that predicts the food reward earlier than the light, for example, the sound of a bell, the animal will learn the new association. The bell will become the new CS and the light will no longer elicit the CR. This is known as secondary conditioning. Accordingly, the earliest stimulus that confidently predicts the reward becomes the CS.

In classical conditioning the animal is passive, but in instrumental conditioning the animal is active; it learns that a response it makes leads to a particular consequence, for example, pressing a bar leads to access to food. Wolfram Schultz and colleagues used classical and instrumental conditioning to demonstrate that dopamine provides a reward prediction error of the difference between current and expected future reward (Section 4.1.2).

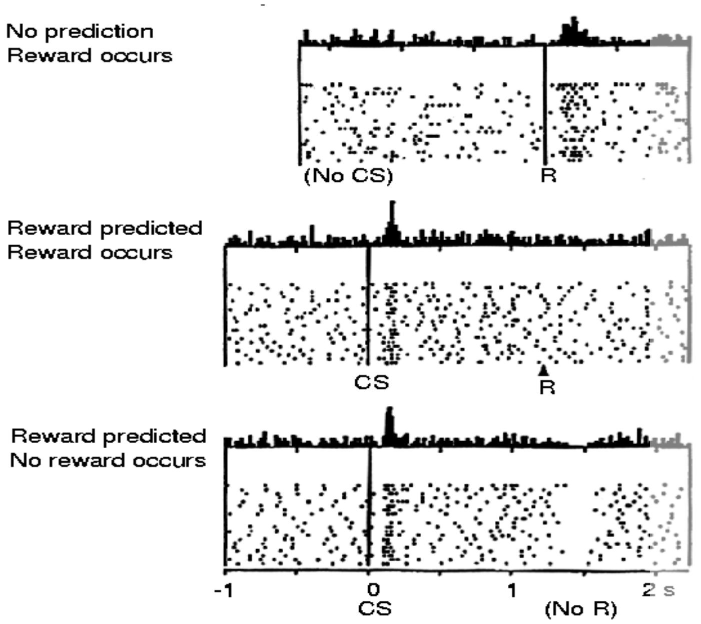
A reinforcer is any event that increases the probability of a response. A positive reinforcer, or reward, is a stimulus that, when presented following a response, increases the probability of the response. An example of positive reinforcement is detailed in Chapter 6, where a rat learns to traverse a maze to receive a juice reward. In Chapter 7 I describe an example of negative reinforcement, where a rat learns to associate a neutral CS, such as a light or tone, with an aversive US or outcome, such as an electric foot-shock, that has preceded the CS. This produces a conditioned avoidance response, thus avoiding the US. A negative reinforcer is a stimulus that, when removed following a response, increases the probability of the response. If conditioned behaviour is not reinforced, the CR gradually diminishes and extinction occurs, where the CS will eventually cease to be a reliable predictor of reward and will no longer take on the specific motivational properties of the US.

4.1.2 Single Cell Recordings of Dopamine Neurons

Physiological recordings of dopamine neurons on alert monkeys, by Wolfram Schultz and colleagues (Hollerman & Schultz 1998; Ljunberg, Apicella & Schultz 1992; Romo & Schultz 1990; Schultz 1986; Schultz, Apicella & Ljunberg 1993; Schultz & Romo 1990) have shown that information about rewarding stimuli is encoded in dopaminergic activity. Figure 4.1, is taken from Schultz et al. (1997) and represents the firing pattern of a single dopamine neuron over time (*x*-axis) and across different trials (*y*-axis). Firing of a single neuron is recorded as a single dot in the raster of impulses and the sum of these firings across trials is recorded in the histogram above. When an unexpected reward is given, such as a drop of juice, dopamine neurons respond with a short phasic response shortly after receipt of the reward. This effect can be clearly identified in Figure 4.1A as an increased rate of firing, above baseline, reflected in the increased density of impulses in the raster and in the height of the associated histogram. Phasic responses occur in the majority of dopamine neurons (55-80%) in a homogeneous fashion and do not discriminate between different types of rewarding stimuli.

When a reward is fully predicted by a CS, for example when a monkey is trained over the course of several days to touch a lever following a burst of light, there is no phasic dopamine activity following receipt of the reward. Instead the timing of the phasic activation changes to just after presentation of the CS (Figure 4.1B). This reflects the transfer seen during conditioning, where an animal's appetitive behavioural reaction transfers from an unconditioned stimulus (the drop of juice) to the CS (the light), as the associations develop over a period of learning.

Figure 4.1C shows the effect of non delivery of an expected reward. The CS elicits the same phasic response, but when no reward is received there is a depression of dopamine firing (below baseline) at the expected time of reward. This demonstrates that the dopamine neurons encode the timing of the expected reward.



- A. Unpredicted rewards result in a positive phasic reward prediction error. Discrepancy: things are better than expected.
- B. Fully predicted rewards result in no reward prediction error. No discrepancy: things are just as expected.
- C. When an expected reward fails to arrive, there is a negative reward prediction error corresponding to inhibition of the neurons. Discrepancy: things are worse than expected.

Figure 4.1 Electrophysiological recordings from a single dopamine neuron in the brain of a monkey taken from Schultz et al (1997). Each dot represents a neuron firing, while the histograms at the top of each diagram represent cumulative values of spiking activity from a population of dopamine neurons. The *x* -axis of each diagram represents time in seconds relating to presentation of the conditioned stimulus (CS) or reward (R) and the *y*-axis relates to the number of trials; with early trials at the top and later trials at the bottom. Diagram A is a selection of trials at the beginning of training, Diagram B represents trials when learning has taken place. Diagram C is also after learning has taken place, but shows the effect on dopamine neuron firing when an expected reward is not given.

The dopamine neurons appear to be encoding a reward prediction error signal of the discrepancy between actual and expected future reward. Unpredicted rewards provide a discrepancy where things are better than expected, resulting in a positive reward prediction error following receipt of the reward (Figure 4.1A). Fully predicted rewards provide no discrepancy as things are just as expected, resulting in no reward prediction error at the time of the reward (Figure 4.1B). When an expected reward fails to arrive there is a discrepancy as things are worse than expected, resulting in a negative reward prediction error at the expected time of the reward (Figure 4.1C).

4.2 Dopamine and Temporal Difference Learning

Life is a chain of actions undertaken to satisfy our basic needs and desires to survive and reproduce (Doya 2007). The results of rewards and punishments given in animal studies (e.g., Section 4.1.2 and Chapters 5.2, 6 and 7.1) are used to explain these goal directed behaviours by linking actions to positive and negative rewards in reinforcement learning, where organisms learn to organise their behaviour under the influence of positive or negative reinforcers. Reinforcement learning maps situations to actions in order to maximise reward, and has two distinguishing characteristics: a trial-and-error search, and a delayed reward (Sutton & Barto 1998). It is a computational framework for an active agent to learn behaviours from a scalar reward signal and provides a coherent account of the basal ganglia (Doya 2007). Reinforcement learning theory models the many internal states between the stimulus and its associated response, and expresses how animals learn to achieve goals and rewards efficiently, using reinforcement signals that maximise total future reward (Montague, Hyman & Cohen 2004).

A detailed account of the various different reinforcement learning algorithms available can be found in Sutton and Barto (1998), but in this thesis I focus on Temporal Difference (TD) (Sutton 1988; Sutton & Barto 1998), a model of classical conditioning, which extends the Rescorla-Wagner model (Rescorla & Wagner 1972) to take account some of the fine temporal structure of conditioning. TD was originally used by engineering systems to optimise actions in complex environments, but following recognition of dopamine as a reward prediction error, it has been used also as an explicit method of modelling and quantifying this error (Montague et al. 1996; Schultz et al. 1997; Hollerman & Schultz 1998). In order to move from passive classical conditioning to the active control seen in instrumental learning it is necessary to incorporate actions into the basic TD algorithm. This can be achieved by adding a policy, a set of rules that dictate which actions are to be taken in each state. Alternatively, direct assessment of state-action pairs, or Q values, obviates the need for a policy (Watkins & Dayan 1992). Both methods are taught by a TD error, but in this thesis I focus on policy optimisation in an actor-critic architecture (Barto 1995), where the aim is to optimise actions in different states to maximise long-term reward. However, I do refer to Q learning briefly in Chapter 7 when I discuss the limitations of TD.

In this thesis I do not attempt a thorough investigation of existing actor-critic models but focus instead on a particular variation of a TD model by McClure, Daw and Montague (2003). This model was inspired by the decision model of Egelman and Montague (1998), and incorporated an actor-critic architecture capable of extending the TD algorithm to address action selection (Section 4.2.2). This model is an interesting exception to the typical actor-critic in that it is capable of addressing free-operant behaviour (Niv et al. 2007). In Chapter 5 of this thesis, I implement a version of this model to conduct an investigation of the effects of dopamine receptor antagonism on running speed in a maze, and in Chapter 6 I extend the model to perform an analysis of the relationship between TD learning and uncertainty coding in a computational model of dopaminergic signaling.

The actor-critic architecture is capable of using the TD learning algorithm to model dopamine neuron firing patterns (Section 4.1) in a manner that can be related to basal ganglia circuits (Section 4.3). The dopamine cells in the ventral tegmental area and the substantia nigra pars compacta are posited to report the same prediction error, but the former are believed to be associated with the critic, controlling the learning of values held in the amygdala and orbitofrontal cortex; and the latter with the actor, controlling the learning of actions in competitive cortico-striato-thalamo-cortical loops (Dayan & Balleine 2002). The TD error signal generated is used in two ways:

(i) The Critic (Section 4.2.1): As a prediction error or learning signal used to create better estimates of future reward.

(ii) The Actor (Section 4.2.2): To bias action selection towards situations that predict the best reward.

The actor-critic is a way of organising the architecture of an agent. The critic learns to anticipate reinforcing events by learning predictions of long term future reward and storing these estimates as values associated with a particular state. The actor adjusts behaviour to maximize the frequency and magnitude of reinforcing events by using a policy to specify action choices. The critic is able to solve the temporal credit assignment problem of determining the contribution of a particular action from general feedback of a full sequence of actions. It assumes some actions are more desirable than others and learns to provide useful immediate evaluative feedback as to which are to be reinforced and which are to be avoided, based on predictions of future reinforcement (Barto 1995). The corresponding biological problem is getting reinforcement signals to the correct synapses at the right time to effectively guide the learning process and to achieve goals by maximizing total future reward.

With regard to the neural underpinnings of the actor-critic architecture, Houk, Adams & Barto (1995) originally suggested a model where the critic and actor may be analogous to striasome and matrix modules in the striatum respectively. Later research by O'Doherty et al. (2004) involving functional magnetic resonance imaging pointed to a dissociation between ventral and dorsal striatum in instrumental conditioning, where dopaminergically controlled plasticity in the ventral striatum is associated with the critic, while the dorsal striatum is better associated with the actor (Section 4.5). Recent research using temporary lesions in rat striatum suggests more of an actor-director-critic architecture in the striatum, where the dorsal striatum only partially conforms to an acting role as it is involved in the performance of actions, but plays no part in learning them. Thus, the ventral striatum, which is responsible for both learning and performance of skills, acts more like a director than a trainer to the dorsal striatum, mediating the effects of the critic (inputs from the ventral tegmental area and substantia nigra) on the actor (Atallah, Lopez-Paniagua, Rudy & O'Reilly 2007).

4.2.1 The Critic: Dopamine as a Reward Prediction Error

The TD algorithm is designed to learn to predict future reward, which can be expressed as a value function *V**, representing expected total future reward, from any state, *s*, (Equation 4.1), where *t* represents a period of time and subsequent arbitrary discrete time steps $t = 1$, $t = 2$ etc. *E* is the expected value and *r* represents the value of the reward. The discounting parameter, *γ*, between 0 and 1, has the effect of reducing previous estimates of reward exponentially with time, so that a reward tomorrow is worth slightly less than the same reward today. Equation 2 is Equation 1 in a recursive form that can be used in the learning process.

$$V^*(s_t) = E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + ...\right] \qquad \text{[Eqn 4.1]}$$

$$V^*(s_t) = E\left[r_t + \gamma V^*(s_{t+1})\right] \qquad \text{[Eqn 4.2]}$$

Because *V** is unknown the TD algorithm proceeds by calculating an estimate *V* of *V**. If *V* is equal to *V**, then the system will have perfect information about the environment and the difference between the estimated value of reward, now and in the future, $V(s_t)$, (value of the current state) and the estimated value of reward in the future, $V(s_{t+1})$, (value of the next environmental state) will be equal to the intrinsic reward associated with the current state, r_t (Equation 4.3). However, this is unlikely in practice, as the environment is usually stochastic and *V* is only an estimate, and so it is usual for an error signal, *δ(t)*, to be generated. The TD prediction error, *δ(t)*, is a measure of the inconsistency for estimates of value at successive time steps and is derived by rearranging Equation 4.2 into Equation 4.4, which is a measure of the relationship between two successive states and the current reward. *δ(t)*, takes into account the current reward, plus the next prediction multiplied by the discounting parameter γ, minus the current prediction. The TD error *δ(t)* is used to create better estimates of future reward by nudging $V(s_t)$ towards a better estimate of the value function $V^*(s_t)$ (Equation 4.5).

$$\text{If } V = V^*\text{, then } V(s_t) - V(s_{t+1}) = r_t \qquad \text{[Eqn 4.3]}$$

$$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t) \qquad \text{[Eqn 4.4]}$$

$$V(s_i) \leftarrow V(s_i) + \alpha\delta(t) \qquad \text{[Eqn 4.5]}$$

For example, suppose an agent believes that they will receive a reward of £20 per annum, now and for the next four years. The estimate of reward now and in the future, $V(s_t)$, will be £100. If they receive an actual reward of £20 now, this is just as expected, $V(s_t)$ is equal to $V^*(s_t)$; they expect to receive a sum of £80 in the future, $V(s_{t+1})$, and there will be no error, *δ(t)*. As long as the £20 is received the following four years, things will be as expected and no TD errors will occur as in Table 4.1. If however, they only receive a sum of £15 in year 3, $V(s_t)$ is NOT equal to $V^*(s_t)$; things will be worse than expected and a TD error will be produced, *δ(t)*, of minus 5, which will be used to modify a proportion of $V(s_t)$, the estimate of expected future reward, now and in the future (Equation 4.5). This scenario is shown in Table 4.2, resulting in a reduction in the estimated future reward for year 3 of 17.5 (40 + (0.5 x -5)), where learning rate, α, is 0.5.

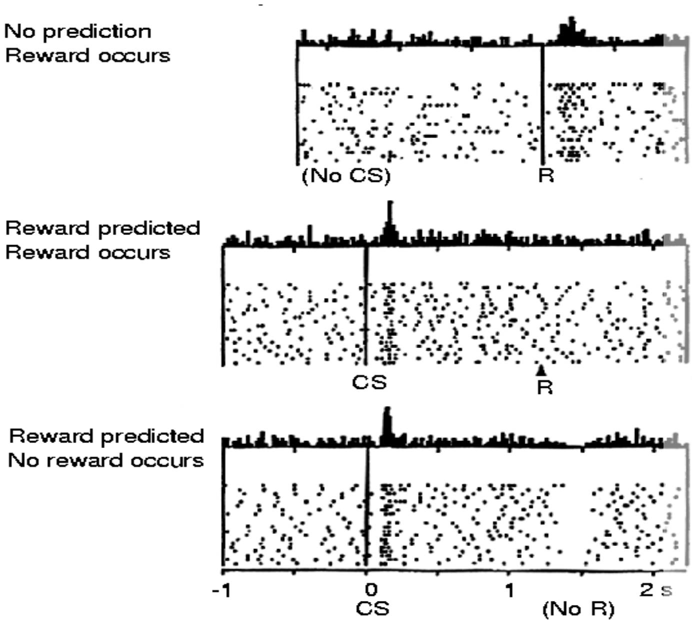
Table 4.1. When $V(s_t)$ is correct at all timesteps and equal to $V^*(s_t)$, there will be no phasic burst of surprise, $\delta(t)$.

Time	Year 0	Year 1	Year 2	Year 3	Year 4
$V(s_t)$	100	80	60	40	20
$V(s_{t+1})$	80	60	40	20	0
r_t	£20	£20	£20	£20	£20
$\delta(t)$	0	0	0	0	0

Table 4.2. $V(s_t)$ is NOT always correct. For example, in year 3 a sum of £15 was received instead of the expected £20, giving rise to a phasic burst of surprise, $\delta(t)$, of minus 5.

Time	Year 0	Year 1	Year 2	Year 3
$V(s_t)$	100	80	60	**37.50**
$V(s_{t+1})$	80	60	40	20
r_t	£20	£20	£20	**£15**
$\delta(t)$	0	0	0	**-5**

The correspondence between artificial models of learning using TD and electrophysiological studies of the role of dopamine in the ventral tegmental area is remarkable. Examples of TD errors can be seen in Figure 4.2, where values of $\delta(t)$ have been assigned to the circumstances in Figure 4.1 above. In Figure 4.2A, unpredicted rewards result in a positive TD error, where $\delta(t)$ is +1, as things are better than expected. In Figure 4.2B there is no TD error; $\delta(t)$ is 0, as fully predicted rewards result in no reward prediction error and no change in firing because things are just as expected. In Figure 4.2C, when expected reward fails to arrive, there is a negative TD error; $\delta(t)$ is -1**,** as things are worse than expected.



- A. **TD error, *δ*, is +1.** Unpredicted rewards result in a positive phasic reward prediction error. Discrepancy: things are better than expected.
- B. **TD error, *δ*, is 0.** Fully predicted rewards result in no reward prediction error. No discrepancy: things are just as expected.
- C. **TD error, *δ*, is -1.**When an expected reward fails to arrive, there is a negative reward prediction error corresponding to inhibition of the neurons. Discrepancy: things are worse than expected.

Figure 4.2 Temporal Difference can quantify the prediction error in Figure 4.1 above. In Diagram A where no reward is predicted, *δ(t)* is +1, in Diagram B where a reward is predicted by a CS, *δ(t)* is 0 and in Diagram C when an expected reward fails to arrive, *δ(t)* is -1.

4.2.2 The Actor: Dopamine to Bias Action Selection

McClure et al. (2003) designed an extension to the basic TD algorithm to include a role for dopamine in biasing action selection using the same prediction error signal, *δ(t)*, to teach a system to take the best actions to maximise the long term accumulated reward. In this policy the actor randomly chooses a possible action and the anticipated TD error, *δ(t)*, is calculated using Equation 4.4. The probability of taking this action is then calculated from this *δ(t)* value using a softmax function in Equation 4.6 (where *m* and *b* are parameters of the softmax curve, a sigmoid, Figure 4.3), which has the effect of squashing any input to the space between 0 and 1. This is a non-linear function, where the effect on the activation value of changing the net input varies at different points along the curve, in a similar manner to biological neural systems (Freeman 1979). Accordingly, changes in net input close to zero (at the centre of the curve) have a greater effect on output than changes in net input when it is very large. In this context the *x*-axis represents the TD error, *δ(t)*, input and the *y*-axis, the probability of action selection. When the TD error is high and positive there will be a greater probability of that action being selected than when the error is low and negative, due to the shape of the curve. Actions are generated with a probability of selection based on the predicted values of their successor states and there is a greater probability of remaining at the same state and not making a move when the error signal is low as all states become increasingly probable. If no action is selected, time is increased by one step and another random action is considered.

$$\textit{Probability (of taking action)} = \left(1 + e^{-m(\delta(t) - b)}\right)^{-1} \qquad \text{[Eqn 4.6]}$$

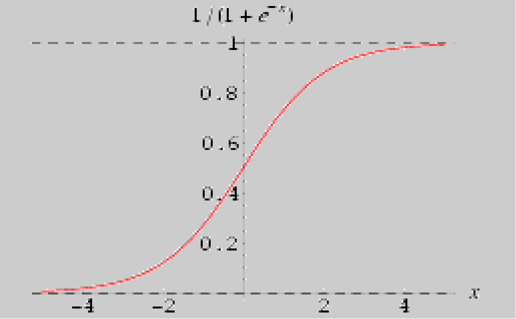


Figure 4.3 showing a sigmoid function. Here the *x*-axis represents the reward prediction error, *δ(t)*, and the *y*-axis, the probability of action selection. When delta is high there will be a greater probability of that action being selected than when delta is low.

Once an action has been selected, learning takes place according to Equation 4.5, where *α* is a learning rate parameter and is used to update the values of states with a proportion of the dopamine reward prediction error. It is noteworthy to mention that as our model was inspired by McClure et al. (2003), which incorporated the ideas of incentive salience (Section 4.4, Chapter 5.1 and Chapter 6), we maximise 'surprise' rather than reward, however, most other policies used in TD models maximise reward, for example Daw, Niv & Dayan (2005) (Chapter 7.4).

4.2.3 Explanation of TD Learning

Consider a one way maze of *n* states where only one intermediate state contains a reward and the value, $V(s_t)$, of all states is initialised to zero, i.e. pre learning. For simplification, the discount factor, *γ*, is 0 and learning rate, *α*, is 0.5. The three states of interest are the pre-reward state A, where no reward is received; the reward state B, where a reward with a value of 1 is received; and the post-reward or initialisation state C (Figure 4.4).

It is necessary to ensure that a virtual rat is always surprised when entering the maze, so that it has no expectation of future reward prior to being placed in the maze. In this way, the act of entering the maze becomes the earliest predictor of reward or the conditioned stimulus. The initialisation state contains no reward and is where $V(s_t)$ for that state is reset to zero every time the rat visits that state. This means whenever the virtual rat is moved into the pre-reward state it will always be surprised and a corresponding TD error will be produced.

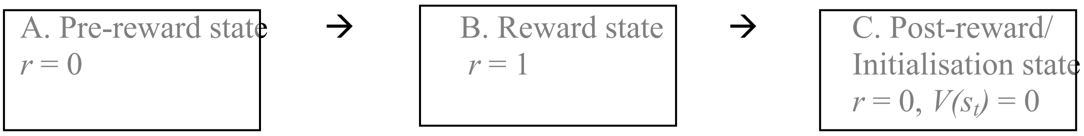


Figure 4.4 showing the three states, in a maze of *n* states that are of interest during TD learning

In this particular model the virtual rat is put in a maze in the pre-reward state where it has the ability to think ahead before selecting an action and making a move to the next state. (It is important to note that considerations of moves are an artefact of our model of action selection. A real rat operates in continuous time and will not plan moves in this manner). A comparison is made (Equation 4.4) between $V(s_t)$ of the current state and $V(s_{t+1})$ of the next state to determine any inconsistency, or TD error. As there is no reward and values of the current state, $V(s_t)$, and the next state,$V(s_{t+1})$, are zero, there is no inconsistency and the TD error is zero, $\delta(t) = 0 + 0 - 0$. This TD value is then used in the process of action selection (Equation 4.5); however, as the TD error is low the action is unlikely to be selected. If no action is selected, the virtual rat will remain in the same state and another move will be considered. An action is more likely to be selected when $\delta(t)$ is high, but even with a TD error of zero, an action will be selected eventually, by chance, and the rat will move to the next state. There may be a number of pre-reward states and the rat will travel to each of these, quite by chance; but until a reward is encountered the value of the reward state cannot be updated and compared with other states, and no TD errors will be recorded.

Eventually, by chance, the rat considers a move to the reward state containing a reward of 1. The value of that state, $V(s_t)$, is zero and, as the reward has not yet been encountered, there is no dopamine burst, $\delta(t) = 0$, $(0 + 0 – 0)$. The dopamine burst is only experienced when the rat is in the reward state, having received a reward, and is considering a move to post reward state, $\delta(t) = 1$, $(1 + 0 – 0)$. The value of the reward state is updated for the first time only when the probability of action selection has been accepted, an actual move is made away from the reward state, and a dopamine burst has been experienced, (Equation 4.6). With a learning rate of 0.5, the update to $V(s_t)$ is $0 + (0.5 \text{ x } 1) = 0.5$. On considering the move from the reward state to the post reward state, there will be no dopamine burst as no reward is received in the post reward state, $\delta(t) = 0 + 0 – 0 = 0$. It will be necessary for the rat to return to the pre reward state, and consider a move to the reward state for the second time, before any more dopamine bursts are experienced.

On the second occasion the reward is received in the reward state, the dopamine burst will be reduced from 1 to 0.5, $\delta(t) = 1 + 0 – 0.5 = 0.5$, as the value of that state was updated on the last run and a reward expected, so actual receipt of the reward is less surprising the second time around. On each successive run the dopamine burst will halve, as a result of the increasing value of that state. Eventually there will be no further TD error at that state and the value of that state will reach a maximum of 1 as the reward becomes fully predicted. The second run through the maze will both increase the value of the reward state and affect the value for the pre-reward state (or the closest of a number of pre-reward states) for the first time. Each subsequent run through the maze will affect earlier pre reward states until all states are affected by the reward received in the reward state. The values of all the pre-reward states will gradually increase to a maximum of 1 (if γ is 0), by which time full learning will have occurred, and the earliest pre-reward state effectively becomes the conditioned stimulus, whereby it is the only state that elicits a maximum dopamine burst of surprise, $\delta(t)$, of 1.

It is important to note that considerations of moves are an artefact of our model of action selection. A real rat operates in continuous time and will not plan moves in this manner. However, these considerations of moves permit the modelling of time and it is this product that is investigated in Chapter 6 where I implement a model by

McClure et al. (2003) and model the effect of dopamine receptor inhibition on the running speed of a rat in a maze.

4.2.3.1 An Example of the Learning Process

Consider a one way maze of eight states plus a post reward initialisation state, with a reward of 1 in state 7, where movement is permitted only in the direction of the arrows (Figure 4.5). Prior to learning the values of all states in the maze, *$V(s_t)$,* are initialised to zero, and the rat has no expectation of reward, in or out of the maze.

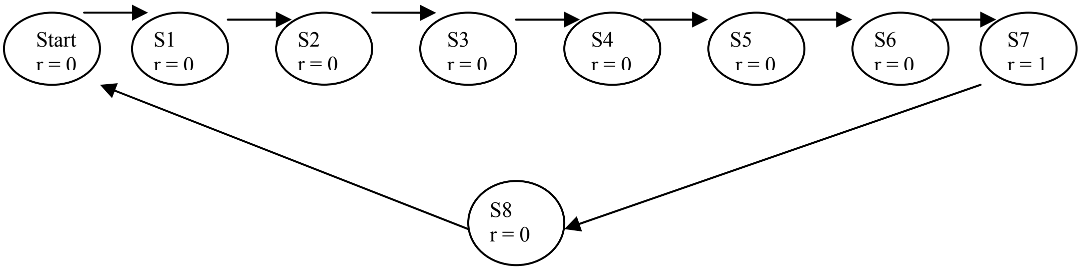


Figure 4.5 Maze with 8 states (S0-S7) plus initialisation state (S8). A reward of 1 is given in S7. All other states provide rewards of zero.

Complete learning takes place over the first 30 runs through the maze, during which time the values of each state converge to 1 (except for reset state S0), and maximum TD error transfers from the reward state to the start state, by which time the start state acts effectively as a conditioned stimulus. The effect of the transfer of the TD error can be seen in Figure 4.6, which shows the TD error, *$\delta(t)$,* for the first 30 runs for two of the state transitions: (A) reward state S7 to initialisation state S8; and (B) initialisation state S8 to start state S0 (the CS). On the first run a large prediction error, *$\delta(t)$*, was recorded for the transition from the reward state to the initialisation state only, and this error, or measure of surprise, reduced with each successive run, until there were no further TD errors for that state from run 9 onwards. The effect of the TD error, the inconsistency between states, was propagated back through all states through time (not shown) and the first effects were seen in the transition to the CS from the initialisation state, in run 10. The TD error in the CS reached a maximum of 1 by run 29, when full learning had taken place and the full effect of the reward was propagated back to the CS, and remained at that level as the only state showing a TD error, or surprise.

Appendix I contains raw data from the simulation for runs 1-3, 15-16 and 27-29 through the maze in Figure 4.5. The data is for actual moves taken and does not include considerations of moves not made, which are an artefact of the model and do not occur in reality. The information includes: (i) the state from which the move is made; (ii) the state moved to (iii) the TD error, *$\delta(t)$*; (iv) The values, *$V(s_t)$*, for each state in the maze; and (iii) a brief explanation of how progression through the maze occurs.

Although Appendix I does not show the considerations of moves not made, prior to learning traversing the maze will be slow. There is a greater probability of remaining in the same state and not making a move when the error signal is low, as all states become increasingly probable. As learning progresses the time taken to reach the reward state (modelled by the ability of the rat to reject a move and remain in the same state) decreases, and reaches a minimum when full learning has taken place. In this way the simulation is capable of modelling time, and advantage of this is taken in Chapter 5.1, where I model the effects of antipsychotic drugs and subsequent dopamine receptor antagonism on the time taken to progress through a similar maze.

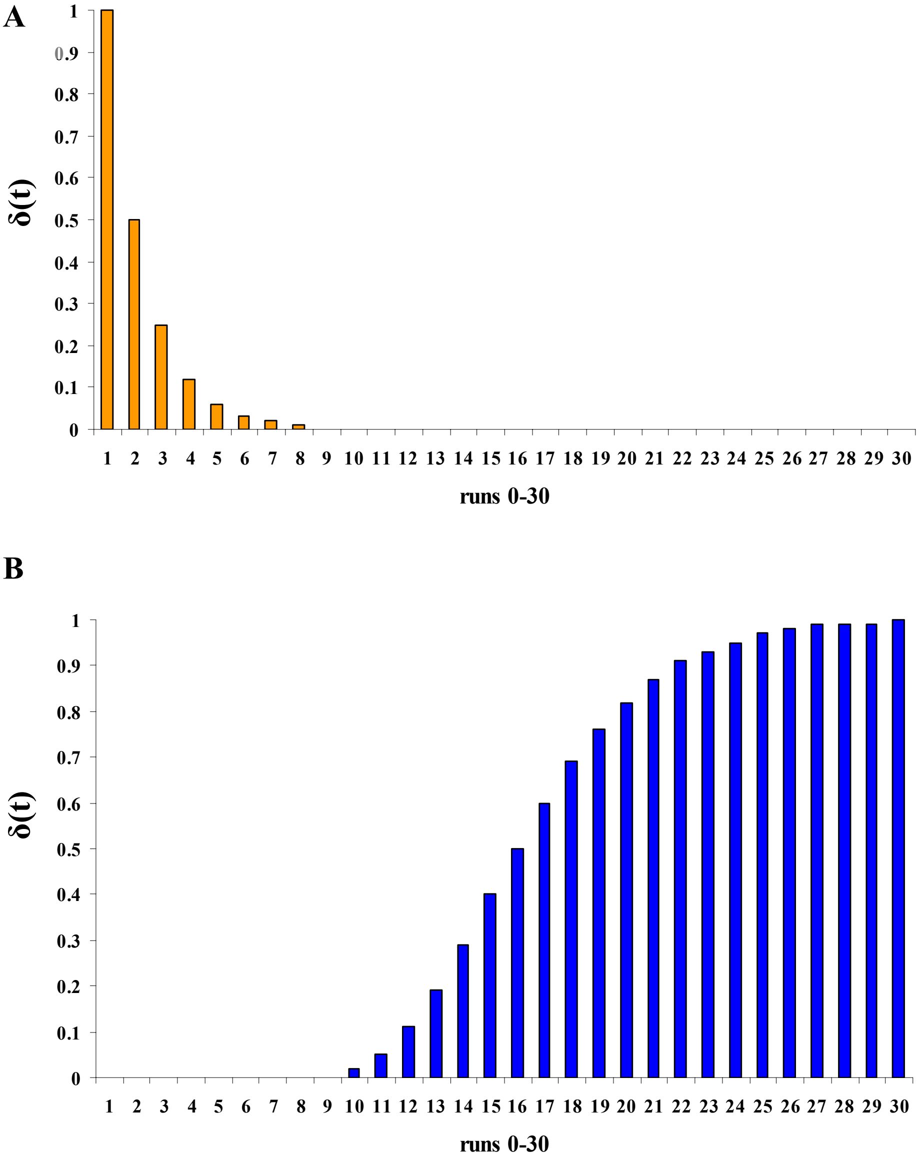


Figure 4.6 A demonstration of learning, where the TD error transfers from the reward delivery state (US – yellow bars) to the CS (blue bars), over time (30 runs through the maze). The TD error, *δ(t)*, is shown over these 30 runs for two state transitions: **A.** S7 (the reward state) to S8 (the initialisation state), beginning at 1 and reducing to zero by run 9; and **B.** S8 (the initialisation state) to (S0) the CS, which begins at zero and increases gradually to 1, from run 10 to run 30. By run 30 the value of all the states are learnt (except initialisation state) and the reward is fully predicted by the CS.

4.3 The Basal Ganglia

In addition to reward processing the basal ganglia have an important role to play in the contextual analysis of the environment by receiving input from diverse areas of the cerebral cortex, generating the dopamine reward prediction error signal and relaying this information for use in planning and behaviour. The basal ganglia form part of the functional cortical/subcortical circuit loops in the brain, also involving cortex, thalamus, and ventral tegmental areas (Alexander et al. 1986). They are a collection of interconnected nuclei comprising striatum (putamen, caudate nucleus and nucleus accumbens); globus pallidus (external and internal segments), subthalmic nucleus and substantia nigra (pars compacta, pars reticulate and pars lateralis), and are associated with motor control, cognition, emotions and learning. The basal ganglia are an area of huge convergence and summation of synaptic input, mainly from the cerebral cortex and thalamus (Figure 4.7), but also from the brainstem. The function of these nuclei involves the coordination of dynamically changing representations of sensory inputs, motor programmes and internal states; and the selection of appropriate behaviours for survival, as a result of those inputs (Yin & Knowlton 2006).

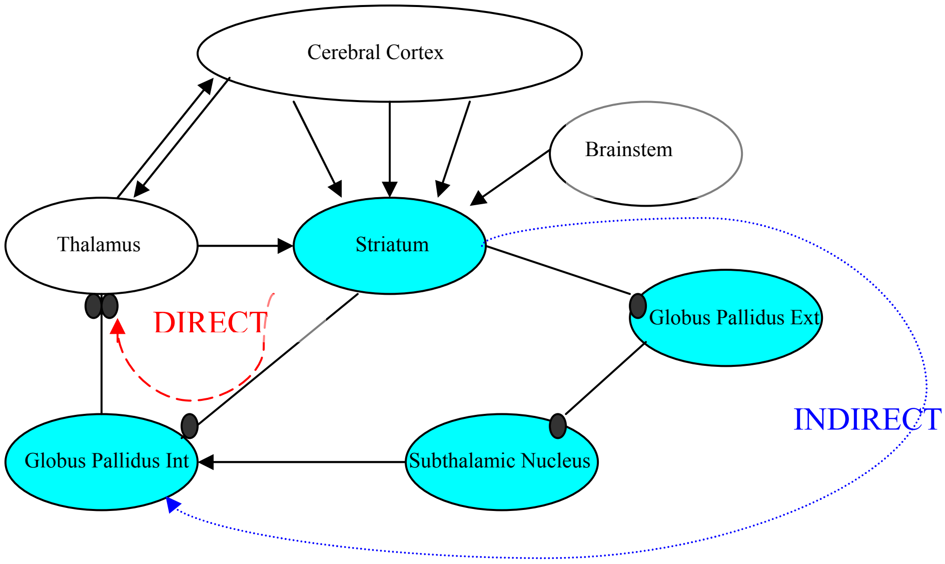


Figure 4.7 Direct (red) and indirect (blue) pathways in basal ganglia. Projections are excitatory (arrows) and inhibitory (circles). Basal ganglia nuclei in turquoise.

The dopamine system provides a reward prediction error signal of the discrepancy between actual and expected future reward (Section 4.1) whereby the spiny input neurons in the striatum recognise complex contextual patterns through the reinforcing influence of dopamine. They are particularly suited to this function because of the density of synaptic spines on the long dendrites in the spiny neurons in the striatum. There are 25-30 dendritic terminal branches and each spiny neuron receives input from about 10,000 different afferent fibres. The combination of the large number of terminals and the heavy afferent innervations enables the basal ganglia to function as a contextual processor (Wilson 1990). In addition to cortical synapses at the tips of the spine, spiny neurons also receive up to 5,000 dopaminergic synapses at the stems of the spines. Specifically, it is the timing of the dopamine phasic bursts that is important; only those inputs that coincide with the dopamine burst are strengthened, imbued with salience and cause striatal neurons to fire. Other non-salient input is inhibited by the strengthened, salient input through lateral inhibition (Schultz, Romo et al. 1995). In the short term dopamine exerts its effects immediately through neuronal activity, but in the longer-term through synaptic plasticity (Wickens & Kotter 1995).

In terms of design, single afferent fibres from each area of the cortex extend over much of the striatum, leading to a widespread, distributed cortical control. As a result, striatal neurons fire infrequently, and only when a particular combination of thalamocortical input, which is salient in terms of survival, is received (Wilson 1990). Evidence suggests that it may be the extensive dopamine projections from the ventral tegmental area and substantia nigra that meet thalamocortical inputs on the same dendrite that determine when these striatal cells fire (Wickens & Kotter 1995).

Output from the striatum is mainly inhibitory and is to other basal ganglia nuclei. The main direct output pathway (Figure 4.7, red line) is to the internal globus pallidus and the pars reticulate of the substantia nigra, which in turn project to areas outside the basal ganglia: the thalamus, and eventually back to the cerebral cortex. In addition, there is an indirect pathway (Figure 4.7, blue line), via the external pallidus to the subthalmic nucleus, before rejoining the main pathway and the internal pallidus (Houk 1995). Contrary to the rare firing of striatal neurons, neurons in the globus pallidus and substantia nigra fire tonically at high rates, which constantly inhibit neurons in the thalamus. When context is detected, the spiny neurons fire in the direct pathway, leading to a pause in the sustained inhibitory output from pallidal neurons and a brief burst of thalamic discharge. Effectively the inhibitory GABA projections from the striatum disinhibit the thalamus and thus responses are facilitated. Alternatively, the indirect pathway provides net excitatory positive feedback, resulting in an inhibition of responses (Houk, Adams & Barto 1995). These two pathways work together as a brake and accelerator system (Carlsson et al. 2001), which is also described by Frank (2005) as 'go' and 'no-go' pathways. Concurrent activity in both pathways thus controls competitive selection.

There are two main dopaminergic pathways in the human brain: the nigostriatal system, projecting from the substantia nigra to the striatum, and the mesocorticolimbic system, from the ventral tegmental area to the prefrontal cortex, hippocampus, amygdala and nucleus accumbens. The second pathway is the focus of this thesis and can be subdivided into the mesolimbic system, associated with reward and locomotion; and the mesocortical system, the modulator of cognitive function (Adell & Artigas 2004).

4.4 Incentive Salience

Evidence suggests that the dopamine system mediates the incentive salience of rewards, modulating their motivational value, which is dissociable from hedonia and reward learning. The Incentive Salience Hypothesis (Berridge & Robinson 1998) distinguishes conscious pleasures, 'liking,' from the unconscious core processes of reward, ‘wanting.’ Liking without wanting occurs as the result of extensive damage to brain dopamine systems, leaving individuals without motivation for any incentive, conditioned or unconditioned, while positive hedonic reactions to sweet tastes remain normal in animals that have lost nearly all of their mesolimbic dopamine neurons that project from midbrain to forebrain (See animal experiment by Ikemoto and Panksepp (1996), described in Chapter 5.1). It is also possible for wanting to occur without liking. Increased appetite (want) is usually accompanied by increased hedonic appreciation of food (like) but it was found that eating caused by electrical stimulation of the lateral hypothalamus was not accompanied by enhanced hedonic reactions. Wanting but not liking can also be triggered by microinjections of amphetamine (Wyvell & Berridge 2000) and microinjections of muscimol (Reynolds & Berridge 2000) that activate dopamine neurons in the nucleus accumbens. There is evidence to believe that the same dopamine-accumbens neural system may become sensitised or hyper-responsive to drugs and conditioned stimuli in the brain of drug addicts, who also demonstrate wanting without liking. This may cause heightened incentive salience to be attributed to drug cues, causing ‘want’ without ‘like’ (Berridge 2001).

The bases of the modern incentive theory came from contributions from Bolles (1972), Bindra (1974) and Toates (1986; 1994), known collectively as the Bolles-Bindra-Toates Theory (Berridge 2001). The Incentive Salience Theory arose as a result of factors that were not accountable under reward theory, such as the existence of salient stimuli, such as lights and tones that result in rapid phasic dopamine release, but are not rewarding (Ljungberg et al. 1992; Horvitz et al. 1997). In addition, novel stimuli, aversive (non-rewarding) stimuli, and even stimuli that predict aversive events, were seen to increase firing of some dopamine neurons (Schulz & Romo 1987; Guarracui & Kapp 1999; Ikemoto & Panksepp 1999).

Implications of the theory suggest that incentive and hedonic value of stimuli rise and fall with changes in internal drive and that the drive states may be potentiated by encountering external incentives or a conditioned stimulus representing those incentives. Furthermore both external incentive stimuli and internal physiological cues are necessary for motivation to occur; it is the combination that is important. The model suggests that Pavlovian incentives become both ‘liked’ and ‘wanted’ as a consequence of reward learning and, consequently, conditioned incentive value is equivalent to conditioned hedonic value (Berridge 2001).

Opposition to the Incentive Salience Hypothesis exists, for example, Ungless (2004) offers an alternative explanation for the anomalies that led to the incentive salience explanation. He argues that midbrain dopamine neurons are activated specifically by reward rather than by all salient stimuli, and that reward theory provides a better explanation of the role of dopamine, and is able to incorporate the effects of salience. Indeed novelty in itself may be rewarding and salience may therefore be accommodated within reward theory. However, since publishing the Incentive Salience Hypothesis (Berridge & Robinson 1998) there have been a considerable number of papers published that incorporate incentive salience into their findings, covering a large range of topics including schizophrenia and attentional disorders, drug addiction and consummation of pleasurable foods, nicotine and alcohol intake.

The incentive salience hypothesis would appear to be a robust finding at the present time and a relatively solid base upon which to build.

In order to link the idea of incentive salience to schizophrenia I describe in Section 4.4.1 a framework by Kapur (2003) that sees psychosis as a state of *aberrant* salience.

4.4.1 Psychosis as a state of aberrant salience (Kapur 2003)

The framework builds on the Incentive Salience Hypothesis and suggests that dopamine is a mediator of contextually relevant saliences by converting stimuli into either attractive or aversive representations which can then be acted on accordingly. Here dopamine production is stimulus-linked and context-driven, converting motivation into action, and has the ability to predict reward and pleasure, allowing us to focus on what we think is important. Kapur posits that during psychosis excess levels of dopamine result in the creation of 'aberrant saliences,' where the release of dopamine is independent of context and no longer stimulus-linked.

The strength of the framework lies in its ability to explain psychosis and its progression, and the effects of antipsychotic drugs exhibited by patients. Psychosis manifests itself in patients as delusions, hallucinations and the secondarily related behaviour arising from those positive symptoms. An individual will be diagnosed as psychotic when the symptoms interfere with their thoughts and actions making it impossible to lead a normal life. The framework accommodates these symptoms and suggests delusions to be a disorder of inferential logic, where the individual will attempt to apply top-down cognitive explanations for the fears and anxieties experienced as a result of the aberrant saliences. Hallucinations are considered to be exaggerated, amplified and aberrantly recognised internal percepts which emerge from the delusions. Secondarily related behaviours and the disruption to daily life arise as a direct result of these manifestations.

Endogenous psychosis evolves in stages over a period of time, and psychosis arising from use of amphetamine will not usually result from one single exposure to the drug. The framework accommodates these findings by suggesting that there is a gradual build-up of exaggerated, stimulus-independent release of dopamine before the onset and diagnosis of psychosis eventually leading to the inappropriate assignment of salience to stimuli. Prior to psychotic episodes, individuals often report feeling a heightened sense of awareness, perplexity and anxiety and are unable to account for their experiences. Over a period of time and with persistently increased levels of dopamine, full-blown psychosis begins to emerge, with delusions and hallucinations, as patients attempt to make sense of their situation. In this way the framework also accounts for the personal nature of the delusions to an individual as each will make sense of their delusions according to their own environment and culture. Differences in the severity of hallucinations, from internal thoughts to alien voices, can also be attributed to these aberrant saliences and their effects on individual's internal percepts. In conclusion the same dopamine imbalance between individuals can lead to many different manifestations within individuals.

When delusions and hallucinations have sufficient impact on the individual, adversely affecting their behaviour, antipsychotics can help to alleviate the symptoms. It was previously believed that antipsychotic drugs did not provide overnight success and, while blocking of the dopamine D2 receptors began within hours of taking the drug, there was a delay of 2-3 weeks before an improvement was seen in positive symptoms. However, Kapur and colleagues (Kapur 2004; Kapur, Mizrahi & Li 2005) found that an anti-'psychotic' effect was evident in patients experiencing a psychotic exacerbation who were treated within the first 24 hours. These findings suggest a close relationship between dopamine, psychosis and the action of antipsychotics and lend further support to the framework.

Antipsychotics do not provide a cure as patients still report a core-belief in the truth of their delusions, the symptoms simply lie dormant. But antipsychotics do help the patient to lead a more normal life by stopping the delusions from interfering with thought and function. Typical antipsychotics are often of limited use as patients often suffer from dsyphoria or a deficit-like state, where they report an indifference to life and the dampening of motivation, drive and pleasure. As a result many patients prefer not to take these drugs and will often stop the medication with the result that the symptoms return as the drugs are only effective while being taken.

The framework proposes that antipsychotics dampen aberrant saliences by blocking excess dopamine, leading to an attenuation of motivational salience of ideas and perceptions. Antipsychotics remove the degree to which symptoms occupy the mind, but not the core content of the symptom. They simply provide a neurochemical balance where dopamine levels return to normal, new aberrant saliences are less likely to form and existing ones are more likely to stop. It is only in the weeks to come that an individual may work through and resolve their delusions in their own time. In this way the delusions and hallucinations may be deconstructed, but this is not always the case as some patients are never able to resolve their symptoms psychologically. Implications arising from the framework suggest that patients need psychological help and time, as well as medication, to work through their remaining delusions and hallucinations once the aberrant saliences have been dampened. In this way the framework accounts for an individual's persistence in the core-belief of their delusions, even when the delusions do not bother them any more. Antipsychotics only provide remission and no cure as symptoms return when the drugs are no longer taken. Without the dampening effect, the aberrant saliences and eventually the same delusions and hallucinations will re-emerge from dormancy.

To summarise, the framework suggests a way of uniting a patient's positive symptoms with the Dopamine Hypothesis and the pharmacological interventions of antipsychotic drugs, and provides a plausible explanation of how excess dopamine levels may lead to psychosis, linking brain and mind. It suggests that under normal conditions dopamine acts as a stimulus-linked mediator of contextually relevant saliences, whereas in psychosis dopamine is a stimulus-independent creator of aberrant saliences. It successfully incorporates the nature, progression and personal nature of psychosis and the effects of antipsychotics on psychosis. However, it is only a framework and does not attempt to specify the nature of the relationship between the brain and the mind. It only looks at the effects of the dopamine imbalance on the positive symptoms of schizophrenia and does not take into account other negative and cognitive symptoms. Current research in schizophrenia is beginning to focus also on the roles of other neurotransmitters (such as glutamate and

GABA), and neurodevelopmental, cognitive and interpersonal deficits that occur before the effects of dopamine. Not all patients respond to antipsychotics and so dopamine transmission via D2 receptors does not fully explain psychosis. Atypical antipsychotics still act on the same D2 receptors but often have less distressing side effects and could provide clues in the future of other possible actions on either the dopamine system or of a different neurotransmitter system.

4.4.2 Interim Conclusions

A model of the role of dopamine in psychosis and the positive symptoms of schizophrenia should ideally incorporate the ideas of incentive salience. In Chapters 5 and 7 I describe models by McClure, Daw and Montague (2003) and Smith, Li Becker and Kapur (2004; 2006), respectively, to demonstrate how the ideas of incentive salience may be addressed by future neurocomputational models of dopamine function. In Chapter 7.1.3 I look at how a model based account of dopamine function by Smith et al. (2006) can relate to understanding psychosis and schizophrenia.

4.5 Functional Magnetic Resonance Imaging Evidence

Dopaminergic models provide a way to understand neuroimaging experiments on reward expectancy and cognitive control in humans (Montague et al. (2004) and, in turn, functional magnetic resonance imaging (fMRI) studies have provided evidence that the reward prediction error model of dopamine activity applies to human reward learning and not just primates. Noninvasive neuroimaging studies have examined brain responses to a broad range of rewarding stimuli, including money, art and social rewards and found a striking consistency in the set of neural structures that respond, which include the orbitofrontal cortex, ventral striatum and ventromedial prefrontal cortex (Montague, King-Casas & Cohen 2006).

McClure, Berns & Montague (2003) and O’Doherty et al. (2003) found transient learning-related changes in both the striatum (putamen), and the orbital frontal cortex of the brains of humans subjected to classical conditioning procedures. These patterns were consistent with predictions from a TD model of learning, confirming the role of both brain areas in reward prediction learning (Braver & Brown 2003). Furthermore, O’Doherty et al. (2004) demonstrated a dissociation between ventral and dorsal striatum in instrumental conditioning, where dopaminergically controlled plasticity in the ventral striatum was associated with the critic, while the dorsal striatum was better associated with the actor. In addition, Seymour et al (2004) used fMRI to show that neural activity in the ventral striatum and the anterior insula corresponded to the signals for sequential learning predicted by TD models, in humans in higher-order learning.

4.6 Chapter Conclusions

In the above five sections I have described:

- The idea of dopamine acting as a reward prediction error.

- TD as an effective method of modelling the dopamine reward prediction error signal, and how outputs from dopamine neurons in the basal ganglia could be used to reinforce behaviours leading to reward, using an actor-critic architecture.
- The functional role of the basal ganglia in contextual analysis.
- The Incentive Salience Hypothesis.
- Evidence of the dopamine reward prediction error in humans from fMRI studies.

Chapters 2, 3 and 4 gave rise to the important research question of 'How do computational models of the role of dopamine as a reward prediction error map on to current dopamine theories.' A paper addressing this question (Thurnham, Done, Davey & Frank 2006a) was published in the proceedings of XXV111 Annual Conference of the Cognitive Science Society, Vancouver, Canada, 26-29 July 2006, pp 2263-2268, Lawrence Erlbaum Associates and can be found in Appendix II.

The sections in this chapter provide the methodology behind a model by McClure, Daw and Montague (2003) that incorporates the Incentive Salience Hypothesis into an Actor-Critic model of dopamine as a reward prediction error. In Chapter 5 I describe and implement a simulation of the Computational Substrate for Incentive Salience by McClure et al. and this model is expanded to look at the relationship between TD learning and uncertainty coding in Chapter 6.

Chapter 5

A Model Free Account of Dopamine Function: Temporal Difference Learning

In Chapter 3, I described the search for an effective connectionist model to be used as a paradigm for schizophrenia, with a view to generating and testing theories of the disorder. My simulations in that chapter pointed to other modelling work regarding the role of dopamine in learning, by Houk, Adams and Barto (1995), Montague, Dayan and Sejnowski (1996) and Schultz, Dayan and Montague (1997). The ideas behind this line of research were explored in Chapter 4, explaining: (i) the idea of dopamine acting as a reward prediction error; (ii) the role of the basal ganglia in the production of that signal; (iii) Temporal Difference (TD) as an effective method of modelling the dopamine reward prediction error signal; (iv) the Incentive Salience Hypothesis; and (v) evidence of the dopamine reward prediction error in humans from fMRI studies.

The theory explored in Chapter 4 pointed to an interesting model by McClure, Daw and Montague (2003) incorporating the Incentive Salience Hypothesis into an Actor-Critic model of dopamine as a reward prediction error; a possible base for my research. This model is described in detail in Section 5.1, and in Section 5.2 I describe my own implementation based on the original model, which replicates and explores the computational substrate for incentive salience by McClure et al. (2003). In particular, I highlight the difference between higher and lower levels of dopamine receptor antagonism in an attempt to reveal the possible dual function of dopamine: as a learning signal and in action selection.

5.1 A Computational Substrate for Incentive Salience (McClure, Daw & Montague 2003)

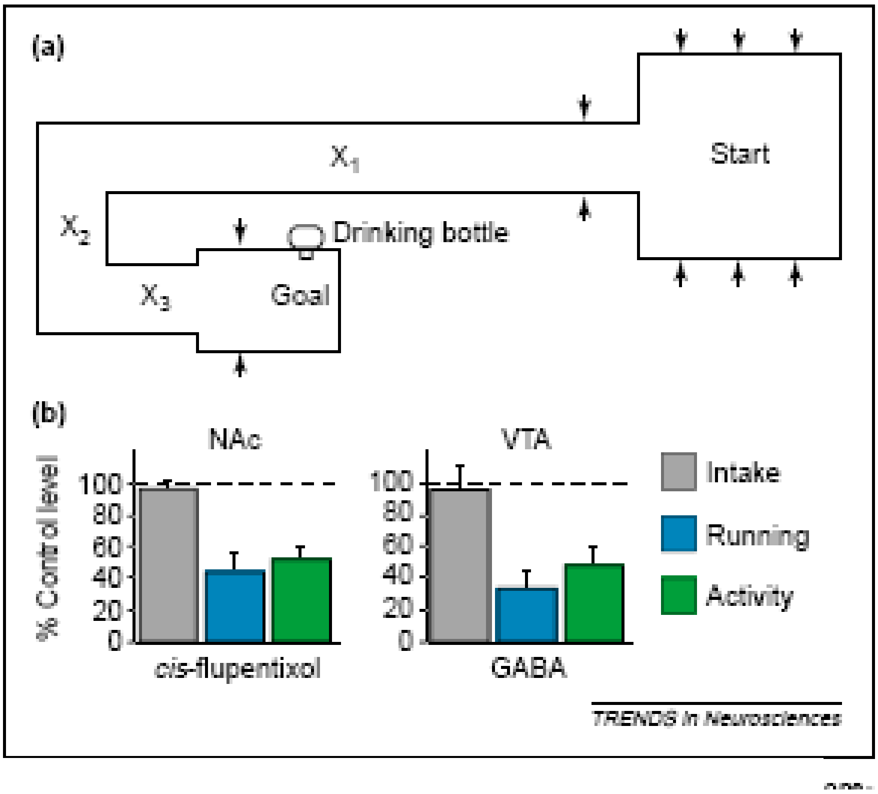
McClure et al. (2003) highlighted a gap that existed at that time between what appeared to be conflicting theories of dopamine function. On the one hand there were computational models of dopamine as a reward prediction error signal with two roles in learning and action selection (e.g., Montague, Dayan & Sejnowski 1996; Schultz, Dayan & Montague 1997), while other pharmacological lesioning studies identified a role for dopamine in the allocation of incentive value rather than reward (e.g., Berridge & Robinson 1998; Ikemoto & Panksepp 1996).

In order to bridge this gap McClure and colleagues incorporated the concept of incentive salience (Chapter 4.4) into a model of dopamine as a reward prediction error (Chapter 4.1), capturing the temporal nature of the relationship with TD and an Actor-Critic architecture (Chapter 4.2), which included the role of the basal ganglia in the contextual analysis of the environment (Chapter 4.3). This computational substrate successfully united these psychological and formal computational theories by interpreting expected future reward as incentive salience.

The Incentive Salience Hypothesis separates hedonic ‘liking’ from ‘wanting,’ and dopamine release is thought to assign incentive value to objects or acts, transforming ‘liked’ into ‘wanted,’ thus enabling reward-seeking behaviours (Berridge & Robinson 1998). The more valuable the action, the more likely it is to be selected by the action selection system. The approach by McClure et al. sees dopamine receptor antagonism, characteristic of the effects of antipsychotic drugs, as the inhibition of the ability to initiate actions necessary for gaining rewards, that is, it affects the ‘wanting’ without having an effect on the actual value of the reward, the ‘liking’. In this way dopamine receptor blockade does not influence the assignment of value, but does inhibit the use of such values.

In order to illustrate the concept of incentive salience with the computational notion of expected future value, McClure and colleagues simulated the findings in two sets of animal experiments. Firstly, they modelled the effects of a large concentration of dopamine receptor antagonist, in accordance with Ikemoto and Panksepp (1996), who found that dopamine receptor antagonism in rats affected the ability to ascribe incentive salience (want) without affecting the motivation for the reward (like). Secondly, they modelled the different effects of lower concentrations of dopamine receptor antagonism in a similar experiment by Wise et al. (1978). The difference in the pattern of the dopamine response between these two simulations provided evidence for the dual role of the dopamine reward prediction error as (i) a learning signal (Chapter 4.3.1), and (ii) in action selection (Chapter 4.3.2). They suggested that this dual role served as the formal counterpart to the ideas of Berridge and Robinson (1998) about the role of dopamine in attributing and using incentive salience.

In Ikemoto and Panksepp (1996), rats were trained in a one-arm maze (Figure 5.1a), to obtain a sucrose reward. Dopamine receptor antagonism, either by application of the dopamine receptor antagonist *cis*-flupentixol in the nucleus accumbens or by injection of GABA into the ventral tegmental area, resulted in (i) significantly slower running speed in the maze immediately after drug delivery (Figure 5.1b, blue bars), severely disrupting the approach (wanting) and (ii) reduced baseline activity in the start box (Figure 5.1b, green bars), outside the context of the task. However, once the rats reached the sucrose reward they still drank the same amount and ‘liking’, or motivation for the reward, was unaffected (Figure 5.1b, grey bars). Dopamine was interpreted as enabling reward-seeking behaviours, supporting the Incentive salience Hypothesis (Berridge & Robinson 1998).



TRENDS in Neurosciences

Figure 5.1. Results of animal experiments by Ikemoto and Panksepp (1996): (a) Rats were trained to traverse a one-armed maze to obtain sucrose solution at the end. Photosensors (arrows) were used to determine running speed [blue bars in (b)], in addition to the baseline level of movement in the start box [green bars in (b)], while access to the runway was blocked by a door. X_1, X_2 *and* X_3 represent intermediate states in the model. (b) After training, dopaminergic neuron activity was reduced (see text). Both manipulations reduced the ability of the rats to initiate the running needed to acquire the sucrose solution [$P<0.01$; loss of ability to ascribe incentive salience (Ikemoto and Panksepp 1996)], while leaving the volume of sucrose they consumed unaffected (grey bars). Figure taken from McClure et al. (2003).

Specifically, McClure et al. modelled the effect of dopamine receptor antagonism resulting in slower running speeds in the maze (Figure 5.1b, blue bars), severely disrupting the approach (wanting). This was achieved using a constant decrease in the TD error signal, and provided a similar pattern of results as the original animal experiment by Ikemoto and Panksepp (1996). McClure et al. achieved a 75% reduction in running speed (Figure 5.2b), which favours well with the animal experiments by Ikemoto and Panskepp, who achieved around a 60% reduction using *cis*-flupentixol in the nucleus accumbens and about 70% reduction in the ventral tegmental area (Figure 5.1b).

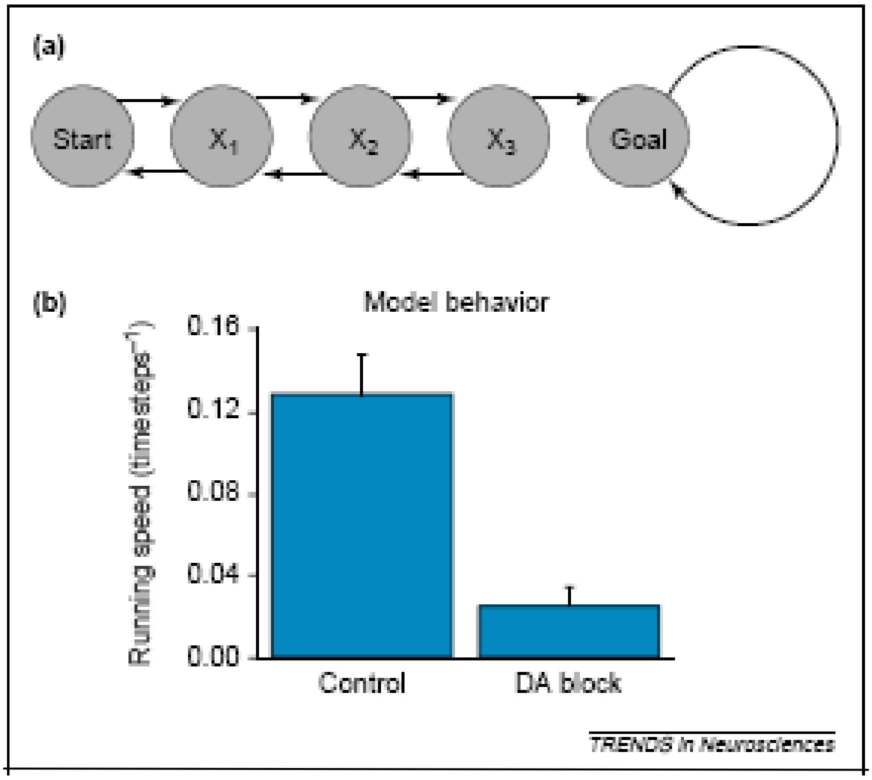


Figure 5.2 Taken from McClure et al. (2003): TD model captures rat behaviour in a maze. (a) The maze for a virtual rat is represented as five states. Movements between states were determined by considering state transitions, which produced a reward prediction error signal, *δ(t)*, and a probability of taking the considered action. (b) When virtual dopamine receptor antagonists are added (Chapter 4.3, Equation 4, where *b*=1), the speed the rat traverses the maze is significantly decreased, as seen in real animals (e.g. Ikemoto and Panksepp 1996).

It was explained that in a dopamine-rich environment, a reliable pattern of values built up over time using the phasic dopamine signal, which predicted the occurrence of the reward, and the rat learned the optimum route to the reward and was able to traverse the maze quickly. However, following dopamine receptor antagonism, in a dopamine-deficient state, the attribution of incentive salience associated with the reward (or conditioned stimulus) was disrupted, which had the immediate effect of reducing running speed in the maze. This reduction in running speed was explained as a discouragement of motivated behaviour through both the *direct* and *indirect* effects of dopamine (see Methods Section). High concentrations of dopamine receptor antagonism caused an immediate decrease in incentive salience via the TD error signal, disrupting the actions that led to reward.

Wise et al. (1978) also used a one-arm maze to test the effects of low concentrations of the dopamine receptor antagonist pimozide on the behaviour of rats. Unlike Ikemoto and Panksepp, they found no immediate effect of the drug on running times (day 1), but effects were seen when tested one week later (day 2), in a similar manner to a different group of rats from which food was withheld at the goal [No reward (NR) condition], to model extinction (Figure 5.3a). McClure and colleagues captured the different patterns seen with low concentrations of antagonists (Figure 5.3b) and described the effect as a discouragement of motivated behaviour through only the *indirect* effects of dopamine (see Methods Section). Here, low concentrations of pimozide or dopamine receptor antagonism caused a progressive decrease in incentive salience through reductions in the TD error signal, disrupting the actions that led to reward, having a similar effect to extinction.

McClure and colleagues claimed that the differences in the patterns of behaviour seen in the above two animal experiments resulting from high and low dopamine receptor antagonism, reflected in the timing of the changes to running speed, reveals the dual function of dopamine, as a learning signal and in the bias of action selection. The immediate effect of the reduction in running speed seen with high dopamine receptor antagonism in Ikemoto and Panksepp (1996) was a result of the discouragement of motivated behaviour through both the *direct* and *indirect* effects of dopamine, while the delayed, progressive effect of low concentrations of dopamine receptor antagonism in Wise et al. (1978) arose as a result of the *indirect* effects of dopamine only, through the slow unlearning of the value estimates, characteristic of the effects of experience-dependent extinction. When the changes in levels of dopamine were too weak to disrupt the *direct* effect on action selection, there was still an *indirect* effect on learning the values of states.

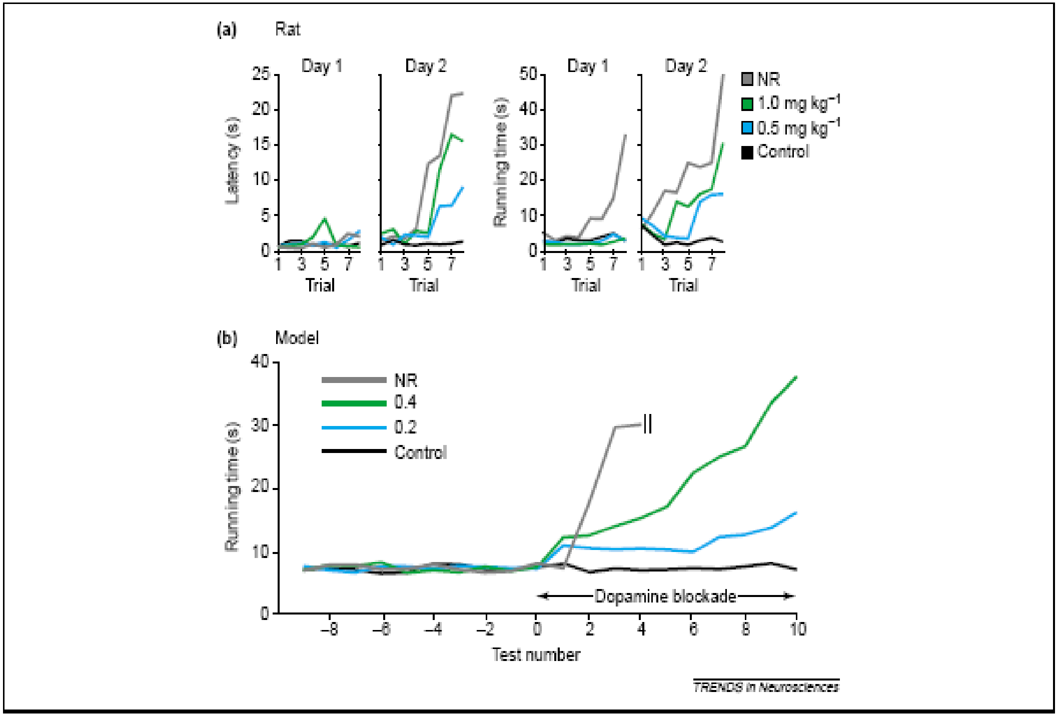


Figure 5.3 Taken from McClure et al. (2003). (a) Using data from Wise et al. (1978): When rats are given low concentrations of the dopamine-receptor antagonist pimozide after learning to run a maze for a food-pellet reward, behavioural effects are not immediately seen. Instead both latency and running speed remain unaffected during repeated trials on the first training day (day 1). When tested again one week later (day 2), deficits become apparent and become stronger in an experience-dependent manner. Rats show a slow increase in latency before entering the maze, and decreased running speed through the maze [0.5mg kg^{-1} (blue) and 1.0 mg kg^{-1} (green) pimozide conditions] in a manner that parallels the effect of extinction (NR condition, grey). (b) This effect is captured by the prediction error hypothesis as a slow decrease in estimated values (V) of each state in the maze. The different concentrations of pimozide were captured in the model as different scalar values, b, subtracted from the dopamine signal, δ[b – 0.2 (green) and b = 0.4 (blue), arbitrary units]. Extinction (NR, grey) is captured by setting to zero the reward value of arriving at the goal box.

5.2 An Investigation of the Effects of Dopamine Receptor Antagonism on Running Speed in a Maze

Having reviewed the computational substrate for incentive salience by McClure et al., a number of interesting research questions arose that warranted further investigation. The current study was influenced by the model of McClure et al., where my aim was to repeat the simulations in order to provide an exploration into the changing parameters of the system. This will help to highlight the difference between the high and low concentrations of dopamine receptor antagonism in the two animal studies, providing a clearer interpretation of the dual function of dopamine as a learning signal and in action selection. My simulations will provide additional weight to the claims of McClure and colleagues that their single model can capture the ideas of dopamine (i) as a reward prediction error and (ii) as a purveyor of incentive salience, by uniting psychological and formal computational theories and interpreting expected future reward as incentive salience.

In order to ascertain that my version of the model was working correctly, I simulated the results of animal experiments by Ikemoto and Panksepp (1996) and Wise et al. (1978) by looking at (A) the effects of a high dopamine receptor antagonism and (B) the effects of lower levels of dopamine receptor antagonism on running speed, respectively. In addition to replicating the simulations of McClure and colleagues, I was able to look into their claims in greater depth than the limitations of a journal publication will allow. While simulating Ikemoto and Panksepp, I looked at the following points which were not addressed in the original paper: (i) the *indirect* effect of the bias, *b* (Equation 5.2), on the values of states in the maze; (ii) the effect of a range of biases on running speed; (iii) the *direct* effects of changing the bias on the shape of the sigmoid decision curve from Equation 5.2; and (iv) modelling dopamine receptor antagonism via changes to the scaling constant *m*, (Equation 5.2).

Modelling the animal experiments of Wise et al. also led to the following additional investigations that were not addressed in the original paper: (i) a comparison between high and low dopamine receptor antagonism; (ii) the effect of extinction on running speed; (iii) a look at the effects of high and low dopamine receptor antagonism on the TD error; and (iv) the *indirect* effects of high and low dopamine receptor antagonism on the values of states in the maze. These additional investigations will provide further evidence for the dual function of dopamine as a learning signal and in action selection.

5.2.1 Methods

The Model

The TD model by McClure and colleagues, inspired by the decision model of Egelman and Montague (1998), is described in detail in Chapter 4.2. The model incorporated the concept of incentive salience as expected future reward, where the dual roles of dopamine: (i) to bias action selection towards situations that predict the best reward; and (ii) as a learning signal used to update the values of states and so create better estimates of future reward, equated with the Incentive Salience

Hypothesis (Berridge & Robinson 1998) and the role of dopamine in attributing and using incentive salience. In both TD and incentive salience, increased dopamine levels increase the likelihood of choosing an action that leads to a reward.

The model is mathematically detailed and is capable of underpinning an artificial neural network. The aim was to model the acquisition process where dopamine signals the difference between predicted future reward and actual reward. McClure et al. used the concept of TD learning using trial by trial computation, which requires less memory and less peak computation than conventional methods (Sutton 1988) and there is no need for any explicit representation other than using received reward to learn a value function which maps current information to a prediction of expected future reward.

I implemented a computer program in LISP (Steele 1990), incorporating the TD algorithm with an Actor-Critic architecture (see Appendix II). The program was capable of modelling the tasks of Ikemoto and Panksepp (1996) and Wise et al. (1978), where rats were trained to traverse a one-arm maze for a reward, in a similar manner to McClure et al. (2003).

The maze
Positions in the maze were captured by McClure et al. as five possible states: start, goal and three intermediate states X1, X2 and X3 (see Figure 5.2a). As in the animal studies, a reward with a value of 1 was given in the goal state. The other four states had reward values of 0.

According to the journal article the virtual rat was permitted to move towards or away from the goal from any position within the maze until it reached the goal, after which it was returned to the start state. However, the diagram of the maze provided (Figure 5.2a) did not show the return to the start state after reaching the goal state. Instead it showed a recurrent connection at the goal state, where it implied the opportunity to revisit that state and obtain the reward of that state an unlimited number of times. Samuel McClure was good enough to clarify the situation and confirmed that he treated the goal state as a terminal state and once the model reached the goal, it received a reward and was then reset (Personal communication 2005). Figure 5.4 shows the maze used in my simulations, which is in accordance with that contained in the text of McClure et al. (2003), but not with their diagram.

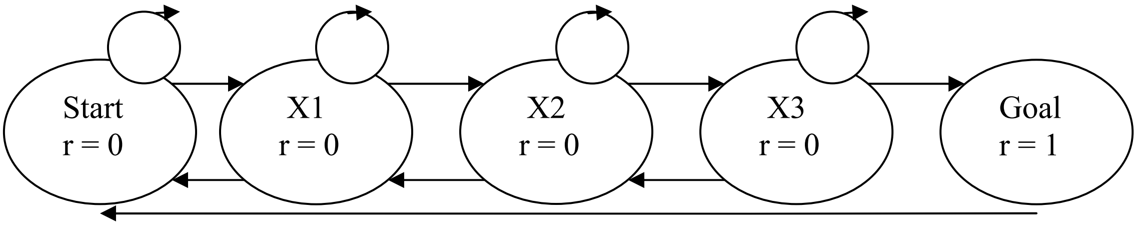


Figure 5.4 showing the maze used in our simulations, containing five possible states: start, goal and three intermediate states (X1, X2, and X3). A reward, with a value of 1, was given in the goal state. The other four states had reward values of 0. Arrows show permitted movement through the maze. There were recurrent connections on all states except for the goal state, where the only option was to return to the start state.

Modelling Time

McClure et al. modelled time by including the possibility of remaining in the same state without having to move on each trial. If dopamine levels decreased as a result of adding a bias, the likelihood of remaining in the same place increased, as the dopamine receptor blockade stopped the phasic dopamine bursts necessary to signal that one position was better than the other. All options became increasingly probable, and thus the time to complete the task increased.

As in McClure et al., time was recorded in this model as a consideration of a move through the maze from one state to another. Each consideration produced an error signal, *δ(t)* [Equation 5.1, (cf Chapter 4.2.1, Equation 4.4)], and was recorded as a timestep.

$$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t) \qquad \text{[Eqn 5.1]}$$

Whether or not the move was taken depended on the probability of taking that action, which depended on the TD error, *δ(t)* [Equation 5.2, cf Chapter 4.2.2, Equation 4.6)]. It is in Equation 5.2, where learned incentive value is converted into a probability of action that is equivalent to the Incentive Salience Hypothesis.

$$P\ (of\ taking\ action) = \left(1 + e^{-m(\delta(t)-b)}\right)^{-1} \qquad \text{[Eqn 5.2]}$$

As detailed in Chapter 4.2.2, the probability of taking this action was calculated from this *δ(t)* value using a softmax function, where the shape of the sigmoid decision curve had an effect on action selection. When the TD error was high and positive there was a greater probability of that action being selected than when the error was low and negative. If the move was not taken, the virtual rat remained in the same state pending another consideration or timestep. A timestep was recorded whether or not a move was actually made.

My simulations followed the methodology of McClure and colleagues as closely as possible and I provide clarification of the following terms which are specific to my simulations.

- A state transition was taken to be an actual movement from one state to another, following a timestep where an action was selected.
- One run through the maze was a series of state transitions from the start state to the goal state, which could be either towards or away from the goal. A run was concluded once the virtual rat was returned to the start state.
- The number of timesteps taken for a run differed according to the probability of a run being selected, which depended on the TD error. The probability of action selection increased as learning progressed and learning progressed with repeated iterations through the maze, as values of states converged to 1.

McClure et al. looked at the effect of increasing dopamine receptor antagonism, on running speed (timesteps^{-1}). In accordance with McClure et al., running speed was taken to be the reciprocal of the average number of timesteps for a state transition.

Modelling Dopamine Receptor Inhibition

In accordance with McClure et al., the effect of dopamine receptor antagonism was modelled: (i) *directly*, by subtracting a constant bias, *b* from the dopamine signal, *δ(t)* [Equation 5.2 (cf Chapter 4.2.2, Equation 4.6)]; and (ii) *indirectly*, by updating the stored value estimates according to Equation 5.3 (cf Chapter 4.2.2, Equation 4.5)].

$$V(s_i) \leftarrow V(s_i) + \alpha\delta(t) - b \qquad \text{[Eqn 5.3]}$$

According to Equation 5.2, fluctuations in dopamine signaling, or TD error, will affect the probability of action selection *directly*, by changing the shape of the sigmoid decision curve (Chapter 4.2.2). This will have an immediate effect on action selection, via the 'actor'.

Indirect effects of dopamine arise from updating the stored values of *V* according to Equation 5.3, where *α* is the learning rate and s_i is the previous state. Following receipt of a reward the values of each state are updated by adding a proportion of the prediction error, *αδ(t)*, where *α* is the learning rate, and subtracting *b*, a constant representing dopamine receptor inhibition. For example, if the current $V_{(s_t)}$ for a state was 0.30 and *αδ(t) - b* = 0.4 (0.5 – 0.2) = 0.12, where *α* = 0.4 and *b* = 0.2, the new value of $V_{(s_t)}$ would be 0.30 + 0.12 = 0.42. Here, it will take longer for the effects of dopamine antagonism to affect the system as the stored value estimates of the 'critic' are updated by successive runs through the maze.

Thus, the same dopamine reward prediction error can be seen to be acting in two different parts of the model: *Direct* effects of dopamine were seen on action selection (the actor) and *indirect* effects from its role in learning the estimated values that underlay the actions (the critic).

Simulations

Simulations were effected as detailed in the following sections:

A. Simulation of Ikemoto and Panksepp (1996)

In accordance with McClure et al. (2003) I investigated the effect of dopamine receptor antagonism on running speed by increasing the bias, *b*, after the value estimates were learned in a maze with a bias of zero, to a bias of 1, in Equations 5.2 and 5.3. Trials consisted of 1000 state transitions through the maze with no bias (*b* = 0), and then a further 1000 state transitions through the maze with a bias of 1 (*b* = 1). The effect of the change in bias was reflected in the number of timesteps taken for the second 1000 state transitions compared to the first 1000 state transitions, and the corresponding running speeds (timesteps^{-1}). Results are shown for the average values recorded for five different trials.

Modelling the animal experiments of Ikemoto and Panksepp also led to the following additional investigations that were not addressed in the original paper:

(i) The *Indirect* Effect of Dopamine Receptor Antagonism on the Values of States in the Maze

I show the effect of dopamine receptor antagonism on the values of states, by comparing the values of all the states in the maze for the first 1000 state transactions, when $b = 0$, to the second 1000 state transitions, when $b = 1$.

(ii) Looking at a Range of Biases

The effect of a range of biases, from 0 to 1.55 was investigated, on the number of timesteps and corresponding running speed for 1000 state transitions through the maze.

(iii) The *Direct* Effect of Adding a Bias to the Sigmoid Decision Curve

Here I demonstrated the *direct* effect of dopamine receptor antagonism on running speed, by showing the effect on the shape of the sigmoid decision curve of changing the bias, b, from Equation 5.2.

(iv) Investigating the Gain Constant, *m*, from Equation 5.2

The effect of changing the scaling constant, m, was investigated (see Figure 5.11 in Results Section). Timesteps and corresponding running speeds were recorded for 1000 state transitions through the maze for values of m, from 1 to 23, for four different biases of 0, 0.5, 1 and 1.5. I also looked at the effect on the sigmoid decision curve of changing the value of m in Equation 5.2.

B. Simulation of Wise et al. (1978)

As in McClure et al., I looked at the effect of lower levels of dopamine receptor antagonism on running speed. The journal paper did not explain precisely how running speed was modelled in this series of simulations, only that the values of bias used were $b = 0.2$ and $b = 0.4$. Their results were shown as increases in running time (unlike the decreases in running speed in Results Section A above) over 8 test numbers prior to a change in bias and for 10 test numbers after the change, modelling dopamine blockade (Figure 5.3b). I have therefore interpreted their running time as the number of timesteps for each state transition (their test number) during the test period, and have not converted these timesteps into running speeds.

My trials consisted of 1000 state transitions through the maze with no bias ($b = 0$), and then a further 1000 state transitions through the maze with dopamine receptor antagonism, modelled with biases of either zero (a control), 0.2, or 0.4. In addition, extinction was modelled by setting the reward in the goal state from 1 to zero, after the first 1000 state transitions, so that there was no reward given during the second set of 1000 state transitions. The effect of the change in bias or reward was reflected in the critical period between the first and second set of 1000 state transitions, where the number of timesteps was recorded for each of a number of state transitions prior to and after this critical period in each trial. In order to make a direct comparison to the results of McClure et al. and Wise et al., I detail the number of timesteps taken for each of the eight state transitions before any change of bias or reward (i.e., at 1001 state transitions), and for ten state transitions after any change. All other parameters of the model remained constant: m (scaling constant) was set to 5; γ (discount parameter) was set to 0.9 and α (learning rate) was set to 0.5.

Modelling the animal experiments of Wise et al. also led to the following additional investigations that were not addressed in the original paper that will provide further evidence for the dual function of dopamine as a learning signal and in action selection:

(i) A Comparison between High and Low Dopamine Receptor Antagonism
Here I compared the results for running speed obtained for high dopamine receptor antagonism, where b = 1, in (A) above, to the results for low dopamine receptor antagonism.

(ii) The Effect of Extinction on Running Speed
I compared the effects of extinction on running speed to the results in (i) above.

(iii) The Effects of high and Low Dopamine Receptor Antagonism on TD error
I looked at the effects of the control (b = 0), low dopamine receptor antagonism (b = 0.2 and 0.4), high dopamine receptor antagonism (b = 1) and extinction (r = 0), on the TD error, $\delta(t)$, for the goal states only, and recorded the results for the critical period for their respective changes (ten runs before and ten runs after the changes).

(iv) The *Indirect* Effects of High and Low Dopamine Receptor Antagonism on Values of States
By looking at the values of just the goal states for (A) 1000 state transitions prior to, and following the critical period, and (B) ten state transitions prior to, and 50 state transitions after the critical period, the effects of low or high dopamine receptor antagonism, or extinction on the values of states in general were ascertained.

5.2.2 Results

A. Simulation of Ikemoto and Panksepp (1996): High Dopamine Receptor Antagonism

In accordance with McClure et al., after the value estimates were learned for the maze with no bias (b = 0 for 1000 state transitions), I modelled dopamine receptor antagonism by increasing the bias from zero to 1 in Equations 5.2 and 5.3 (b = 1 for a further 1000 state transitions). The effect of the bias on the number of timesteps for 1000 state transitions through the maze for each bias can be seen in Figure 5.5, where the average number of timesteps rose from 2337 with no bias, to 11,347 with a bias of 1.

As a direct comparison to the results of McClure et al., Figure 5.6 shows the effect of the bias on the corresponding running speed (timesteps^{-1}). My simulations show a reduction in running speed of nearly 80% which is comparable to the 75% reduction achieved by McClure and colleagues (Figure 5.2b). Although, both sets of results are slightly higher than the original animal experiments by Ikemoto and Panskepp (1996), who achieved around a 60% reduction using *cis*-flupenthixol in the nucleus accumbens and about 70% reduction in the ventral tegmental area (Figure 5.1b). However, the figures for running speeds are parameter dependent and changes to m could give a better fit to the experimental data (see (iv) below).

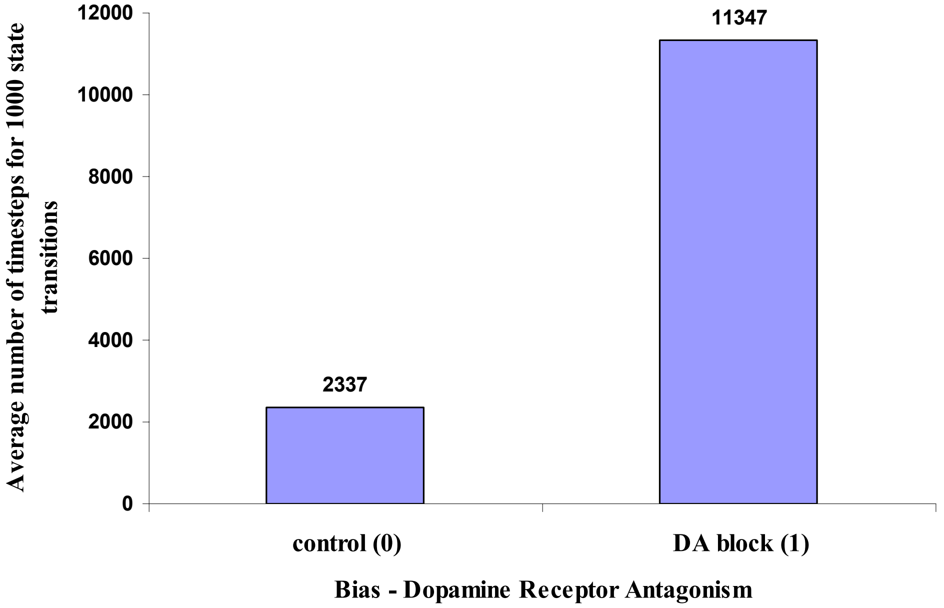


Figure 5.5 showing the effect of adding a bias of 1 in Equations 5.2 and 5.3, on the average number of timesteps for 1000 state transitions through the maze. ($\alpha = 0.5$, $\gamma = 0.9$ and $m = 5$).

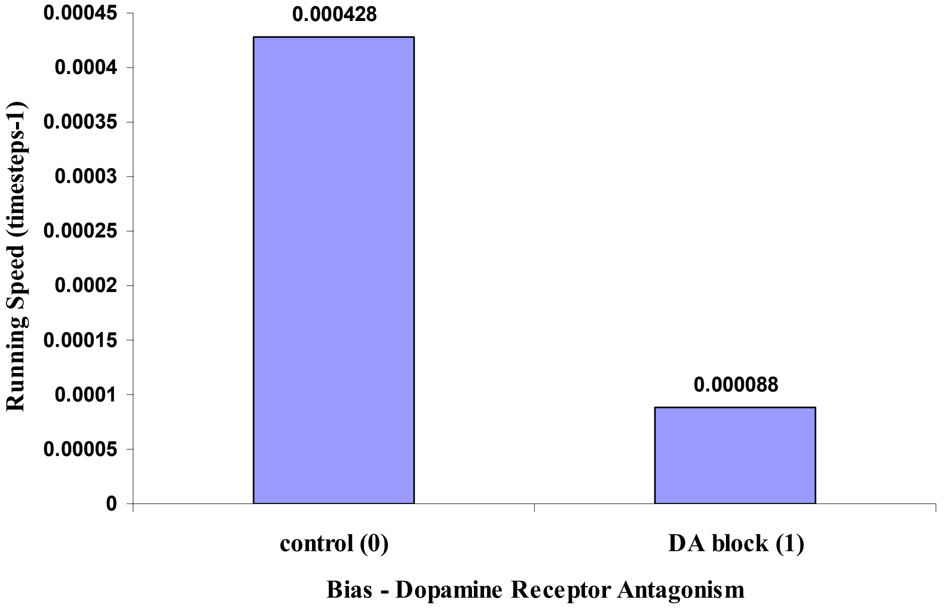


Figure 5.6 When virtual dopamine receptor antagonists are added (b=1), the speed the rat traverses the maze is reduced. This figure can be directly compared to the results of McClure et al. (2003) in Figure 5.2b and is intended to model the results of animal experiments by Ikemoto and Panskepp (1996) in Figure 5.1b. ($\alpha = 0.5$, $\gamma = 0.9$ and $m = 5$).

(i) The *Indirect* Effect of Dopamine Receptor Antagonism on the Values of States in the Maze

Figure 5.7 shows the values of states (A) for the first 1000 state transitions with no bias and (B) for the second 1000 state transitions with a bias of 1. The effect of adding a bias to the second 1000 state transitions is reflected in the first 300 or so state transitions in Figure 5.7B, where values of states change from a maximum value of +1.99 to a minimum of -8.31 (for the goal state). Each graph shows the *indirect,* gradual build up of values over time to their maximum value for each state. These *indirect* effects arise from Equation 5.3, by updating the stored value estimates according to a fraction of the TD error, $\delta(t)$. It can be seen that learning in both graphs took place over the first 300 or so state transitions and after learning had taken place, the values fluctuated around these maximum levels for the remaining 700 state transitions. The fluctuations seen thereafter resulted from the variability introduced to the model in respect of action selection and choice of route through the maze.

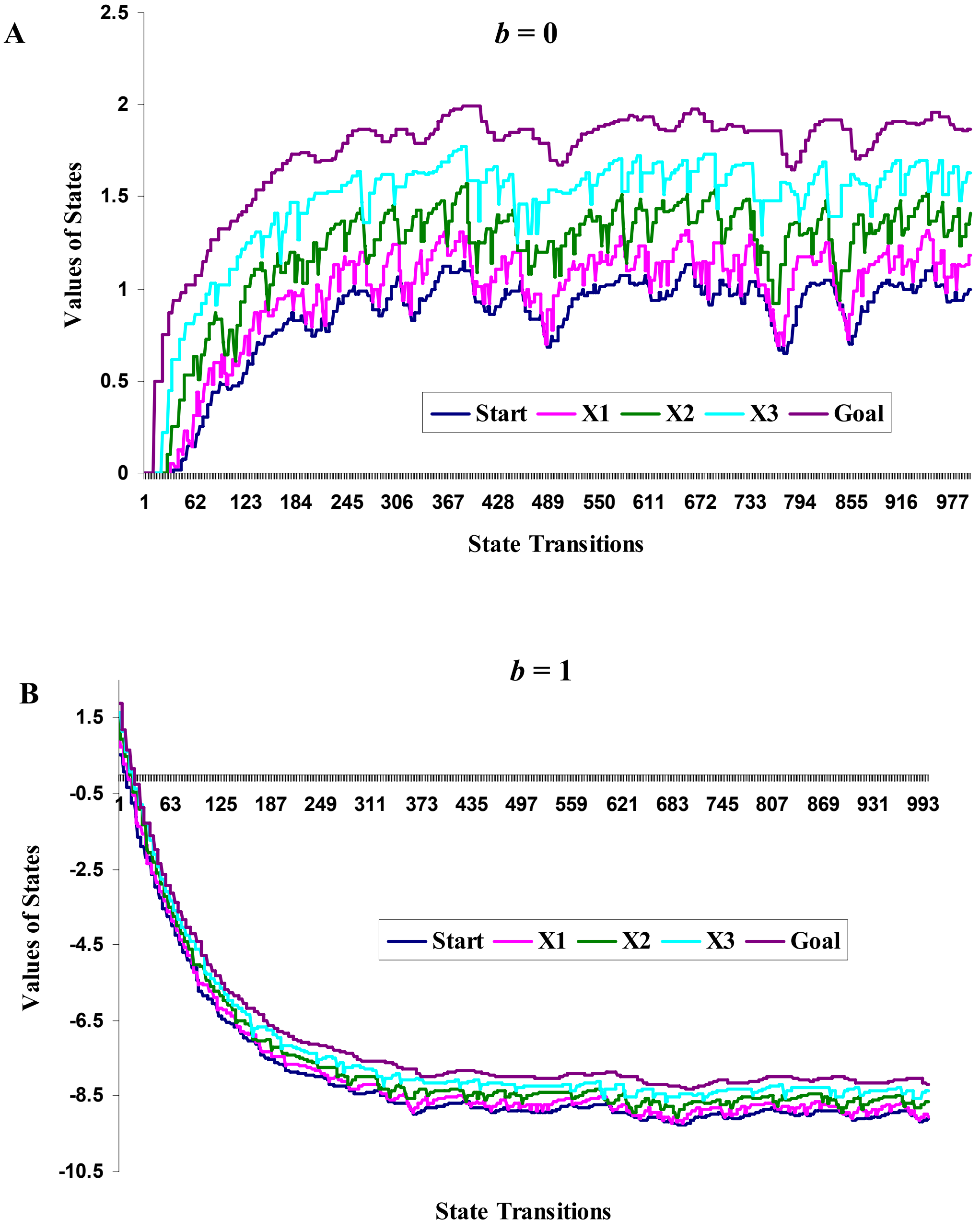


Figure 5.7 showing the values of states in the maze (A) for the first 1000 state transitions with no bias and (B) for the second 1000 state transitions with a bias of 1. (α = 0.5, γ = 0.9 and m = 5). NB note difference in scale of y-axis.

With the discounting factor, gamma, at 0.9, a preceding state will only achieve 90% of the total reward value from the state to its right. For example, the value of state 3 for a bias of 1 will reach 90% of the value for the goal state, and the value of state 2, 90% of the value for state 3, and so on.

(ii) Looking at a Range of Biases

Without modelling dopamine receptor antagonism, I investigated the effect of increasing values of bias, *b*, for 1000 state transitions through the maze for each value.

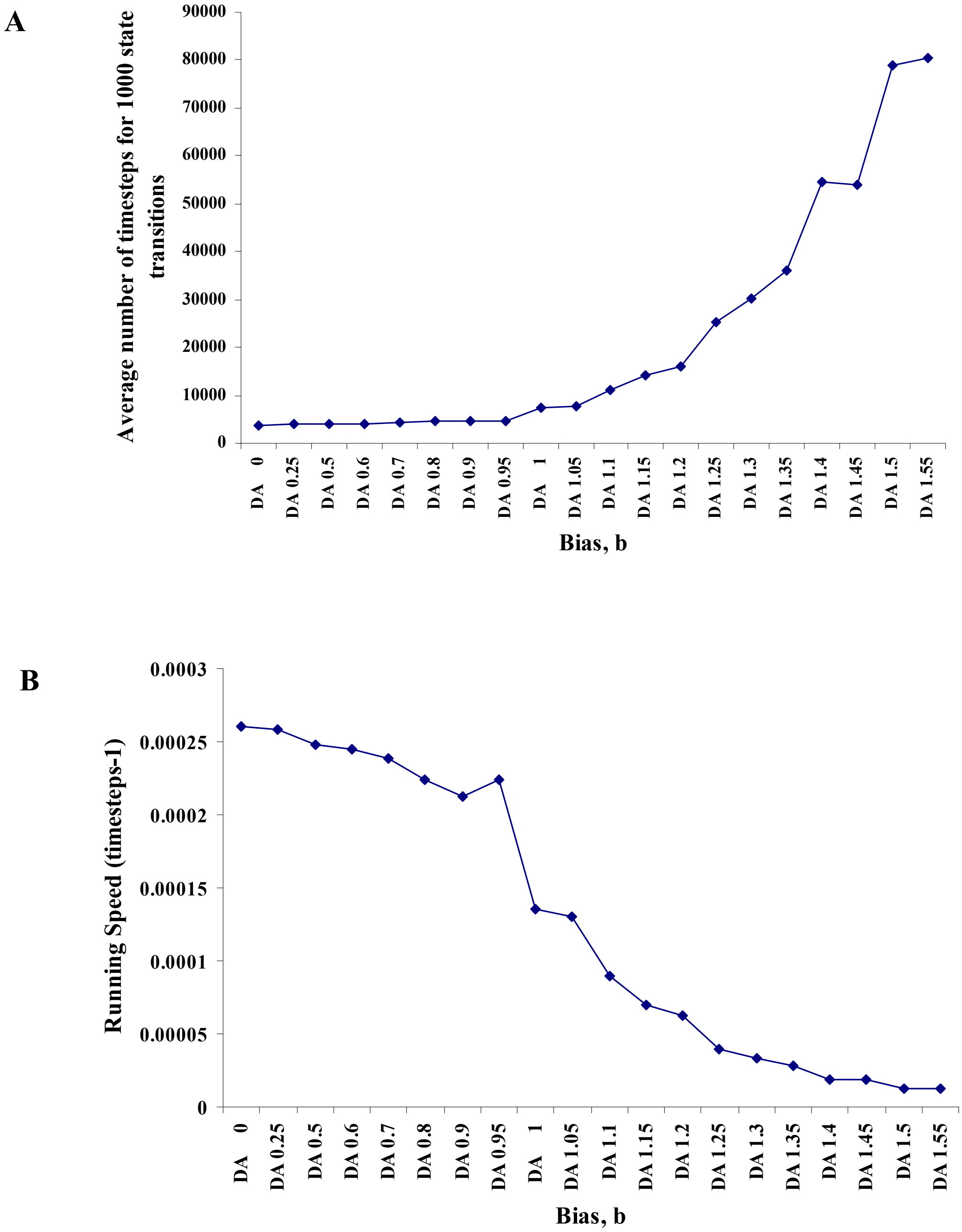


Figure 5.8 showing the exponential effect of increasing bias to model dopamine receptor antagonism on (A) the average number of timesteps for 1000 state transitions and (B) corresponding running speed (timesteps^{-1}). ($\alpha = 0.5$, $\gamma = 0.9$ and $m = 5$).

With this particular model I was able to increase the bias from a range of 0 to 1.55, but the program was unable to cope with a bias of over 1.55, as the number of timesteps became exponentially high. Figure 5.8 shows the effect of a using a bias on (A) timesteps and (B) corresponding running speed, respectively, for bias values of zero to 1.55. There is an exponential increase in timesteps with increasing bias and a corresponding exponential decrease in running speed with increasing bias.

(iii) The *Direct* Effect of Adding a Bias to the Sigmoid Decision Curve

A high level of dopamine receptor antagonism, where b = 1, will have a *direct* effect on action selection, resulting in slower running speed through the maze. According to Equation 5.2, fluctuations in dopamine signaling, or TD error, will affect the probability of action selection *directly*, by changing the shape of the sigmoid decision curve (Chapter 4.3.2). The effect of changing the bias, to the sigmoid function can be seen in Figure 5.9, where an increase in bias, from 0 to 2 results in an overall shift of the sigmoid to the right, altering the position in relation to the x-axis, but not the shape of the curve.

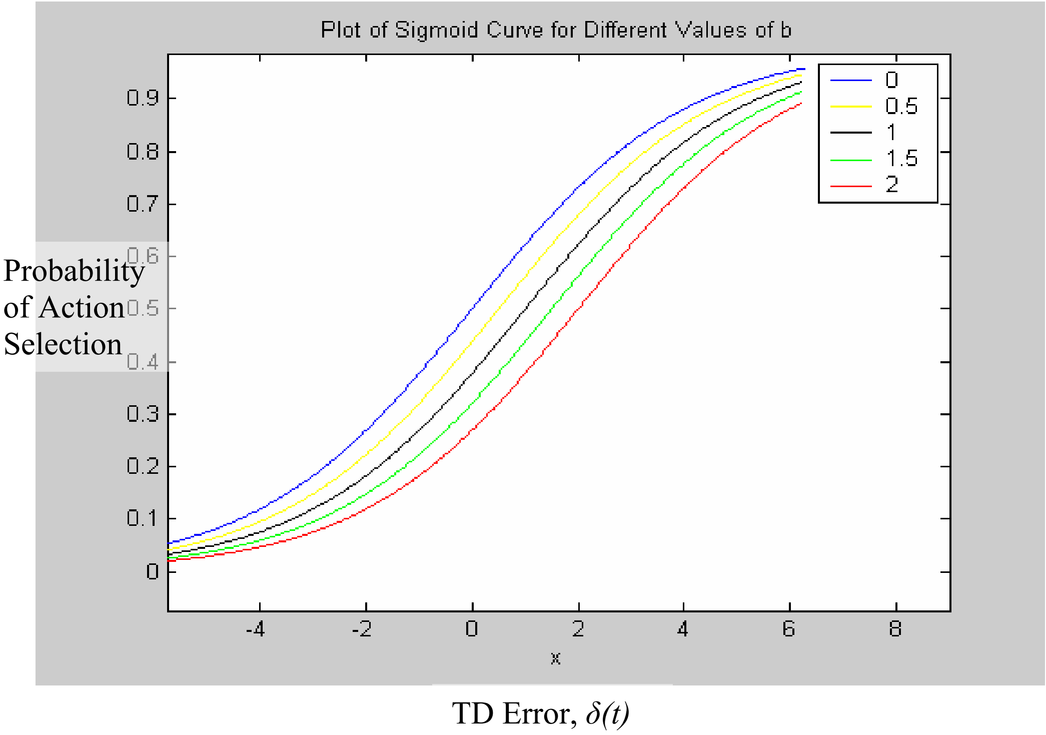


Figure 5.9 showing effect of increase in bias, on position of curve relative to x-axis, of sigmoid function. Here the x-axis represents the reward prediction error, *δ(t),* and the y axis, the probability of action selection. With an increase in bias there is a shift of the sigmoid to the right, resulting in a decrease in running speed.

Actions are generated with a probability of selection based on the predicted values of their successor states, preferring those actions that give a high burst of dopamine, or TD error signal. When the error signal is low there is a greater probability of remaining in the same state as all states become increasingly probable. During a timestep with a bias, there is an even greater probability of remaining in the same state and not making a state transition, as the shift in the sigmoid decision curve results in a lower probability of that particular action being taken. In effect, the part of the curve most sensitive to action selection, at around 0.5 on the *y*-axis, is shifted to the right as the bias increases. It can be seen from Figure 5.9 that the difference in the positions of the two curves, for a bias of 0 (blue line) and a bias of 2 (red line), would be reflected in the decisions made to make a state transition, or remain in the same state for another timestep. For example, reading off the *x*-axis for a TD error of zero would give a *y*-axis value for probability of action selection of around 0.5 with a zero bias, compared to a value of about 0.25 with a bias of 2. Thus an action would be more likely to be selected with a zero bias than with a bias of 2.

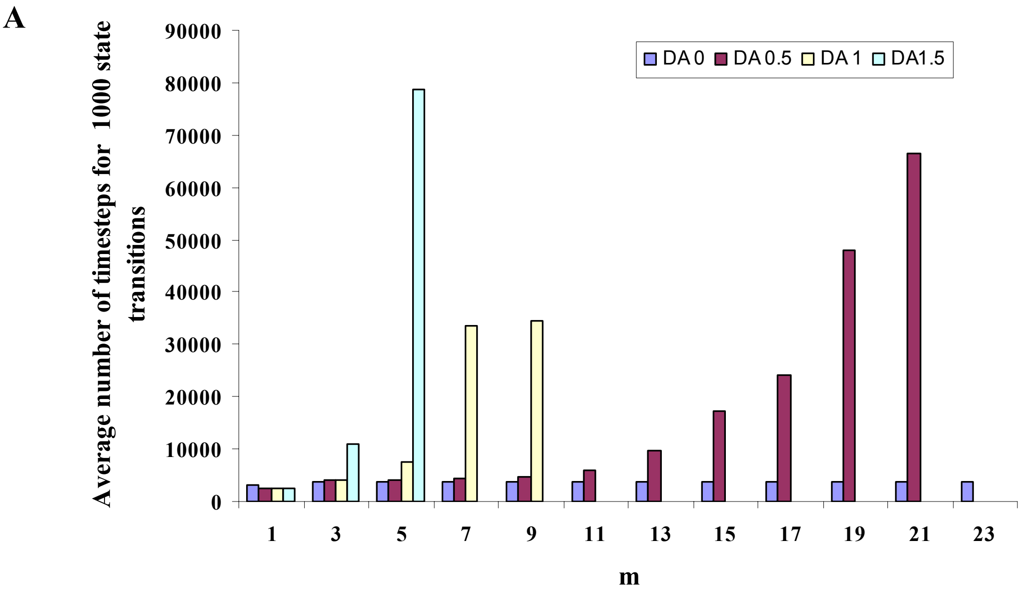
The overall effect of introducing a bias will result in a greater number of timesteps as the bias increases, which is reflected in Figures 5.5 and 5.8A. This will have a corresponding effect on running speed through the maze, reflected in Figures 5.6 and 5.8B, and is in accordance with the findings of Ikemoto and Panskepp (1996) and McClure et al. (2003).

(iv) Investigating the Scaling Constant, *m*, from Equation 5.2

I looked at the effect of increasing values of the scaling constant (*m*), from 1 to 23, for four different biases, 0, 0.5, 1 and 1.5, for 1000 state transitions through the maze. The program was unable to deal with any further increases in *m* above 5 for a bias of 1.5, 9 for a bias of 1, and 21 for a bias of 0.5, as the number of states examined above these values of *m* became exponentially high. The number of timesteps for 1000 state transitions and corresponding running speeds (timesteps^{-1}) were recorded in Figure 5.10, from which it can be seen that as values of *m* increase within biases of 0.5, 1 and 1.5, the average number of timesteps increases (Figure 5.10A) and the corresponding running speed decreases (Figure 5.10B).

In order to model the behavioural effect of cortical dopamine (Chapter 3.2), Cohen and Servan-Schreiber (1992; 1993) originally modified the net output of a neuron by increasing the ‘gain’ or steepness in slope of the sigmoid curve using *G*, analogous to the constant *m* in Equation 5.2, while keeping the bias constant. Although, the present study models a different process (the probability of action selection as opposed to the role of dopamine in optimising the signal-to-noise ratio thought to enhance working memory by reducing interference or noise), the two studies are comparable as they use either the ‘gain’ or the scaling constant *m* to change the ‘decision’ of the sigmoid function. Accordingly, it can be seen that there are two possible ways of modelling dopamine receptor inhibition: via the gain (*m*) or the bias (*b*).

The result in Figure 5.10, where *m* is 5 and *b* is 1, is comparable to that seen in Figure 5.8B, where *m* is 5 and *b* is also 1, as both figures show the same running speed of 0.000135. Increasing *m* would have a similar effect to modelling increased dopamine receptor inhibition to increasing the bias alone; although this was only effective when there was already a bias (i.e., 0.5, 1 or 1.5), as when the bias was 0, running speeds remained unchanged (approximately 0.00025 for each value of *m*, Figure 5.10B). By keeping the bias constant, the effect of a higher value of *m* alone modelled dopamine blockade by modifying the bias.



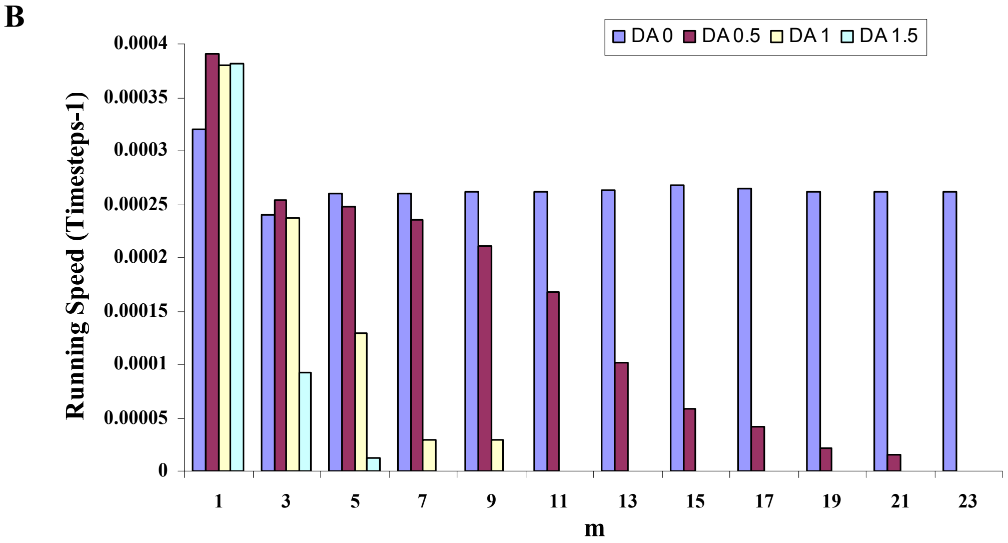


Figure 5.10 showing effect of *m* on (A) Number of timesteps for 1000 state transitions and (B) corresponding running speed, for biases of 0, 0.5, 1 and 1.5. ($\alpha = 0.5$, $\gamma = 0.9$).

The effect of changing the scaling constant, *m*, to the sigmoid function can be seen in Figure 5.11, where an increase in *m*, from 1 to 9 resulted in a change in the shape of the curve without altering the position with respect to the *x*-axis. With increasing values of *m*, the shape changes from a deterministic straight line at *m* = 1, to the typical *s*-shaped curve, characteristic of the sigmoid function and associated with biological systems, at *m* = 9.

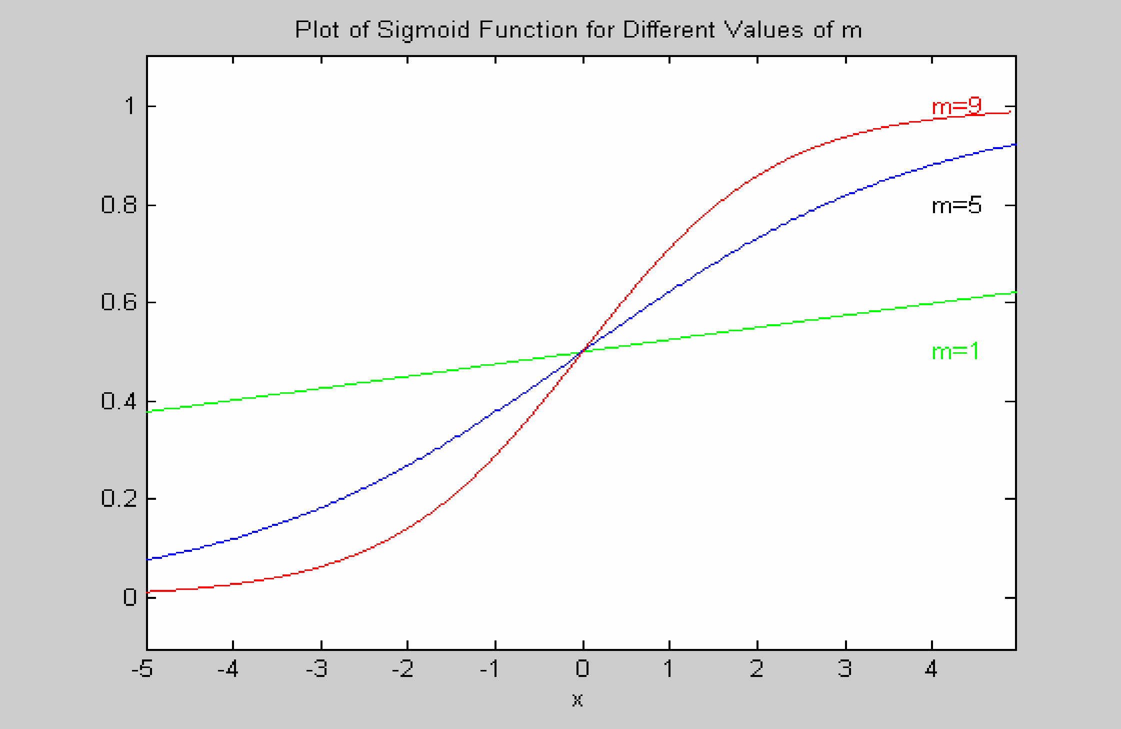


Figure 5.11 showing the effect of an increase in the scaling constant, *m*, on the shape of the sigmoid curve (Equation 5.2). Here the *x*-axis represents the reward prediction error, *δ(t*), and the *y*-axis, the probability of action selection. When the TD error is high there will be a greater probability of that action being selected than when the error is low.

B. Simulation of Wise et al. (1978): Low Dopamine Receptor Antagonism

In a similar manner to McClure et al., I modelled lower levels of dopamine receptor antagonism after the value estimates were learned for the maze with no bias (b = 0 for 1000 state transitions), by increasing the bias from zero to 0.2 (Simulation 2) and from zero 0.4 (Simulation 3) in Equations 5.2 and 5.3 (i.e., b = 0.2 or 0.4) for a further 1000 state transitions. Extinction was modelled by setting the reward in the goal state from 1 to zero, after the first 1000 state transitions, so that there was no reward given during the second set of 1000 state transitions (Simulation 5). These results were compared to a control condition, which consisted of no change in bias for the second set of 1000 state transitions through the maze, where the bias remained at zero (Simulation 1). These four simulations are summarised in Table 5.1.

Table 5.1 summarising the four simulations to determine the effects of lower levels of dopamine receptor antagonism. Simulation colours match those in the graphs in the figures below and changes between the first and second set of 1000 state transitions are highlighted in red.

	First 1000 state transitions	**Second 1000 state transitions**
Simulation 1 Control	No bias (b = 0) Reward given in goal state	No bias (b = 0) Reward given in goal state
Simulation 2 Low Dopamine receptor antagonism	No bias (b = 0) Reward given in goal state	Bias (b = 0.2) Reward given in goal state
Simulation 3 Low dopamine receptor antagonism	No bias (b = 0) Reward given in goal state	Bias (b = 0.4) Reward given in goal state
Simulation 4 Extinction	No bias (b = 0) Reward given in goal state	No bias (b = 0) No reward given in goal state

For Simulations 2 to 4, the critical period for examination was the change between the first and the second set of 1000 state transitions, where the changes of bias or reward occurred. The effects of these three simulations on the number of timesteps were compared to the control (Simulation 1) and recorded in Figure 5.12 for eighteen state transitions; eight state transitions prior to the changes and the ten state transitions after the changes which simulated dopamine receptor blockade. Figure 5.12A shows all four simulations in one graph, but due to the different scales it is difficult to see the effects of the lower levels of dopamine receptor antagonist, and so Simulations 2 and 3 are plotted separately In Figure 5.12B and compared to the control, Simulation 1.

The results in Figure 5.12 can be compared directly to those of Wise et al. and McClure et al. in Figure 5.3a and 5.3b, respectively. Figure 5.12 reflects the results of the simulations of low concentrations of dopamine receptor antagonism by McClure and colleagues, where a bias of 0.4 (green line) produced a greater increase in running time than a bias of 0.2 (blue line), which, in turn, produced a greater increase in running time than the control, with no change in bias (black line) and no increase. This is most clearly seen in Figure 5.12B.

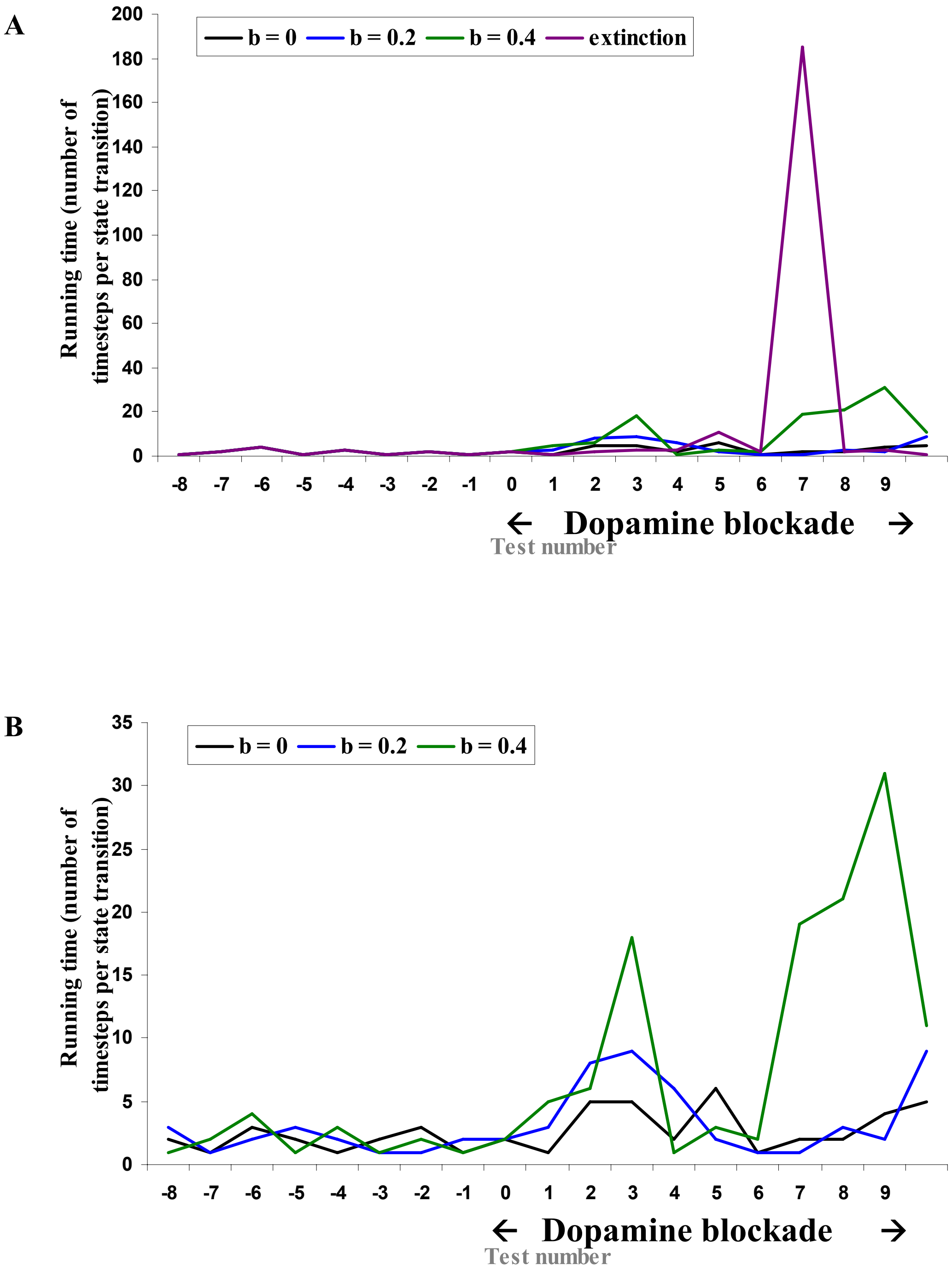


Figure 5.12 shows the effect of dopamine receptor antagonism on running time (number of timesteps for each state transition) at the critical period of the change of bias or reward, between the first and the second set of 1000 state transitions for eighteen state transitions (eight state transitions prior to the changes and ten state transitions after the changes), for: (A) all four simulations and (B) Simulations 2 and 3 for low dopamine receptor antagonism only, compared to a control. Please note the change in scale between (A) and (B).

However, there is a difference between the simulated results of McClure and colleagues and my results, as my simulations do not show a steady increase in running time; rather there are dips seen that return to baseline levels between the periods of increased running speed. No mention was made in the McClure paper of the exact periods of time between their test numbers, but whatever periods of time I used; I still did not obtain the steady increase seen in the simulations of McClure and colleagues. For example, Figure 5.13 shows the same results of the same simulations, for fifty state transitions before and fifty state transitions after the change of bias or reward. As in Figure 5.12B, Figure 5.13B shows that the larger of the two biases is the first to affect running speed, but clearly identifies the dips seen in running speed between the increases. The effects of biases of 0.2 and 0.4 are evident by state transitions 3 and 4 after dopamine blockade, when running speed per state transition increases to 9 and 18, respectively, but overall the effect of increased running time is seen earlier for a change in bias to 0.4, than for a change in bias to 0.2. This is in accordance with the animal experiments of Wise et al. and the simulations of McClure et al.

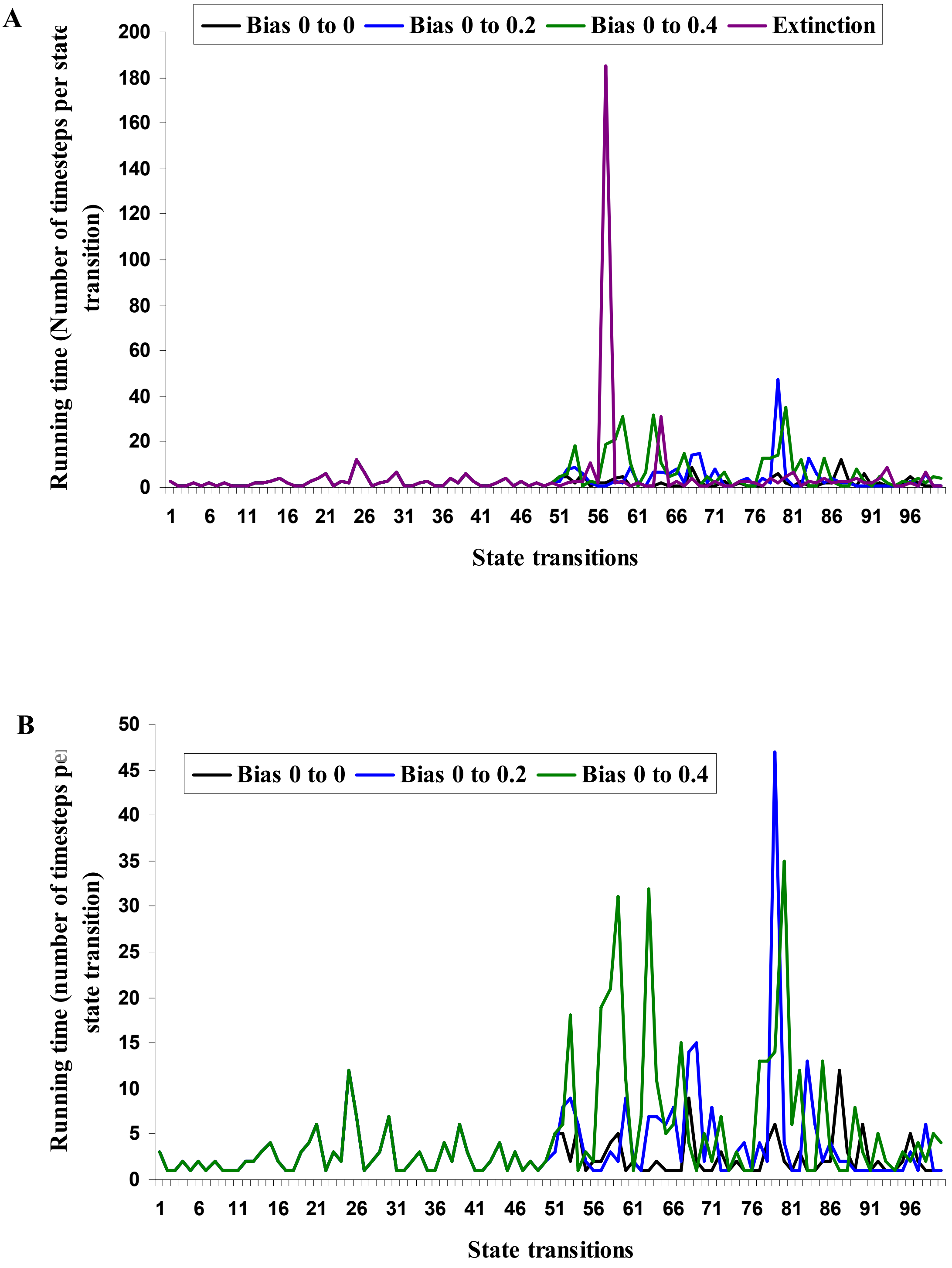


Figure 5.13 shows the effect of dopamine receptor antagonism on running time (number of timesteps for each state transition) for fifty state transitions before and fifty state transitions after the change of bias or reward (the critical period of the change of bias or reward is at state transition 50) for: (A) all four simulations and (B) Simulations 2 and 3 for low dopamine receptor antagonism only, compared to a control. Please note the change in scale between (A) and (B).

(i) A Comparison between High and Low Dopamine Receptor Antagonism

It is useful to compare the results for low dopamine receptor antagonism in Figure 5.13A to the results in Results Section A above, of a higher level of dopamine receptor antagonism, with a change in bias from zero to 1 (Figure 5.14, light blue line). Here the larger increase to the bias results in a much higher running speed, which is seen immediately when the bias changes, with the number of timesteps increasing from an average of 2 per state transition and peaking at 208, 404, 568 and 1591 timesteps for the first, fifth, fourteenth and twenty-seventh state transition following dopamine blockade, respectively. The effects of the lower levels of dopamine blockade take a little longer and are not so pronounced as with a bias of 1 and can be interpreted as being due to the *indirect* effects of updating the stored values of states, only, as opposed to both the *direct* and *indirect* effects associated with higher dopamine receptor antagonism (McClure et al. 2003).

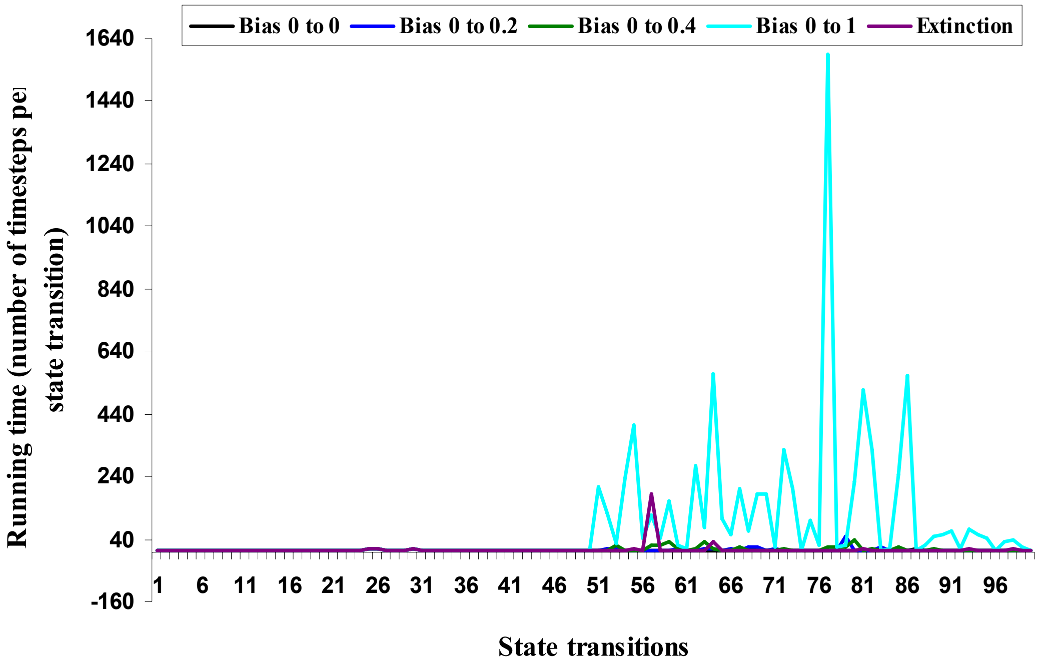


Figure 5.14 shows the effect of adding a higher level of dopamine receptor antagonism (a change in bias from zero to 1, light blue line) to Figure 5.13A. Please note the change in scale.

(ii) The Effect of Extinction on Running Speed

It is also interesting to look at the effect of extinction on running speed, which is only implied in McClure et al. and not shown in detail. In Figures 5.12A and 5.13A the delayed effect of extinction is evident, as running speed is not affected until seven state transitions after the reward is withdrawn, at which time 185 timesteps are recorded for one state transition. One further large running speed of 31 is seen in state transition fourteen, but all other values of running speed remain at control levels.

(iii) The Effects of high and Low Dopamine Receptor Antagonism on TD error

When the effects of the control (*b* = 0), low dopamine receptor antagonism (*b* = 0.2 and 0.4), high dopamine receptor antagonism (*b* = 1) and extinction (*r* = 0), on the *δ(t)* recorded for the goal states only, are plotted for the critical period for their respective changes (ten runs before and ten runs after the changes), the differences at the time of the change become apparent. In particular, it can be seen in Figure 5.15 that the effect on values of *δ(t)* immediately after the change are greatest for extinction, which results in a negative TD error of -0.94 (almost a maximum value of -1). Second, is high dopamine receptor antagonism with a TD error of -0.39, followed by low dopamine receptor antagonism with a bias of 0.4 and a TD error of -0.12, and finally, the lowest dopamine receptor antagonism tested with a bias of 0.2 and a TD error of -0.03, which was well within the range of values of plus or minus 0.09 recorded for the control condition.

The delayed effects of lower dopamine receptor antagonism compared to higher dopamine receptor antagonism are also evident in Figure 5.15. With a bias of 1, the maximum negative TD error is recorded in Run 1, immediately following the change, while the maximum negative TD errors for biases 0.2 and 0.4 are not recorded until runs 3 and 6, respectively. This is in line with the animal experiments of Wise et al. and supports the claims of McClure et al. that lower levels of dopamine receptor antagonism have a delayed effect on running speed, as opposed to the immediate effects seen with higher levels of antagonism. In my simulations, plotting the TD error for each run of the ten runs prior to and after the critical period of change provides additional evidence for the difference between high and low dopamine receptor antagonism. With the smaller TD errors seen, when *b* = 0.2 and *b* = 0.4, there will be less of a *direct* effect on the probability of action selection (Equation 5.2) and more reliance on an *indirect* build up through relearning of the value weights (Equation 5.3).

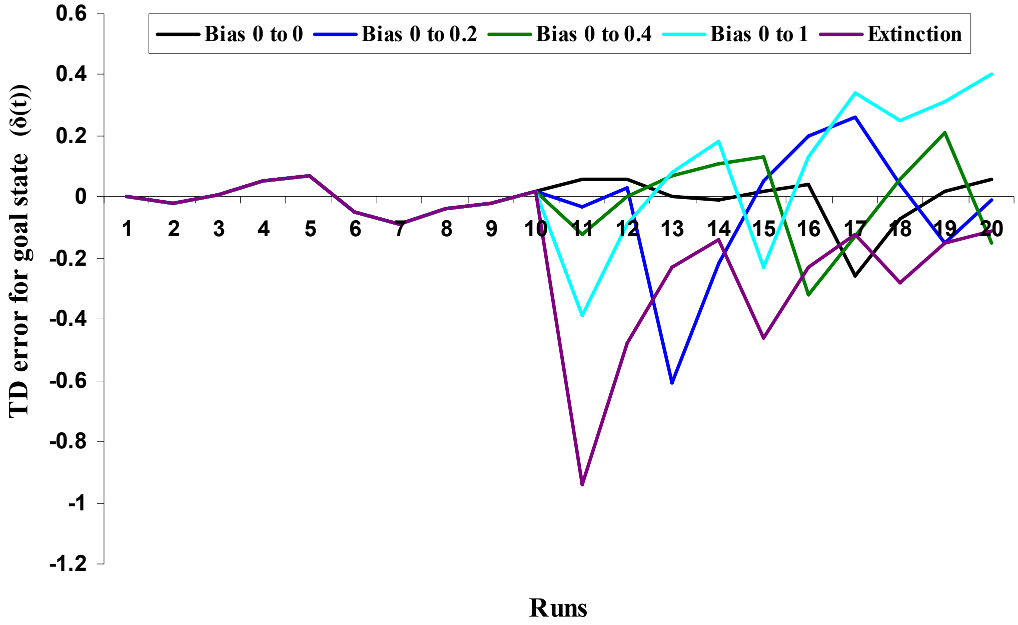


Figure 5.15. Plotting the TD error recorded for the goal state 10 runs before and 10 runs after any changes shows the differences between the control (*b* = 0), low dopamine receptor antagonism (*b* = 0.2 and 0.4), high dopamine receptor antagonism (*b* = 1) and extinction (*r* = 0) at the time of the change, and the delayed effects of lower dopamine receptor antagonism compared to higher dopamine receptor antagonism.

(iv) The Effects of High and Low Dopamine Receptor Antagonism on Values of States

Finally, by looking at the values of just the goal states for 1000 state transitions prior to, and following the critical period, the effect of low or high dopamine receptor antagonism, or extinction on the values of states in general can be ascertained (with the discounting factor, gamma, at 0.9, a preceding state will only achieve 90% of the total reward value from the state to its right). It can be seen in Figure 5.16 that high dopamine receptor antagonism, with a bias of 1 has the greatest effect on the values of states, with a change from a maximum of +1.99 to minimum of -8.31. For lower dopamine receptor antagonism, a bias of 0.4 produces the next greatest effect of the values of states, with a drop in range from +1.99 to -2.23, and a bias of 0.2 follows, with a drop in range from +1.99 to zero.

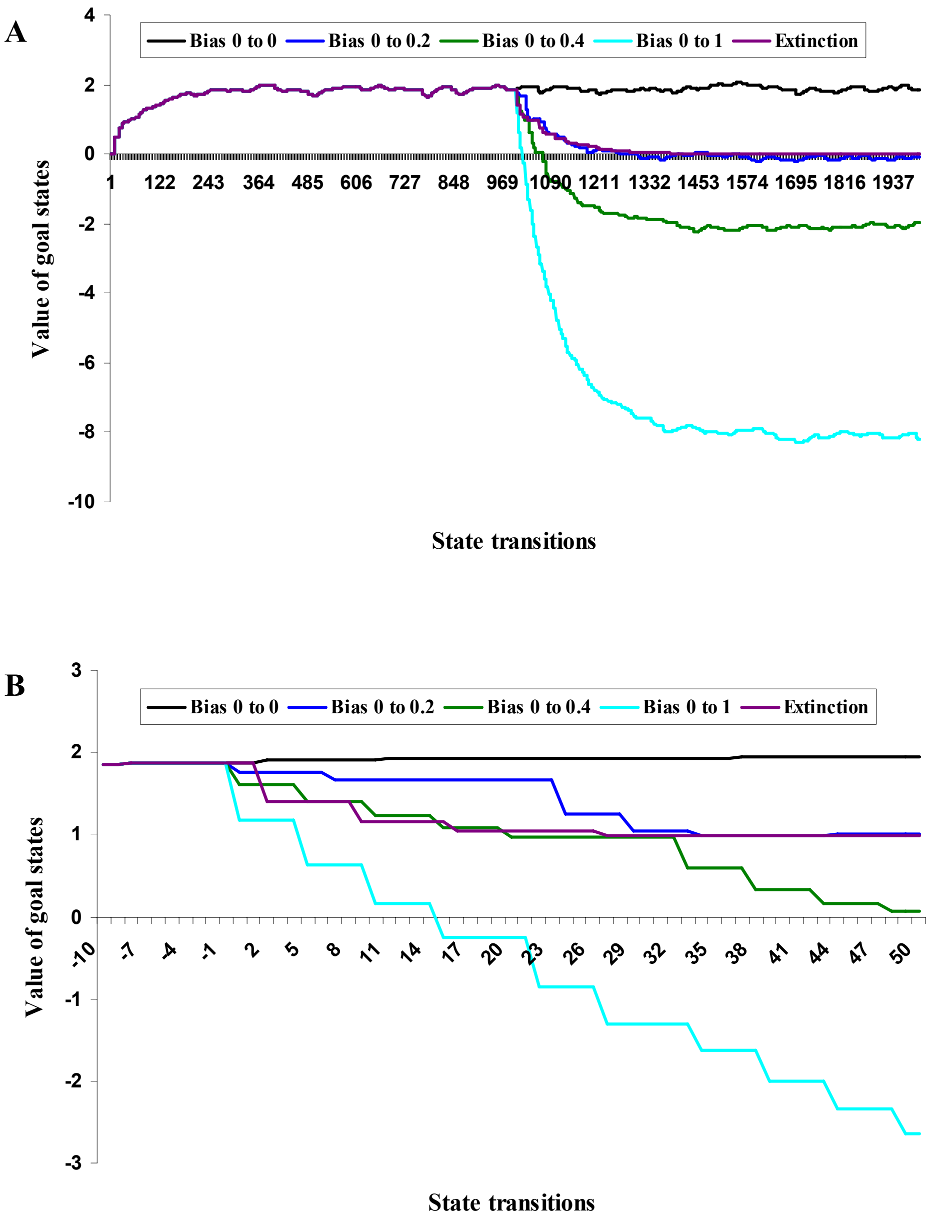


Figure 5.16. Plotting the value of the goal states only (A) 1000 state transitions prior to, and 1000 state transitions after the critical period, and (B) ten state transitions prior to, and 50 state transitions after the critical period, shows the differences between the control (b = 0), low dopamine receptor antagonism (b = 0.4), high dopamine receptor antagonism (b = 1) and extinction (r = 0), and the similarities between low dopamine receptor antagonism (b = 0.2) and extinction (r = 0) following the changes.

A bias of 0.2 gives a most interesting result, which is similar to the drop seen with extinction, where all values of states progressively fall to zero, following withdrawal of the reward. Extinction arises through the withdrawal of a reward and involves a gradual unlearning of the values of states so that no reward is predicted in the future. This is achieved through negative TD errors, signalling that things are worse than expected, where a proportion of this TD error, subject to the learning rate, is used to reduce the values of states (Equation 5.3). With extinction it takes 523 runs for the values of the goal state to steadily fall to zero and then remain at that level, but with a bias of 0.2, values first fall to zero after approximately 250 runs, after which time they hover between a range of +0.04 and -0.23. The difference between a bias of 0.2 and extinction in the early stages is seen more clearly in Figure 5.16B. Nevertheless, despite the differences between a bias of 0.2 and extinction, reflected in Figure 5.16B, I have demonstrated in my simulations that a lower dopamine receptor antagonism has a similar effect on the values of states as extinction. This is in line with the claims of McClure et al.

Summary of Results

A. Simulation of Ikemoto and Panksepp (1996): High Dopamine Receptor Antagonism

I replicated the simulations of McClure et al. (2003) and modelled the results of animal experiments by Ikemoto and Panksepp (1996) of the effect of dopamine receptor antagonism, which severely disrupted the approach (wanting) to a reward in a maze. The reduction in running speed in the maze I achieved of almost 80% (Figure 5.6) obtained with a bias (b = 1), is comparable to the 75% reduction achieved by McClure and colleagues (Figure 5.2b), although, both sets of results are slightly higher than the original animal experiments by Ikemoto and Panskepp (1996), who achieved around a 60% reduction using *cis*-flupentixol in the nucleus accumbens and about 70% reduction in the ventral tegmental area (Figure 5.1b).

While simulating Ikemoto and Panksepp, I was able to address a number of research questions which were not addressed in the original paper. I successfully demonstrated both the *direct* effect of dopamine on action selection, by showing the resulting impact of the lower TD error on the sigmoid decision curve; and the *indirect* effect of dopamine, from its role in learning the estimated values that underlay the actions, on the update of the values of states in the maze. My simulations looked at the exponential decrease in running speed in a maze with increasing levels of bias (where b ranged from 1 to 1.55). I also found that it would be possible to model dopamine receptor antagonism using the scaling constant, m, in Equation 5.2, and showed the effect of increasing values of m on the sigmoid decision curve.

B. Simulation of Wise et al. (1978): Low Dopamine Receptor Antagonism

I also replicated the simulations of McClure and colleagues relating to the animal experiments by Wise et al. (1978) using lower doses of dopamine receptor antagonism. I had a similar pattern of results to the animal experiments and to the simulations of McClure and colleagues, where, unlike with a higher level of dopamine receptor antagonism, a change in running speed was not seen immediately after drug delivery (modelled by an increase in bias), but emerged through repeated exposure, and was less marked, the lower the bias. My results were comparable to those of McClure et al., where a bias of 0.4 produced a greater increase in running time than a bias of 0.2, which in turn, produced a greater increase than the control, with no change in bias and no increase.

While I did not obtain the smoother, steadier increase seen in McClure et al., investigation of other parameters of the model not addressed in the original paper provided evidence for a greater reliance of lower dopamine receptor antagonism on the *indirect* effect on the update of the values of states in the maze, than on the *direct* effect of action selection seen for higher levels of dopamine receptor antagonism. In particular, my comparison between the effects of higher and lower dopamine receptor antagonism showed that the effects of the lower levels of dopamine blockade took a little longer to develop and were not so pronounced as with a high level of bias. In addition, I showed how lower levels of dopamine receptor antagonism produced smaller, delayed TD errors, which, in turn, amounted to less of a reliance on the *direct* effects of action selection on running speed, and more reliance on the *indirect* build up of the new values for states through relearning of the value weights. Finally, I showed that lower dopamine receptor antagonism, particularly with a bias of 0.2, resulted in a similar pattern for values of states in the maze as extinction, which was in line with the claims of McClure and colleagues, and suggested a similar pattern of unlearning.

5.2.3 Discussion

My simulations have allowed me to explore the changing parameters of the computational substrate for incentive salience by McClure and colleagues, and to answer a number of interesting research questions that warranted further investigation. I have provided additional evidence of the difference between the high and low concentrations of dopamine receptor antagonism in animal studies and given an interpretation of the dual function of dopamine as a learning signal and in action selection. My simulations provide additional weight to the claims of McClure and colleagues that their single model can capture the ideas of dopamine (i) as a reward prediction error and (ii) as a purveyor of incentive salience, by uniting psychological and formal computational theories and interpreting reward prediction errors as incentive salience.

The differences in the patterns of behaviour seen in animal experiments resulting from high and low dopamine receptor antagonism, reflected in the timing of the changes to running speed, revealed the dual function of dopamine, as a learning signal and in the bias of action selection. The same dopamine reward prediction error was seen to be acting in two different parts of the model: D*irect* effects of dopamine were seen on action selection (the actor) and *indirect* effects from its role in learning the estimated values that underlay the actions (the critic). The immediate effect of the reduction in running speed seen with high dopamine receptor antagonism in Ikemoto and Panksepp (1996) was shown to result from both the *direct* and *indirect* effects of dopamine, while the delayed, progressive effect of low concentrations of dopamine receptor antagonism in Wise et al. (1978) arose mainly as a result of the *indirect* effects of dopamine, through the slow unlearning of the value estimates, characteristic of the effects of experience-dependent extinction.

Across all of the simulations, an increase in dopamine receptor antagonism, modelled by increasing the parameter, b, in Equations 5.2 and 5.3, produced an increase in the time taken to traverse a maze, with a corresponding decrease in running speed through the maze. This was an example of a robust effect of a computer simulation encompassing animal experiments on Reinforcement Learning Theory.

My simulations for the higher level of dopamine receptor antagonism show a reduction in running speed of nearly 80% which corresponds well with the 75% reduction achieved by McClure and colleagues. However, both sets of results are slightly higher than the original animal experiments by Ikemoto and Panskepp (1996), who achieved around a 60% reduction using *cis*-flupentixol in the nucleus accumbens and about 70% reduction in the ventral tegmental area. It is important to note that the figures for running speeds are parameter dependent and changes to the scaling constant, *m*, could give a better fit to the experimental data.

In addition, my simulations of low dopamine receptor antagonism did not achieve the smoother, steadier increase in running time seen in McClure et al., but this was possibly due to my interpretation of the methodology from the journal paper. The original article gave no specific instruction on how this was to be modelled and so I had to make some assumptions, which are detailed in my Methods Section. However, as I am not limited to the space of a journal article, my additional investigations of the other parameters of the model not addressed in the original paper all pointed to a greater reliance of lower dopamine receptor antagonism on the *indirect* effect on the update of the values of states in the maze, than on the *direct* effect of action selection seen for higher levels of dopamine receptor antagonism, thus demonstrating that my model had captured the effects posited by McClure and colleagues.

Furthermore, the success of this method of modelling has allowed me to build upon these results and adapt this approach to investigate a new topic of the relationship between TD learning and uncertainty coding in dopamine neuron firing, in a novel way. This research is detailed in Chapter 6 and an early version has been published in the proceedings of XXV111 Annual Conference of the Cognitive Science Society (Appendix III), with two further poster presentations at the International Conference on Cognitive and Neural Systems, Boston, Massachusetts, USA, May 17-20, 2006, and the International Conference on Schizophrenia Research, Colorado Springs, Colorado, March 28 – April 1 2007.

In their concluding remarks, McClure and colleagues referred to a hypothesis for incentive salience by Ikemoto and Panksepp (1999), which better accounts for the results of Ikemoto and Panksepp (1996) than the original incentive salience hypothesis of Berridge and Robinson (1998); where dopamine receptor antagonism affects running speed in the maze but not the consumption of the reward. The later hypothesis suggests that dopamine may underlie appetitive approach behaviours but not consummatory behaviours such as licking. This theory is in line with other areas of research that suggest that dopamine in the nucleus accumbens/ventral striatum is important for responding to conditioned stimuli and stimuli that are spatially and temporally distant (distal) rather than proximal to the organism (e.g., Daw, Niv & Dayan 2005; Maffii 1959; Salamone, Cousins & Snyder 1997; Smith, Becker & Kapur 2005; Yin, Knowlton & Balleine 2004). These results pose a problem for the basic TD framework, which does not distinguish between the two motor actions of running and licking, and these ideas are developed further in Chapter 7, where I discuss the limitations of TD learning.

This computational substrate for incentive salience by McClure and colleagues has successfully bridged the gap between conflicting theories of dopamine function, uniting formal computational and psychological theories by interpreting expected future reward as incentive salience. Uniting TD learning and incentive salience has permitted the separation of appetitive and consummatory behaviours that cannot be achieved by TD alone.

5.3 Chapter Conclusions

The main focus of this thesis, so far, has been on modelling dopamine function using TD, and I build upon this further in Chapter 6, with an extended version of the model detailed in this chapter in an attempt to answer one of the key questions in the current debate over whether or not dopamine encodes uncertainty. The models in Chapters 5 and 6 are not specific to schizophrenia, the initial focus of this thesis, but they are models of the specific firing patterns of dopamine, a possible mechanism in the midbrain and cortex for the symptoms and cognitive deficits of schizophrenia (Chapter 2.2).

There are drawbacks to using a TD approach and I highlight some of these in Chapter 7, where I also introduce the concept of model based learning (Sutton & Barto 1998) as an alternative to TD. In particular, I investigate models by Smith, Li, Becker & Kapur (2004; 2006) that, like McClure et al. (2003), use the concept of incentive salience and expected future reward to account for behaviour in a Reinforcement Learning paradigm, but unlike McClure and colleagues these model based accounts can provide an explanation for the effects of dopamine manipulation on distal rather than proximal reward.

Chapter 6

An Analysis of the Relationship between Temporal Difference Learning and Uncertainty Coding in a Computational Model of Dopaminergic Signaling

In Chapter 4, I described the function of dopamine as a reward prediction error in animal studies; the role of the basal ganglia in the production of that signal; temporal difference (TD) as an effective method of modelling the dopamine reward prediction error signal; the Incentive Salience Hypothesis and evidence of the dopamine reward prediction error in humans from fMRI studies. Chapter 4 contained the methodology behind my simulations in Chapter 5, and, in particular, a model by McClure, Daw and Montague (2003) incorporating the Incentive Salience Hypothesis into an Actor-Critic model of dopamine as a reward prediction error. This model successfully simulated the results of animal experiments by Ikemoto and Panksepp (1996), and Wise et al. (1978). The former study found that dopamine receptor antagonism in rats affected the ability to ascribe incentive salience (want) without affecting the motivation for the reward (like). It is believed that dopamine enables reward-seeking behaviours and does not encode the pleasure associated with reward; instead it assigns incentive salience, which maps 'liked' to 'wanted.' It is therefore important that the ideas of incentive salience be incorporated into future neurocomputational models of the role of dopamine.

The work of Wolfram Schultz and colleagues is central to this thesis (Chapter 4.1.2) and a paper by Fiorillo, Tobler and Schultz (2003) came to my attention that claimed to have found a new role for dopamine, in the coding of uncertainty. This was reflected in the sustained activation recorded from dopamine neurons during the delay period between presentation of a conditioned stimulus and receipt of a reward that appeared to increase with increasing uncertainty. A reply by Niv, Duff and Dayan (2005) questioned this claim, and produced a computational model showing that the 'ramping' effects seen were actually due to due to backpropagating TD prediction errors (Chapter 4.3), and not to uncertainty. Fiorillo replied to these counter claims (Fiorillo, Tobler & Schultz 2005) but their arguments did not appear convincing to me. As I already had a computer program capable of modelling dopamine neuron firing, I helped to modify that program to look at the relationship between TD learning and uncertainty. It soon became apparent that I could replicate the effects highlighted by both Fiorillo and Niv, making it possible to investigate the claims of both parties in more detail.

It was necessary for some alterations to be made to the model described in Chapter 5.1.1. Firstly, the maze only allowed travel in one direction. The unnecessary complications of traveling both towards and away from the goal state were avoided, making it easier to compare the effects of different probabilities of reward on TD error. Secondly, the length of the maze was increased from five to nine states, which were deemed an optimal number for the purposes of this research. Finally, in order to model the breaks between maze runs in real rats, it was necessary to insert an initialisation/satiety state into the maze, between the goal and the start. This had the effect of resetting the value of start State 0 to zero, acting as a 'resting' state and ensuring that the 'rat' was always surprised when starting the maze. Without this additional state, the simulated rat learnt the value of the start state, and in effect, there was no conditioned stimulus.

My investigations focused on the key to this debate, namely, how frequently sustained activation or ramping occurs in individual trials, rather than averaging over trials. If there is more sustained activation when probability, p = 0.5 (maximum uncertainty) than when p = 0.25 and 0.75, then it could be said that dopamine is encoding uncertainty in the delay period as a product of TD. It was also important to distinguish between what constitutes sustained activation (Fiorillo et al 2003; 2005) and ramping Niv et al. (2005). My efforts resulted in an alternate model supporting the claims by Niv et al., published in the proceedings of XXV111 Annual Conference of the Cognitive Science Society, Vancouver, Canada, 26-29 July 2006, pp 2263-2268, Lawrence Erlbaum Associates (See Appendix IV). In addition, the research gave rise to two further poster presentations at the International Conference on Cognitive and Neural Systems, Boston, Massachusetts, USA, May 17-20, 2006, and the International Conference on Schizophrenia Research, Colorado Springs, Colorado, March 28 – April 1 2007.

6.1 Abstract

Does dopamine code for uncertainty or is the sustained activation recorded from dopamine neurons a result of Temporal Difference (TD) backpropagating errors? An answer to this question could result in a better understanding of the nature of dopamine signaling, with implications for the psychopathology of cognitive disorders, like Schizophrenia, for which dopamine is commonly regarded as having a primary role. The key to this debate appears to be whether or not sustained activation or ramping occurs during the delay period between presentation of a conditioned stimulus and receipt of a reward, and, if so, whether this increases with increasing uncertainty. A computer simulation of uncertainty incorporating TD Learning and an Actor-Critic architecture successfully modelled a Reinforcement Learning paradigm and the resulting sustained activation in the delay period as demonstrated in single dopamine neuron recordings. Analysis of single trials in our simulations allowed us to make predictions about sustained activation and ramping during the delay period in single trials in vivo. This alternate model showed that while both sustained activation and ramping are common in single trials, both increased with increasing reward probability to a maximum at a probability of 0.9. As there is maximum uncertainty at a probability of 0.5, our TD simulations demonstrate that neither sustained activation nor ramping during the delay period appear to be encoding uncertainty. Our predictions can be tested and verified with behavioural data.

6.2 Background

It has been posited that dopamine codes for uncertainty (Fiorillo, Tobler & Schultz, 2003) as observations of single cell recordings have shown that sustained activity of dopamine neurons precedes uncertain rewards. Fiorillo and colleagues recorded the activity of dopamine neurons in two primates during a delay paradigm of classical conditioning to receive a fixed juice reward, while manipulating the probability of receipt of the reward. Two related but distinct parameters of reward were identified from the activation produced, after learning had taken place, which were found to occur independently within a single population of dopamine neurons: (i) A phasic burst of activity, or reward prediction error, at the time of the expected reward that increased as reward probability decreased (Figure 6.1A). This was in accordance with previous literature, and reinforced the position that midbrain and striatal dopamine systems calculate probabilities of future reward in both learning and decision making tasks (Montague, Dayan & Sejnowski, 1996; Schultz, Dayan & Montague, 1997; Schultz, 1998; Waelti, Dickinson & Schultz, 2001; McClure, Daw & Montague, 2003; Montague et al., 2004). (ii) In addition, Fiorillo and colleagues identified a new slower, sustained activity above baseline lasting from presentation of a conditioned stimulus (CS) to the expected time of a reward, which developed with increasing levels of uncertainty (Figure 6.1B and 1C). This was found to be related to motivationally relevant stimuli and varied with reward magnitude.

In the presence of uncertainty, the sustained activation began on presentation of a CS and increased in strength until a reward was due, at which point the activation ceased (Figure 6.1C, where probability = 0.25, 0.5 and 0.75). This activation was greatest when uncertainty of reward was at a maximum, i.e., when the reward was received on only 50% of occasions and $p = 0.5$. Sustained activation was also seen at lower values of uncertainty, for example, when p was 0.25 and 0.75, but to a lesser extent. No sustained activation was seen when the outcome was certain at probabilities of either zero or 1, suggesting that the sustained activation coded for uncertainty. Out of a total of 188 dopamine neurons they found 29% showed significant increases in activity before potential reward at $p = 0.5$ (3% showed decreases), while only 9% showed increases at $p = 1$.

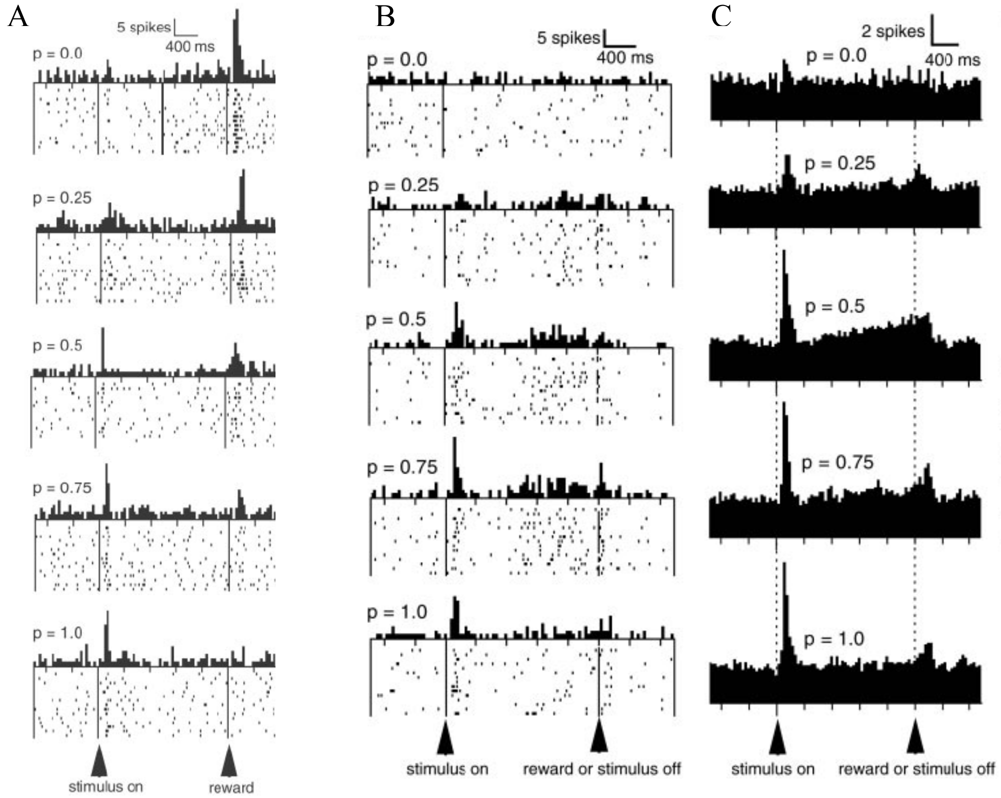


Figure 6.1. Sustained activation of dopamine neurons with uncertainty taken from Fiorillo et al. (2003). (A) Rasters and histograms of single cell activity. Rewarded trials only shown at intermediate probabilities. (Their Figure 2A) (B) Rasters and histograms of single cell activity. Both rewarded and unrewarded trials shown. (Their Figure 3A) (C) Population histograms of B. (Their Figure 3B)

It was concluded that the activity of dopamine neurons may carry information about two distinct parameters of reward: (i) The phasic response at the expected time of reward as a teaching signal for Reinforcement Learning in accordance with the principles of Rescorla and Wagner (1972), and (ii) the sustained activation as information about uncertainty, facilitating attention, and therefore learning, which is more in line with the Pearce-Hall Theory (Pearce & Hall, 1980).

However, the view that the sustained activity encodes uncertainty is controversial as Niv, Duff and Dayan (2005) have suggested that the sustained activation, or what they term the *ramping* effect in the delay period, is due to backpropagating Temporal Difference (TD) prediction errors, and not to uncertainty. TD Learning (Sutton, 1988; Sutton & Barto, 1998), a form of Reinforcement Learning, and a time dependent variant of the Rescorla-Wagner model (Rescorla & Wagner, 1980), provides an explicit method of modeling and quantifying the dopamine reward prediction error (Hollerman & Schultz, 1998; Schultz et al., 1997). Specifically, Niv and colleagues claim it is the asymmetric coding of reward prediction errors that gives rise to the effects seen in time, over consecutive CS presentations, due to a low baseline rate of activity in dopamine neurons. Firing rates corresponding to positive prediction errors typically rise to about 270% above baseline, while those for negative errors only fall to approximately 55% below baseline (Fiorillo et al. 2003). During uncertainty, these asymmetrical positive and negative errors, when summed, will not cancel each other out, as predicted by the TD algorithm, even after extensive training periods. The overall effect, as seen in Fiorillo et al., will be of (i) a positive response across trials at the expected time of reward, and (ii) a ramping effect from presentation of the CS to the expected time of reward, described by Fiorillo and colleagues as sustained activation.

Using TD, Niv and colleagues successfully modelled both effects identified by Fiorillo et al. (2003) during uncertainty. They explained that the ramping was an artifact of the asymmetric coding of reward prediction errors and does not directly encode uncertainty. Furthermore, they claimed that the resulting effects arose as a result of averaging across multiple trials and were not a within trial phenomenon. They also showed that the shape of the ramp depended on the learning rate, and that the difference in the steepness of the ramp between delay and trace conditioning could be accounted for by the low learning rates associated with trace conditioning, resulting in smaller or even negligible ramps.

Fiorillo and colleagues defended their original claim that the sustained activity encoded uncertainty about reward (Fiorillo, Tobler & Schultz, 2005), and two of the five points they raised were of particular interest to the present study. Firstly, they referred to the difficulty in determining what a single trial increase in the delay period activity should look like. They specified that it was difficult to equate positive spike trains with the positive and negative reward prediction errors generated from the TD algorithm, as any spike could represent a backpropagating error, while any absence of spikes, a negative error. However, based on averaged firing rates over tens of milliseconds, they maintained that firing rate appeared to increase during the delay period over single trials and gave two examples in support of their argument. Their argument would suggest that if sustained activation or ramping is common in single trials, the extent of the validity of TD as a model of dopamine as a reward prediction error is called into question as it does not account for this activity. However, if such instances in single trials are rare then claims by Fiorillo et al. that the sustained activation encodes uncertainty are unfounded.

Secondly, they suggest that according to TD, activity in the last part of the delay period in a trial should reflect the reward outcome in the previous trial to that same CS. However, they found no evidence in their data of a dependence of neural activity on the outcome of the preceding trial. This also calls into question the extent to which dopamine firing patterns can be represented by TD. We suggest that this is not what is predicted by TD; specifically, it is the history of all preceding trials, and not just the activity of the last trial that determines activity.

6.3 This Study

The key to this debate appears to be how frequently sustained activation or ramping occurs in individual trials. Furthermore, if there is more sustained activation when $p = 0.5$ (maximum uncertainty) than when $p = 0.25$ and 0.75, then it could be said that dopamine is encoding uncertainty in the delay period as a product of TD. It is also important to distinguish between what constitutes sustained activation (Fiorillo et al, 2003; 2005) and ramping (Niv et al., 2005).

The aim of this study was to seek answers to the above questions in order to shed some light on whether or not dopamine encodes uncertainty. To this end I helped to create a novel model of dopamine neuron firing, incorporating TD backpropagating errors. Inspired by McClure, Daw and Montague (2003) and an alternative to Niv et al. (2005), the model linked the ideas of reward prediction error and incentive salience, and captured the following effects seen in single cell recordings of reinforcement by Fiorillo and colleagues: (a) The phasic activations at the expected time of reward; (b) the sustained increase in activity from the onset of the conditioned stimulus until the expected time of reward; and (c) the sustained activation increasing with increasing reward magnitude. A successful model would enable a single trial analysis; which is the cornerstone of the debate over whether or not the sustained activation or the ramping effect, seen in dopamine neuron firing is encoding uncertainty. Specifically, if it were possible to build the inaccuracies of single cell recordings into a TD model where sustained activation or ramping was seen in single trials and was not just an artefact of averaging reward prediction errors over trials, there would be evidence for dopamine encoding uncertainty and TD would remain a viable model of dopamine neuron firing patterns. Furthermore it would be possible to consider the effects of interrupting the sequence of rewarded and non rewarded runs or trial history.

6.4 Methods

The Maze

A computer simulation was constructed of a ‘rat’ learning to traverse a one-arm maze to receive a reward, using the TD algorithm with *Actor-Critic* architecture. The maze modelled was inspired by McClure, Daw and Montague (2003) linking the ideas of reward prediction error and incentive salience, but contained an additional initialisation/reset state and only allowed travel in one direction. Figure 6.2 shows a typical maze with positions modelled as eight states, starting at State 0 (the CS) and progressing through intermediate states to receive a simulated reward in State 7 (the reward state). In order to model the breaks between maze runs in real rats, it was necessary to insert an initialisation/satiety state (State 8) into the maze, between the goal (State 7) and the start (State 0), where the transition between that state and State 0 remained at zero so that no learning could take place. This had the effect of resetting the value of start State 0 to zero, acting as a ‘resting’ state and ensuring that the ‘rat’ was always surprised when starting the maze. Without this additional state, the simulated rat learnt the value of the start state, and in effect, there was no CS. Intermediate states were added and removed, as required to make mazes of different lengths.

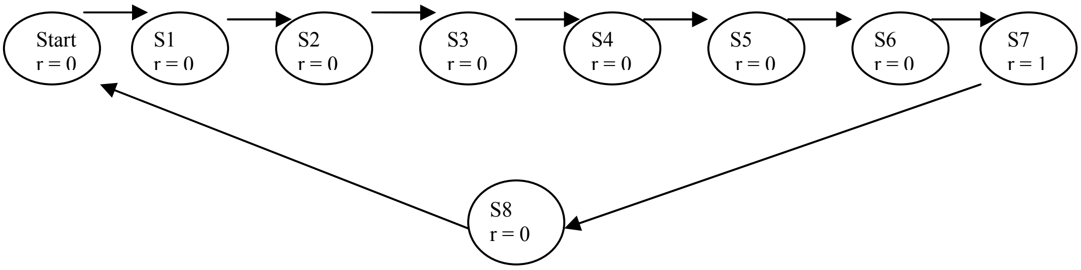


Figure 6.2. Schematic representation of a maze with eight states (S) plus 'satiety' state (S8). A reward (r) of 1 can be provided in S7 only.

The Model

I implemented and modified an existing computer program written by a member of the research team in LISP (Steele 1990), of a maze incorporating the TD algorithm with an Actor-Critic architecture (Chapter 4.3); a form of reinforcement TD learning where a *critic* computed a reward prediction error, which was used by the *actor* to choose those actions that led to reward (McClure, Daw & Montague, 2003; Montague et al., 2004; Sutton & Barto, 1998). The model incorporated the concept of incentive salience as expected future reward, where the dual roles of dopamine: (i) to bias action selection towards situations that predict the best reward; and (ii) as a learning signal used to update the values of states and so create better estimates of future reward, equated with the Incentive Salience Hypothesis and the role of dopamine in attributing and using incentive salience.

The program code can be found in Appendix V and is a modification of that used in my simulation of a computational substrate for incentive salience in Chapter 5.1.1. The original model by McClure, Daw & Montague (2003) is described in detail in Chapter 4.2.

Simulations

Uncertainty – Degree of Probability

The ranges of probabilities used for trials were 0.25, 0.5 (maximum uncertainty), or 0.75. The *δ(t)* values were recorded for each state transition, for a single probability in each trial. Each trial consisted of 1000 runs through the one-way maze, with a step being a transition from one state to the next, and a run being one complete journey through the maze, from start to finish. At the beginning of each trial the values of each state in the maze (*V*) were set to zero. Movement to the next state in the maze was selected according to the effect of TD learning on different probabilities of receiving a reward for each run.

To investigate:

(a) The phasic activations at the expected time of reward

(b) The sustained increase in activity or ramping effect

In keeping with the pharmacokinetics of dopamine, namely the asymmetry in coding of positive and negative errors, any negative prediction errors were scaled by a factor of one sixth, the scaling factor used by Niv et al. (2005), based on data from Fiorillo et al. (2003). The scaled *δ(t)* values were then averaged across 1000 consecutive runs for each state, where $\gamma = 1$ [Chapter 5, Equation 5.1, (cf Chapter 4.2.1, Equation 4.4)] and $\alpha = 0.5$ [Chapter 5, Equation 5.3 (cf Chapter 4.2.2, Equation 4.5)], and the magnitudes of the scaled values compared. This averaging corresponded to the summing of peri-stimulus-time-histograms of activity over different trials and inter-trial averaging used by Fiorillo et al. (2003).

(c) Reward Magnitude

Individual reward magnitudes of 0.5, 1 and 2 were compared at $p = 0.5$ in different trials to see the effect on the sustained activation.

(d) Single Trial Analysis

In order to address two of the points raised in Fiorillo et al. (2005), it was possible to undertake single trial analysis, by recording the scaled prediction errors, *δ(t)*, for each state transition, for several single runs through the maze. One run was analogous to the activity of a single neuron in a single trial over time and was simply a snapshot of the *δ(t)* values for each state. Each state was equivalent to a period of time, usually represented by a number of bins in spike trains. While the model does not represent firing patterns in real time, the juice reward in Fiorillo et al. (2003) was delivered after a two second delay, and so each of the nine states in the maze could be seen to be representing 0.25 seconds.

(e) Trial History

Different series of consecutive runs were examined to see if the activity in the last part of the delay period in a trial should always reflect the reward outcome in the previous trial to that same CS. The effects of interrupting the sequence of rewarded and non rewarded runs in the trial history were investigated.

6.5 Results

An Example of Learning

With a probability of 1 and a maze of eight states plus a ‘satiety’ state, complete learning took place over the first thirty runs. On the first run a large prediction error, *δ(t)*, was recorded at the expected time of the reward (S7-S8), and as runs progressed, this *δ(t)* was transferred back to the CS (S8-S0). When full learning had taken place only the CS elicited a reward prediction error. This effect is demonstrated in Figure 6.3, which shows the *δ(t)* at the expected time of reward beginning at 1 and reducing to zero by run 9 at which point the value of the state is learnt and the reward fully predicted. The *δ(t)* at the CS begins at zero and increases gradually to 1, from run 10 to run 30. An intermediate state transition S3-S4 is included which records the *δ(t)* backpropagated from the reward state by run 5. The error increases until run 8 and then reduces to zero by run 21.

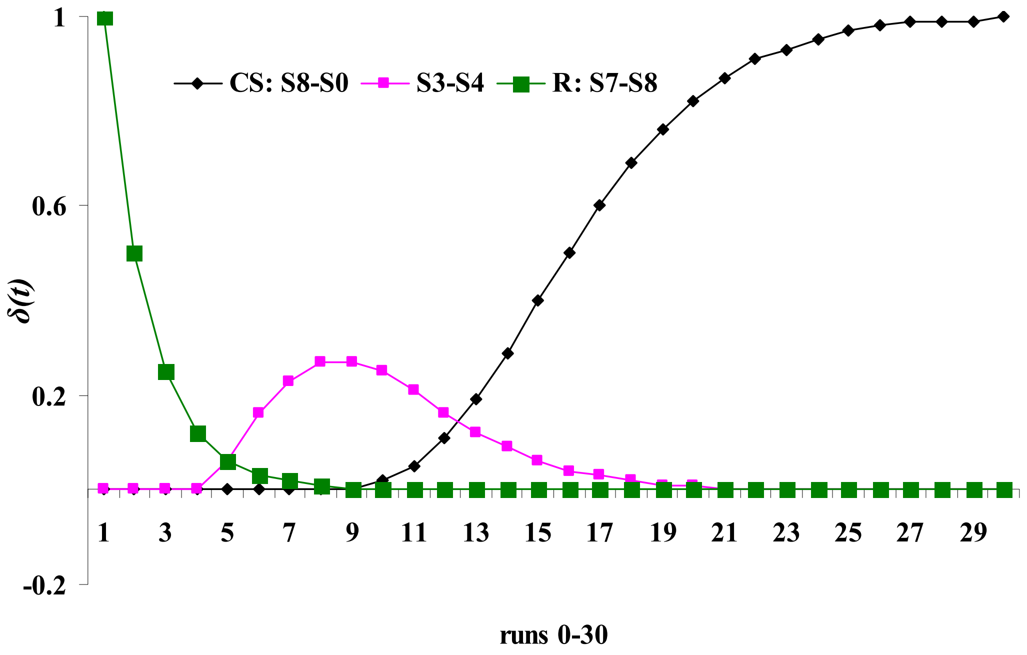


Figure 6.3. A demonstration of learning showing the reward prediction error, *δ(t)*, at the expected time of reward (R: S7-S8) beginning at 1 and reducing to zero by run 9. The *δ(t)* at the CS (S8-S0) begins at zero and increases gradually to 1, from run 10 to run 30. At this point the value of the state is learnt and the reward fully predicted. An intermediate state transition (S3-S4) is included which records the *δ(t)* backpropagated from the reward state by run 5. The error increases until run 8 and then reduces to zero by run 21. (*p* = 1, *r* = 1).

As expected, when uncertainty was introduced the CS (S8-S0) was not fully predictive of the reward, as in the case of *p* = 1. With probabilities (*p*) 0.25, 0.5 and 0.75 all states, and not just the CS, continued to elicit a reward prediction error. Figure 6.4 shows the CS (S8-S0) over 30 runs for the three different probabilities plotted on the same graph as *p* = 1. While the CS at *p* = 1 reached a *δ(t)* of 1, the others averaged over 0.25, 0.5 and 0.75 respectively.

Training was assumed to have occurred to a sufficient level when the CS (S8-S0) prediction error, δ(t), averaged 0.25, 0.5 and 0.75 for p = 0.25, 0.5 and 0.75 respectively. All the following tests with uncertainty were done post training.

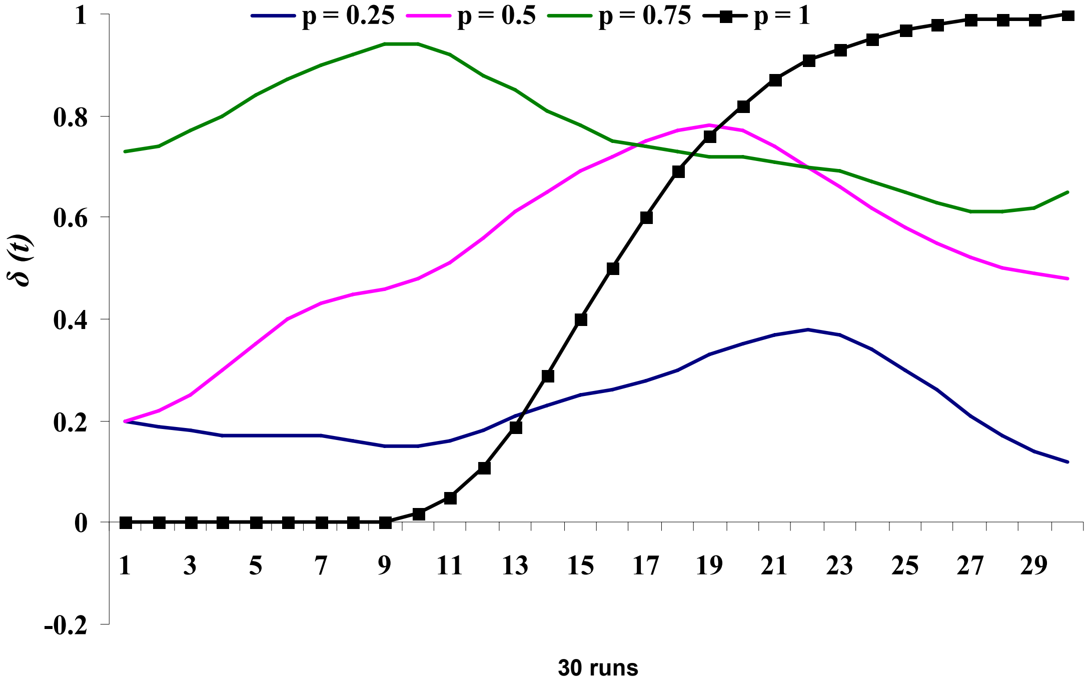


Figure 6.4. With uncertainty (p = 0.25, 0.5 and 0.75) the CS is never fully predictive of the reward as with p = 1. Actual runs 0-30 for p = 1 and runs 36-65 for p = 0.25, 0.5 and 0.75.

Uncertainty – Degree of Probability

Eventually, by chance, actions were selected in trials for the entire range of probabilities, 0.25, 0.5 or 0.75, and progression was made along the maze towards the reward state where the reward (r) of that state, r = 1 was received for the first time. Once a reward has been received, learning could begin to take place. This was achieved by visiting a state (by chance) and comparing the values of that state with the state just visited [Chapter 5, Equation 5.1, (cf Chapter 4.2.1, Equation 4.4)]. The TD algorithm ensures that values of states become consistent between states, and if there was a difference, i.e., a reward prediction error, then the value of that state was updated by α, a fraction of that reward prediction error. On subsequent runs, learning occurred as the value of the reward was propagated backwards, updating earlier states.

As the probability of obtaining a reward increased, from 25% to 50% to 75%, so did the level of phasic activation at the CS (S8-S0) (Figures 6.5 and 6.6), with average *δ(t)* values of 0.25, 0.46 and 0.76 respectively over 1000 runs.

(a) The phasic activations at the expected time of reward

Without scaling the *δ(t)* values recorded for each state transition to compensate for the biologically asymmetric coding of positive and negative prediction errors, no average positive phasic activation was seen at the expected time of reward (Figure 6.5, S7-S8). However, after scaling rewarded and unrewarded *δ(t)* values by a factor of one sixth and averaging *δ(t)* values over consecutive trials, positive phasic activation was seen at the expected time of reward of 0.15, 0.21 and 0.15, for probabilities of 0.25, 0.5 and 0.75 respectively (Figure 6.6, S7-S8). Positive phasic activation was highest for maximum uncertainty (p = 0.5).

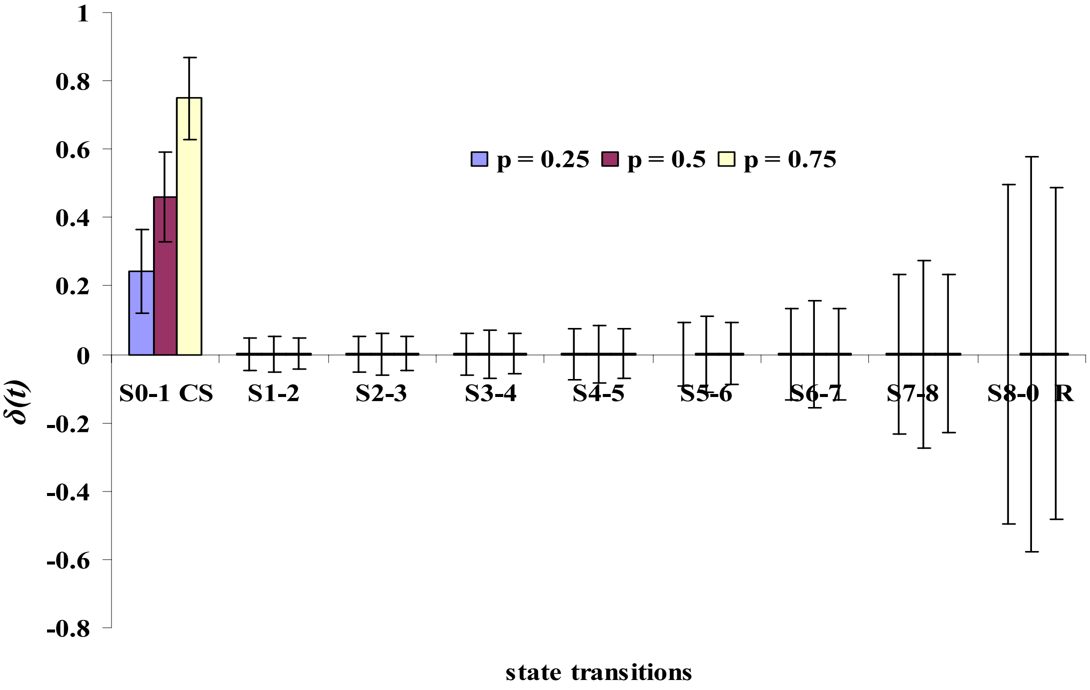


Figure 6.5. Average δ (t) values, with standard deviation error bars, for rewarded and unrewarded runs, **before scaling**, for each state transition over 1000 runs, when p = 0.25, 0.5 and 0.75, α = 0.5. There is no sustained activation or ramping, that is the $\delta(t)$ values on all, except the CS state, are zero.

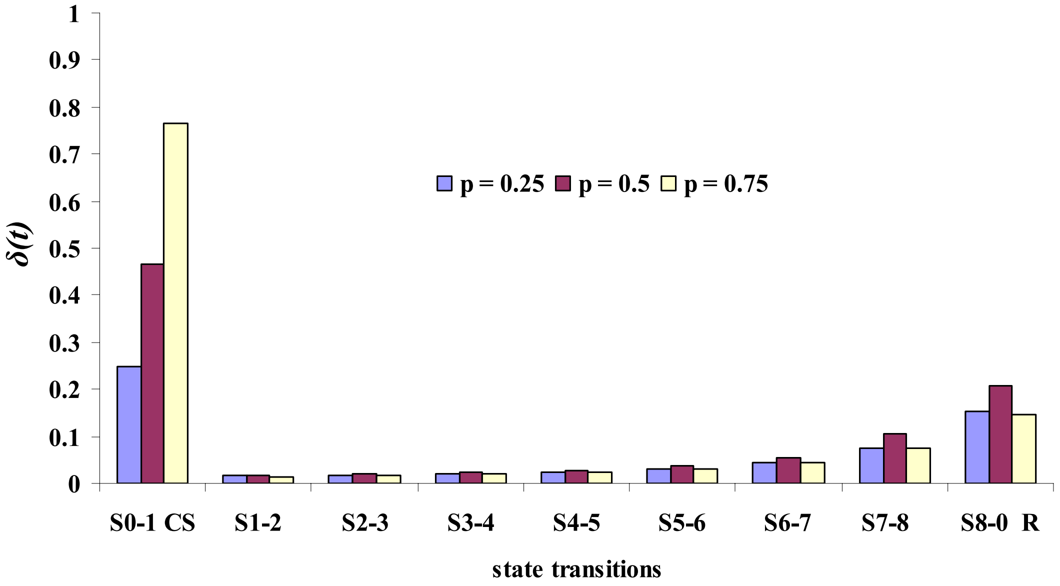


Figure 6.6. Average $\delta(t)$ values, with standard deviation error bars, for rewarded and unrewarded runs, **after scaling**, for each state transition over 1000 runs, when p = 0.25, 0.5 and 0.75, α = 0.5. Sustained activation/*ramping* is seen which is greatest for maximum uncertainty (p = 0.5).

This is contrary to the findings of Fiorillo et al. (2003), who showed that the magnitude of reward responses increased as probability decreased. However, their Figure 2 (Figure 6.1A) showed rewarded runs only and, after removing all the negative *δ(t)* values, our Figure 6.7 shows comparable results to their Figure 2. So it can be said that when rewarded trials only are shown (Fiorillo et al. (2003), their Figure 2) the magnitude of the reward responses increased as probability decreased. My Figure 6.6 is comparable to their Figure 3 (Figure 6.1B and C) which includes both rewarded and unrewarded trials. This highlights the danger of averaging non consecutive runs without considering the reward history (see Trial History below).

In conclusion, uncertainty would appear to be coded in the reward averages of consecutive rewarded and unrewarded trials. This information becomes finely tuned over a period of time and could be used as a training signal by another system, such as a working memory buffer, to utilise the information provided in future trials.

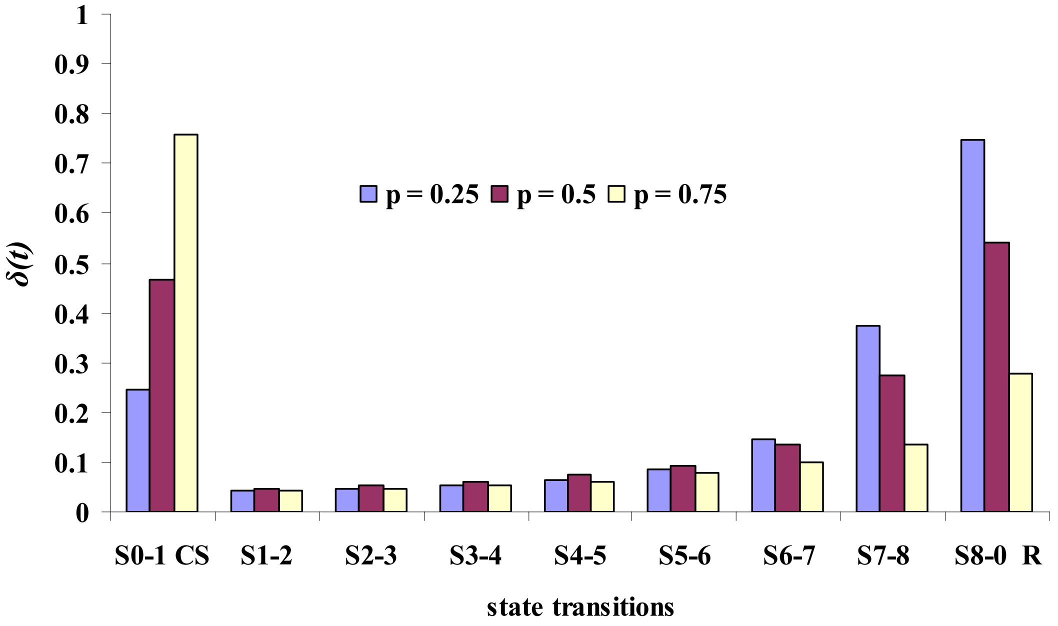


Figure 6.7. Average *δ(t)* values for rewarded runs only, **after scaling**, for each state transition over 1000 runs, when *p* = 0.25, 0.5 and 0.75, *α* = 0.5. Sustained activation/*ramping* is seen which increases as probability decreases.

(b) The sustained increase in activity

Figure 6.5 showed that, in the absence of asymmetric scaling, no ramping effect was seen from plotting the average *δ(t)* values obtained for probabilities of 0.25, 0.5 and 0.75 for each state transition. The symmetrical positive and negative errors effectively cancelled each other out, in accordance with the TD algorithm. However, when the *δ(t)* values were scaled by a factor of one sixth to compensate for the biological asymmetric coding of positive and negative errors, and averaged across consecutive runs, positive *δ(t)* values were seen that corresponded to the sustained activation and ramping effects reported in Fiorillo et al. (2003) and Niv et al. (2005) respectively (Figure 6.6). In accordance with the findings of Fiorillo et al. (Figure 6.1C), the magnitude of the ramping effect was greater for maximum uncertainty ($p = 0.5$) than for the lower uncertainties of $p = 0.25$ and $p = 0.75$, which both had similar values.

(c) Reward magnitude

The value of the reward at $p = 0.5$ was manipulated in three different trials of 30 runs, with rewards given of 0.5, 1 and 2. The size of the reward had an effect on the range of *δ(t)* values available for each state. With a larger reward comes a larger range of possible *δ(t)* values, and, accordingly, larger ramping effects (Figure 6.8). Therefore, the sustained activation increased with increasing reward magnitude, in accordance with Fiorillo et al. (2003).

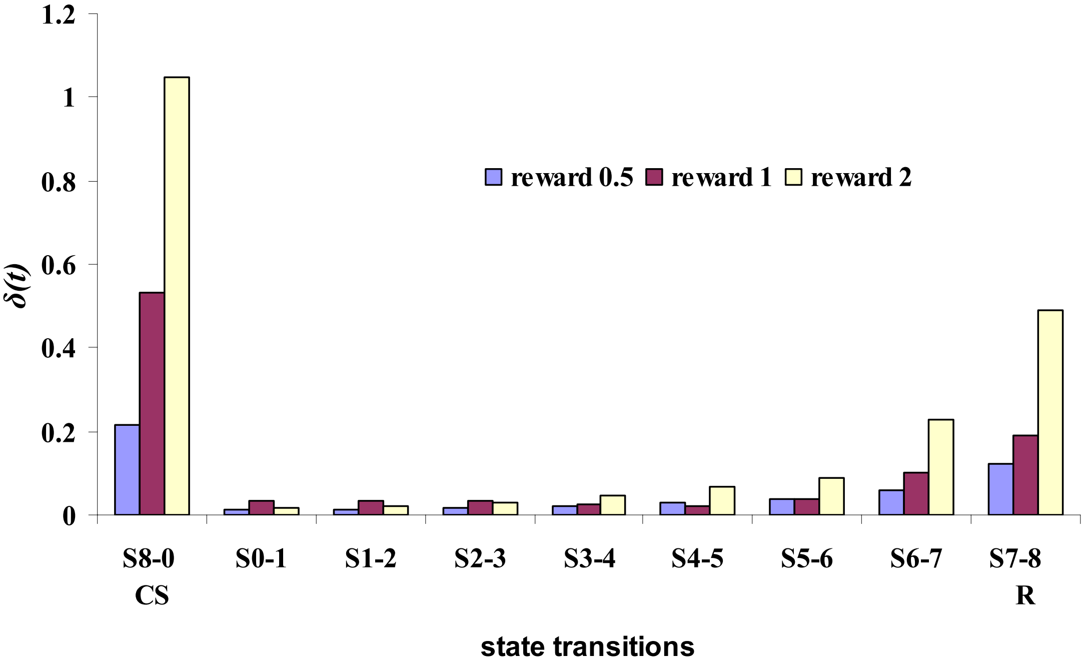


Figure 6.8. Scaled average *δ(t)* values over 30 runs for reward values of 0.5, 1 and 2, $p = 0.5$, $\alpha = 0.5$. The sustained activation/*ramping* increased with increasing reward magnitude.

(d) Single Trial Analysis

Criteria for sustained activation

Fiorillo et al. (2003, page 1900) described the apparent coding of uncertainty as '…a sustained increase in activity that grew from the onset of the conditioned stimulus to the expected time of reward,' where '…the peak of the sustained activation occurs at the potential time of reward, which corresponds to the moment of greatest uncertainty'. The sustained activation was maximal at $p = 0.5$, less pronounced at $p = 0.25$ and 0.75 and absent at $p = 0.0$ and 1. They provided an example of activity in a single dopamine neuron which corresponds to a single trial (Figure 6.9). The horizontal dashed line is my guess at where baseline firing could be (not to scale).

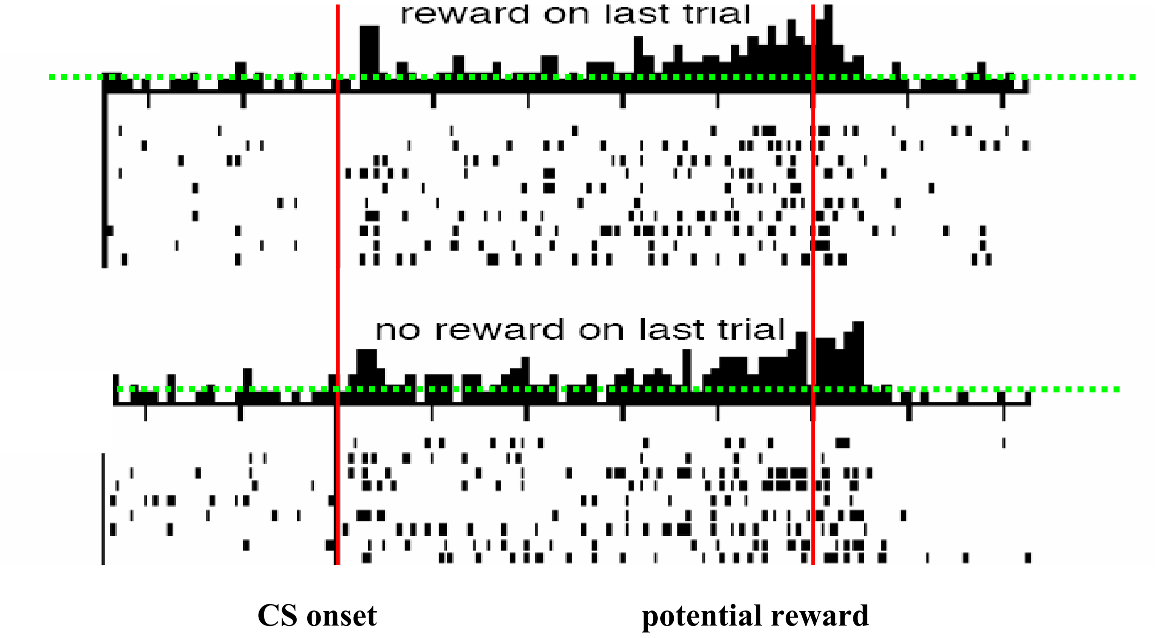


Figure 6.9. Activity in a single dopamine neuron taken from Fiorillo et al. (2005). Dotted horizontal line shows our guess at baseline firing level.

Fiorillo et al. (2005) report that it is difficult to be certain whether or not activity increases on single trials and refer therefore to strong and sustained activation within single trials. Based on the evidence in Figure 6.9, the present simulations were undertaken with three criteria for the sustained activation or ramping effect:

> **Criterion 1 for sustained activation only:** Four of the eight possible transition states should have positive reward prediction errors/activations (not including the CS, transition state S8-S0).
>
> **Criterion 2 for Ramping:** The sum of activation for states S5-6, S6-7 and S7-8 should be greater than for states S2-3, S3-4 and S4-5.
>
> **Criterion 3 for Ramping:** The peak of the sustained activation should occur in states S5-6, 6-7 or S7-8; close to the potential time of reward and the moment of greatest uncertainty.

If all three of the above criteria were satisfied the single trial was deemed to have met with the strict criteria for sustained activation/ramping. Criterion 1 alone was sufficient to identify sustained activation.

An example of sustained activity/ramping in scaled single trials

While this research supports the argument by Niv et al. (2005) that the ramping effect seen during uncertainty, when averaging across multiple trials, is a result of backpropagating TD errors, I show that it is also possible for sustained activation/ramping to occur in single trials in accordance with Fiorillo et al. (2005). Single trials in our simulations are represented by recording the scaled prediction errors, *δ(t)*, for each state transition, for one run through the maze. This run is analogous to the activity of a single neuron in a single trial over time and is simply a snapshot of the *δ(t)* values for each state, which may be either positive or negative with respect to baseline firing, depending on the history of previous runs. They are directly comparable to Figure 6.9 taken from Fiorillo et al. (2005), where baseline firing is defined as zero and corresponds to the dashed horizontal line estimate of baseline that has been added to Figure 6.9.

My simulations allow me to predict that sustained activation/ramping may occur when several rewarded runs are received in succession. In order to demonstrate the methodology I take an extreme case, when p = 0.5 and eight rewarded runs are received in succession by chance. Single runs 81 to 92 are shown in Table 6.1 where the sequence of rewarded (R) or non rewarded (N) runs is ---NRRRRRRRRNNN, where --- represents learning prior to the sequence under scrutiny (---NRNRNN on this occasion). It can be seen from Table 6.1 that five of the twelve runs met all three of our criteria, satisfying the strict criteria for sustained activation/ramping, while ten of the twelve runs met criterion 1 for sustained activation alone.

Table 6.1. How a sample of consecutive runs met with the three criteria for sustained activation/ramping (actual runs 81-92). When criteria are satisfied the results are in bold. See text for definition of criteria and Figure 6.10 for an example of run 83.

Run Number (N-non rewarded R-rewarded)	**Criterion 1 met for Sustained Activation?**	**Criterion 2 met for Ramping?**	**Criterion 3 met for Ramping?**	**Strict criteria met? (criteria 1, 2 and 3 all satisfied)**
81 (N)	2/8 No	No	No	No
82 (R)	3/8 No	**Yes**	**Yes**	No
83 (R)	**4/8 Yes**	**Yes**	**Yes**	**Yes**
84 (R)	**5/8 Yes**	**Yes**	**Yes**	**Yes**
85 (R)	**6/8 Yes**	**Yes**	**Yes**	**Yes**
86 (R)	**5/8 Yes**	**Yes**	**Yes**	**Yes**
87 (R)	**7/8 Yes**	**Yes**	**Yes**	**Yes**
88 (R)	**8/8 Yes**	No	No	No
89 (R)	**8/8 Yes**	No	No	No
90 (N)	**7/8 Yes**	No	No	No
91 (N)	**6/8 Yes**	No	No	No
92 (N)	**5/8 Yes**	No	No	No

As an example of a ramping trial, run 83 from Table 6.1 is presented graphically in Figure 6.10. Run 83 is an example of a single trial that satisfies all three criteria for sustained activation/ramping, and thus conforms to our strict criteria. Firstly, four of the eight possible states, S2-3, S3-4, S6-7 and S7-8 showed positive reward prediction errors/activations (not including the CS: S8-0), satisfying Criterion 1 for sustained activation. Criterion 2 for ramping required the activation for states S5-6, S6-7 and S7-8 to be greater than for states S2-3, S3-4 and S4-5 and in the case of run 83 average values were 0.265 and 0.029 respectively. Finally the peak of the sustained activation occurred in state S7-8, the reward state, satisfying Criterion 3, which required the peak to be in one of states S5-6, 6-7 or S7-8.

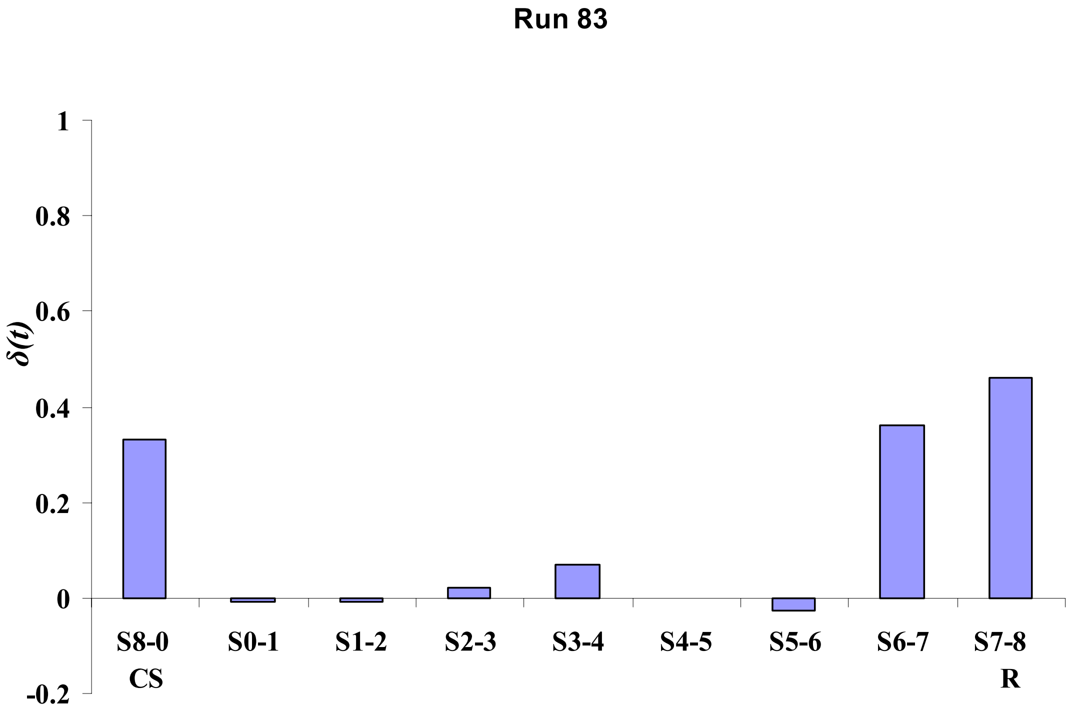


Figure 6.10. Actual run 83 from Table 6.1. An example of a single trial that satisfies the strict criteria for sustained activation/ramping. Criterion 1: Four of the eight possible states showed positive reward prediction errors/activation (not including the CS: S8-0). Criterion 2: The activation for states S5-6, S6-7 and S7-8 was greater than for states S2-3, S3-4 and S4-5. Criterion 3: The peak of the sustained activation occurred in one of states S5-6, 6-7 or S7-8.

Does Dopamine encode Uncertainty?

I have demonstrated above that positive sustained activation can be seen in single trials of dopamine neuron firing at p = 0.5. Evidence of sustained activation/ramping was also evident for lower levels of uncertainty, when p = 0.25 and 0.75, and if it can be shown that sustained activation/ramping is greater and more frequent in single trials when p = 0.5 than for 0.25 and 0.75, as uncertainty increases, then it can be said that dopamine is encoding uncertainty, supporting the claims of Fiorillo et al. (2003; 2005).

I therefore examined 200 consecutive single trials for different probabilities, where p = 0.25, 0.5, 0.75, 0.9, 0.95 and 1, after learning, and applied my criteria to determine whether or not sustained activation or ramping effects were seen. Figure 6.11 shows the number of states showing positive sustained activation for each of 200 runs or single trials with a maximum of 1600 states (200 runs each containing 8 states, not including CS); while Figure 6.12 shows the total number of ramps, according to my criteria, for each probability out of a maximum of 200 runs or single trials. Figures 6.11 and 6.12 show the same trend, where both sustained activation and ramping increase with increasing reward probability to a maximum at p = 0.9, demonstrating that TD error does not appear to be encoding uncertainty in either the sustained activation or the ramping effects.

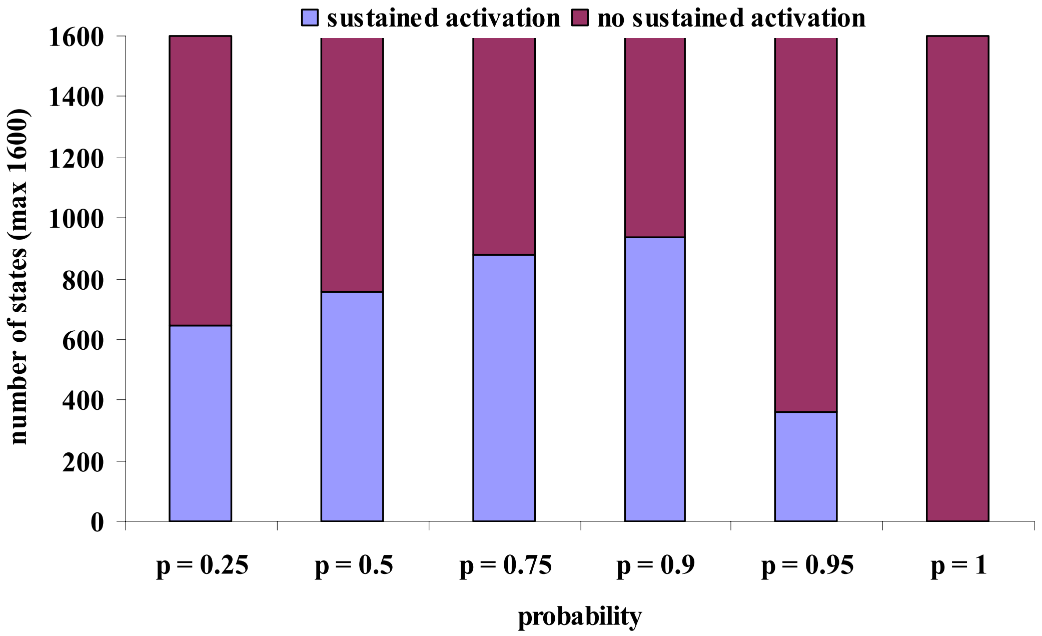


Figure 6.11. Number of states for each probability, showing both positive sustained activation and no sustained activation, out of a maximum of 1600 (i.e., 200 consecutive runs with eight possible states per run).

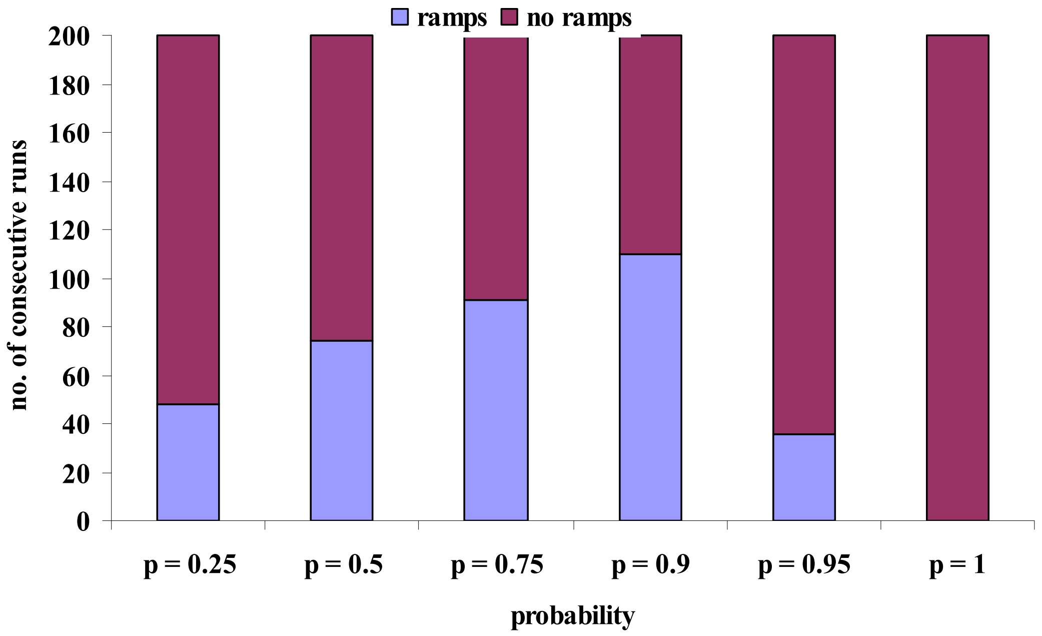


Figure 6.12. Number of runs, out of a maximum of 200, for each probability, showing both ramps and no ramps.

It is interesting to note that sustained activation and ramping decrease rapidly at around p = 0.95 when probability approaches 1, and disappear at p = 1, when the system becomes deterministic and full learning has taken place. In reality a probability of 1 is never reached as there will always be an element of uncertainty.

However, this effect is a direct result of the asymmetric coding of reward prediction errors, where we have scaled the negative errors by a factor of one sixth to simulate nature's own bias. A trade-off is seen between the frequency of receiving a reward and the magnitude of the surprise when a reward is seen. Without scaling, symmetry would be seen, where sustained activation and ramping would be similar for $p = 0.25$ and 0.75 with a maximum for $p = 0.5$, and dopamine would be encoding for uncertainty in the sustained activation/ramping in the delay period.

To conclude, I have addressed one of the points raised in Fiorillo et al. (2005) by demonstrating that both sustained activation and ramping are common in single trials during uncertainty, where probability is 0.25, 0.5, 0.75 and 0.9. However, using TD my simulations indicate that information about uncertainty is not available in the interval between presentation of the CS and receipt of the reward. Contrary to Fiorillo et al. (2003; 2005), both the sustained activation and the ramping effects do not appear to be coding for uncertainty. However, as concluded in (a) above, uncertainty would appear to be coded at the expected time of reward, in the averages of consecutive rewarded and unrewarded trials. This information on uncertainty builds up over a period of time and is, therefore, of limited use for prediction purposes by dopamine neurons, which require a more immediate assessment of ongoing reward prediction. This suggests that while TD provides information about uncertainty, it is likely to be optimally exploited by a postsynaptic system, possibly a working memory buffer that can monitor this information and utilise it in future trials in a more efficient manner.

(e) Trial History

In support of their original claims that dopamine encodes uncertainty, Fiorillo et al. (2005) suggested that if activity during the delay period is due to backpropagating error signals that originated on previous trials, then the activity in the last part of the delay period of each individual trial should reflect the last reward outcome. Specifically, if the preceding trial was rewarded, there should be more activity at the end of the delay period, and less activity if it was not rewarded. However, they found no evidence of this in their trials and provided an example of a recording showing no dependence of neural activity on the outcome of the preceding trial (Figure 6.9).

Figure 6.13 shows a history of rewarded and non-rewarded runs ---RNRNRNNNNN. After scaling, large $\delta(t)$ values were seen for the first six runs because alternate rewards and non-rewards were given, but runs 7-10 were not rewarded and, consequently, gradual extinction of the negative prediction error occurred. This example allows me to address a second point raised by Fiorillo et al. (2005) by showing that it is not always the case that less activity will be seen if a trial is not rewarded (and vice versa), as runs 8-10 show an increase in firing (towards baseline) following non-rewarded runs. It is necessary, therefore, for more of the history of previous runs to be taken into consideration than just the last reward outcome, when analysing reward prediction errors.

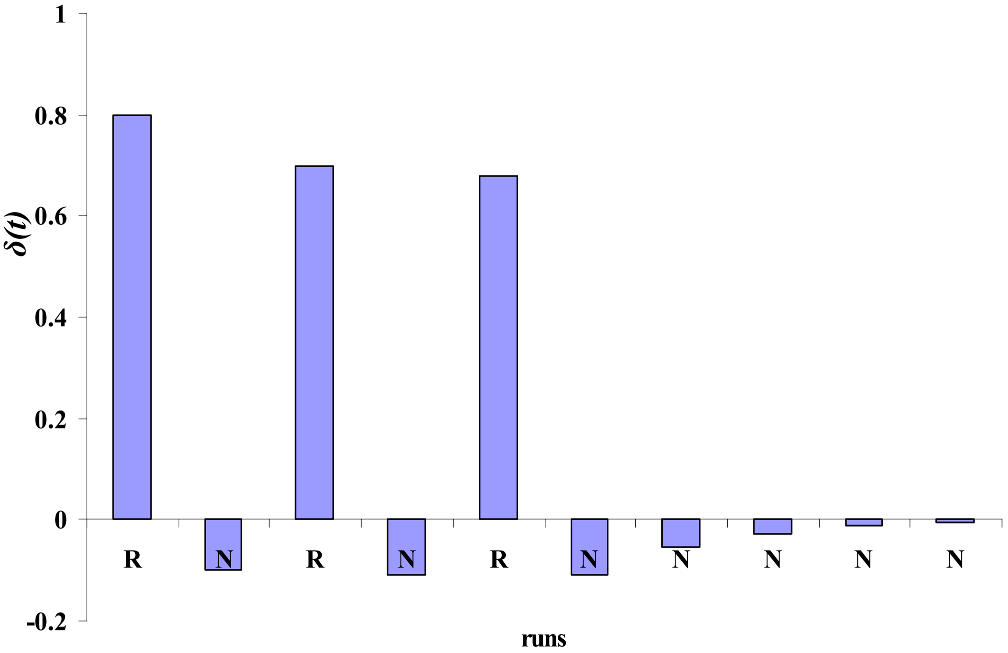


Figure 6.13. Scaled average *δ(t)* values at expected time of reward (S7-S8) recorded over 10 runs when *p* = 0.5.

Referring to the example provided by Fiorillo et al. (2005) shown in figure 6.9, the model would predict that the runs preceding the last rewarded trial in Diagram A were rewarded, as the last reward produced less activation than previous trials. In addition, in spite of possible noise in the data, the effect of no reward on the last trial (Diagram B) can be identified as a dip below baseline, just before the time of potential reward. This is a very large dip and so the model would predict that several rewarded runs preceded that particular non rewarded trial.

6.6 Discussion

This study has attempted to answer one of the key questions in the current debate over whether or not dopamine encodes uncertainty. Contrary to the claims of Fiorillo et al. (2003; 2005), single trial analysis in a simulation of reinforcement learning using TD reveals the possibility that dopamine may not encode uncertainty in the delay period of midbrain dopamine neurons in the form of either sustained activation or ramping. However, it would appear that TD does code for uncertainty, not in the inter trial interval, but at the expected time of reward, in the averages of consecutive rewarded and unrewarded trials. This information on uncertainty builds up over a period of time and occurs in a well learned state and hence, is of limited use for prediction purposes by dopamine neurons. Such information is only useful when the calculation is made in advance of a response, suggesting that while TD provides information about uncertainty, it could better do so to another system, possibly a working memory buffer that can monitor this information and utilise it in future trials in a more efficient manner.

I have produced a simple model that encoded values of states rather than weighted neurons and implicitly propagated reward prediction errors backwards in time using TD. This novel model successfully captured the following properties of dopaminergic activity seen in single cell recordings of reinforcement by Fiorillo et al. (2003): (a) the phasic activations at the expected time of reward; (b) the sustained increase in activity from the onset of the conditioned stimulus until the expected time of reward, during uncertainty; and (c) the sustained activation increasing with increasing reward magnitude. These findings support the claims of Niv et al. (2005) that the ramping effect seen in the delay period between presentation of a CS that predicts an uncertain reward and the expected time of receipt of that reward was due to backpropagating TD errors that arose as a result of averaging across multiple trials and was not encoding uncertainty.

What is new about this study is that I have drawn up criteria for sustained activation and ramping in single trials which permit analysis of single trials in our simulations. While these criteria are subjective, they are relatively liberal, which has allowed us to make predictions about sustained activation and ramping during the delay period in single trials in vivo. Importantly, these predictions can be tested and verified with behavioural data.

I have addressed one of the points raised in Fiorillo et al. (2005) by demonstrating that both sustained activation and ramping are common in single trials during uncertainty, where reward probability is 0.25, 0.5, 0.75 and 0.9. However I have shown that both effects increase with increasing probability to a maximum of 0.9, after which the effects decrease dramatically towards p = 1. Therefore, as neither sustained activation nor ramping is greater with maximum uncertainty we cannot support the claims by Fiorillo et al. (2003; 2005) that dopamine is encoding uncertainty during the delay period between CS and receipt of reward using TD.

In reply to a further point raised by Fiorillo and colleagues, I have demonstrated that activity in the last part of the delay period does not always reflect the reward outcome that followed the last exposure to that same CS. Specifically; the history of consecutive trials should be taken into consideration when analysing reward prediction errors and not just the last trial. In the presence of uncertainty, the particular course taken through a series of trials is different in each simulation, as it depends on the exact order of rewarded and non-rewarded runs, which are delivered randomly. It is important that these factors should be taken into consideration when interpreting data from peri-stimulus-time-histograms of activity over different trials and inter-trial averaging, such as that in Fiorillo et al. (2003).

The importance of taking the succession of trials into account is also evident from my results of the phasic activations at the expected time of reward. When averaging rewarded runs only the magnitude of the reward responses increased as probability decreased, but when averaging both rewarded and non rewarded runs, the magnitude of the ramping effect was greatest for maximum uncertainty, at a probability of 0.5.

While the value of this model is in its simplicity and transparency, which permits the interpretation and prediction of dopamine firing patterns, the model is parameter dependent and discrete, containing a set number of states. In reality neuron firing is noisy and therefore less predictable and a spiking form of this model could contain more realistic noise and more closely resemble dopamine neuron firing in vivo. In addition, the current model does not allow for more than one CS in any one trial and at the present time I am therefore not able to address a further point by Fiorillo et al. (2005) concerning a dissociation between the size of the ramp and the sustained activation at the estimated time of reward, identified in Tobler, Fiorillo and Schultz (2005).

Some evidence suggests that the persistent reward responses of dopamine cells during conditioning are only accurately replicated by a TD model with long-lasting eligibility traces, such as TD(λ) where λ has non-zero values (Pan, Schmidt, Wickens and Hyland 2005). It would be interesting to implement future versions of the model using this version of the TD algorithm to see the effect of different strengths of eligibility trace.

In addition, there is evidence from experiments involving decision-making in macaque monkeys that the activity of dopamine neurons reflects future choice of action as early as 122ms after the presentation of the CS (Morris, Nevet, Arkadir, Vaadia & Bergman 2006), presenting the possibility that dopamine neurons receive information about decision-making from another structure. This has implications for this particular model, where the dopamine signal is used to directly select possible actions. Furthermore, the results of Morris et al. favour the SARSA (state-action-reward-state-action) algorithms that assign a separate value (Q value) to each possible behavioural choice in every state as a better alternative in decision-making than using an actor-critic algorithm. However, the maze in this model only allows movement in one direction and there are no decisions other than whether to move to the next state or not, so this does not present a problem in these particular simulations.

In spite of the above evidence against models that use the dopamine signal to directly select possible actions; these simulations have combined the bias of nature towards the firing of dopamine neurons above baseline with TD to capture the detailed electrophysiological recordings of dopamine neuron firing. This strengthens the argument for TD as a valid method of modelling and quantifying the dopamine reward prediction error. While I appreciate that TD must be one of many algorithms working simultaneously in the brain, I agree with Niv et al. (2005) that the ramping signal, both in single trials, and when averaged over multiple trials, is strong evidence for the nature of the learning mechanism of a shift in dopamine activity from expected time of reward to the CS, using TD. I suggest that it is both reasonable and biologically plausible for future models of dopamine to include TD learning. This chapter demonstrates the value of computer modelling in that our model has generated testable predictions that can be verified with behavioural data.

6.7 Chapter Conclusions

The simulations detailed in this chapter are an example of science through simulation and demonstrate how computational modelling can help to clarify a position by generating testable predictions that can be verified with behavioural data (Chapter 2.3). The arguments presented strengthen the use of TD as a valid method of modelling and quantifying the dopamine reward prediction error. It is therefore both reasonable and biologically plausible for future models of dopamine to include TD learning.

However, there are certain drawbacks to using a TD approach, which are discussed in the next chapter. In Chapter 7, it is my intention to draw attention to those limitations and to shift the focus of this thesis back to my longer-term aims of the effect of dopamine dysfunction in schizophrenia. In particular, I describe an alternative body of research, where it is assumed that an animal builds an explicit internal model of its environment during conditioning. I investigate a computational model of the function of dopamine in a Reinforcement Learning paradigm, by Smith, Li, Becker & Kapur (2006), which, like the models described in Chapters 5 and 6, incorporates both dopamine and the Incentive salience Hypothesis. However, unlike the models previously described, the model based account also gives an account of the tonic firing of dopamine neurons and its effect on the expression of previously acquired behaviour.

Chapter 7

Limitations of Temporal Difference: Towards a Dual Systems Approach

The long-term aim of this research is to explore the application of connectionist models as a paradigm for schizophrenia, with a view to generating and testing theories of the disorder. It is clear that any explanation of the symptoms and cognitive deficits associated with schizophrenia should include dopamine dysfunction (Chapter 2) and my research so far has led to a thorough investigation of TD as an explicit method of modelling and quantifying dopamine as a reward prediction error. In Chapter 5 I described and implemented a computational substrate for incentive salience by McClure et al. (2003), and in Chapter 6 this model was extended to conduct an analysis of the relationship between TD learning and uncertainty coding in a computational model of dopaminergic signalling. While the focus of this thesis so far has centred on using TD as a model of dopamine function in reinforcement learning, it is important to note that there are limitations to using TD alone to account for action control in the brain. In this chapter I draw attention to those limitations and shift the focus of this thesis back to my longer-term aim of implementing a connectionist model of the effect of dopamine dysfunction in schizophrenia.

One of the major benefits of using TD is that it utilises minimal computation by caching or storing prediction values of estimated future reward or incentive value, over successive timesteps. However, this caching of values is also a limitation that can manifest itself in the following ways:

- As TD only stores values of expected future reward or incentive, the basic TD model is unable to distinguish between different rewards with a similar value that are preceded by an appropriate CS (Smith Li, Becker & Kapur 2006). In particular, the basic TD model is unable to distinguish between the effects of dopamine manipulation on distal rather than proximal rewards (e.g., running speed in a maze versus consumption of reward, Chapter 5.3) (Daw, Niv & Dayan 2005; Smith et al. 2004; 2006).
- TD can be brittle as the prediction chains used by TD over successive time-steps break down when the CS-US relationship is unreliable, such as in the complicated 1-2-AX working memory task, which involves maintaining both subgoals and higher order goals (O'Reilly, Frank, Hazy & Watz 2007).
- The simple actor-critic is insensitive to motivational state and fails to take into account some of the psychological differences between Pavlovian and instrumental conditioning (Dayan & Balleine 2002).
- Any change in task will have to be relearned explicitly, which will take time. In reality, relearning often needs to take place quickly, so current TD models do not account for all types of learning (Daw et al. 2005).

A model based approach (Sutton & Barto 1998), where it is assumed that an animal builds an explicit internal model of its environment during conditioning (Smith, Li, Becker & Kapur 2004; 2006), can address some of the problems associated with TD. However, model based reinforcement learning carries its own limitations associated with searching a large environment and the possible errors incurred as a result of using strategies to reduce the search space to manageable proportions. An alternative is to combine the benefits of both TD and explicit internal model representations in a dual system action control in the brain (e.g., Daw et al 2005; Dayan & Balleine 2002).

TD learning is often referred to as a model free form of reinforcement learning, but in reality it stores a rudimentary model of the environment as a series of states and their associated values. However, unlike model based reinforcement learning, TD does not represent the underlying cause-effect contingencies of the environment as it does not store the consequences of taking an action from each state (Smith et al. 2006). In addition, there are different degrees of model based learning, ranging from basic models which store simple state-action pairs, for example Q-learning (Watkins & Dayan 1992), to more sophisticated tree-searches (Daw et al. 2005) that are more computationally expensive. When I refer to the distinctions between model free and model based reinforcement learning in the remainder of this thesis, I place my arguments within the framework of Daw et al. (2005) (Section 7.4), and refer to TD as model free, where TD and model based approaches are two opposite extremes in a trade-off between computational efficiency and the statistically efficient use of experience. However, it should be noted that there is a less clear-cut distinction between the two, which are better described as being separated on a continuum between model free and model based reinforcement learning.

In the following sections, I elaborate further on the arguments against TD listed above, and start by looking at the alternative model based account of dopamine function by Smith and colleagues in Section 7.1. Their line of reasoning bears important similarities to mine as we both have the long-term goal of a better understanding of schizophrenia through the ideas of incentive salience and dopamine dysfunction. Section 7.2 contains a brief account of a new algorithm developed by O'Reilly and colleagues as an alternative to TD, which is not developed further in this thesis, and in Sections 7.3 and 7.4 I look at two models by Dayan and Balleine and Daw et al. that combine the benefits of both model free and model based accounts of reinforcement learning.

In Section 7.5 I describe a framework by Yin and Knowlton (2006) for the role of the basal ganglia in habit formation. Unlike the work in the previous sections, it is not a model addressing the limitations of a model free TD approach, but, it gives an account of the cortico-basal ganglia networks and the distinction between goal directed actions and stimulus response habits, and thus builds upon the argument for a dual system of action control in the brain. These five sections have inspired my connectionist model with dual weights, which is developed in Chapter 8.

7.1 A Model Based Account of Dopamine Function

A series of models have been developed by Smith and colleagues using a model based approach, as an alternative to TD, where it is assumed that an animal builds an explicit internal model of its environment during conditioning (Smith, Becker & Kapur 2005; Smith, Li, Becker & Kapur 2004; 2006). Their latest model extends TD to include action at the level of dopamine receptors (Smith Li, Becker & Kapur 2007), but here I focus on the model based approach. This body of research is particularly interesting to me as the area of interest bears important similarities to some of the work detailed in previous chapters: (i) we both have the long term goal of obtaining a better understanding of schizophrenia through dopamine dysfunction; (ii) both approaches, whether model based or model free, accept the importance of dopamine acting in the estimation of future reward and in the generation of a prediction error; (iii) expected future reward is interpreted in both models as 'wanting' and not 'liking' in accordance with an incentive salience approach; (iv) Smith and colleagues began with a model based account of dopamine function, but aspired to incorporate the benefits of TD into future models, while I have begun with a TD model free account of dopamine function, but now seek model based amendments in order to overcome the limitations of TD described above.

In Section 7.1.1 I look at the earlier model of antipsychotic action in conditioned avoidance that distinguishes between the effects of dopamine manipulation on distal rather than proximal rewards (Smith et al. 2004). It is a model based account of the tonic function of dopamine in the expression of previously acquired behaviour, and does not attempt to model the phasic dopamine bursts associated with the learning process. Despite the difference in the modelling techniques, this model bears important similarities to the computational substrate for incentive salience by McClure, Daw & Montague (2003) detailed in Chapter 5 and I compare the different model free and model based accounts. A later model of both phasic and tonic dopamine neuron firing (Smith et al. 2006) is discussed in Section 7.1.2, which they claim offers a parsimonious distinction between phasic and tonic dopamine function and is again able to distinguish between the effects of dopamine manipulation (via antipsychotic action) on distal rather than proximal rewards in animal studies. In Section 7.1.3 I look at how this model based account can relate to understanding psychosis and schizophrenia.

7.1.1 A Model Based Account of Antipsychotic Action in Conditioned Avoidance (Smith, Li, Becker & Kapur 2004)

Like McClure et al. (2003), detailed in Chapter 5, Smith et al. (2004) modelled the effects of antipsychotic drugs on behaviour, but this time using a model based account of reinforcement learning on a conditioned avoidance response, or a negative reinforcer (Chapter 4.1.1), rather than using model free reinforcement learning on a positive reinforcer.

The model could account for two important features seen in animal experiments: (i) the effects of dopamine manipulation on the expression of behaviours independently of their acquisition; and (ii) it could distinguish between the effects of dopamine manipulation on distal rather than proximal rewards. Both of these effects are illustrated in an animal study by Cousins et al. (1996) where, after training, rats showed a preference for a larger, but obstructed reward (distal), over a smaller, unobstructed reward (proximal). Following dopamine depletion a switch in preference was seen from the larger to the smaller reward. This shift from a distal to a proximal reward showed the effect of dopamine depletion on previously learned behaviour independently of the acquisition process.

The Task

The scenario modelled was an animal experiment by Maffii (1959) of a rat in a two-compartment shuttle box that had learned to associate a prior neutral CS, an auditory tone, with an aversive unconditioned stimulus or outcome, an electric foot-shock. The rat learned to *avoid* the shock by moving away from the area as soon as the CS was presented, rather than just exhibiting *escape* behaviour when the shock was administered. Furthermore, it was shown that, with time, the rat learned to avoid the foot-shock area as soon as it entered the cage, before the CS signal, demonstrating second order conditioning, where the cage itself became the CS.

A switch in behaviour was seen following dopamine receptor blockade, when the rats became less likely to avoid the shock with increasing levels of antipsychotics, and more likely to escape in the presence of the shock itself. This was in accordance with the hypothesis that blockade of the dopamine D2 receptors by antipsychotics reduced the incentive salience and thus the motivation to escape by avoiding (wanting/not wanting) (Berridge & Robinson 1998), without affecting the actual escape (liking/disliking).

Smith and colleagues demonstrated the accepted finding that low, non-cataleptic doses of antipsychotic drugs affected an animal's ability to perform the avoidance response (distal negative reward), but had no effect on the escape response (proximal negative reward). The model also accounted for the fact that lower doses of antipsychotics were needed to disrupt secondary avoidance conditioning (being put in the cage) than for primary avoidance conditioning (the auditory tone), and was able to predict dose-dependent effects of antipsychotic drugs on avoidance and escape response latencies seen in novel latency data.

The Model

This model was of the expression of previously acquired behaviour, associated with tonic dopamine function and, unlike McClure et al. (2003); no attempt was made to model the acquisition process associated with the phasic dopamine function. It was assumed that the animal had already formed an explicit internal model of its environment through trial and error interactions with the task and that dopamine had more of a gating role (Chapter 3.2) for salient information.

The model consisted of states, rewards and transitions, and an early version had five states for the animal to occupy, including a state for hearing the tone (CS), a state for receiving the shock (US), a safety state (the termination state), and two wait states, where the animal did nothing but remain in the same place, representing the delay between the onset of the CS and the shock (Figure 7.1). The negative reward, of -1 representing the shock was delivered in the shock state, and a reward of 0 at the other

4 states. Transitions, modulated/gated by dopamine, represented the consequences of taking each action.

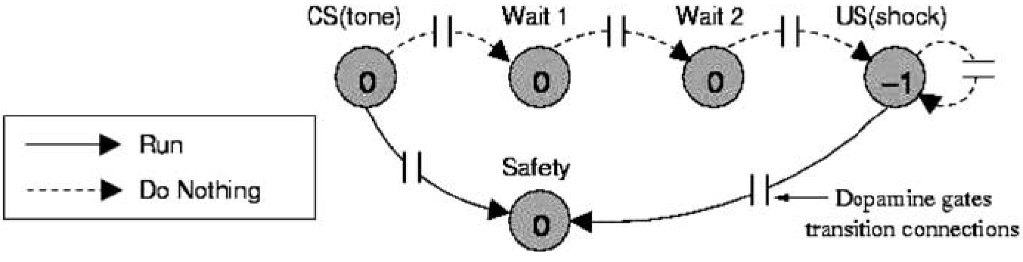


Figure 7.1 Representation of the internal model of the environment, built up by trial and error interaction. The circles are the possible states, and the arrows denote the consequences of taking an action in each state. Each state contains a value for reward (-1 represents the punishment/negative reward in the shock state). The arrows represent the transition function, which is modulated (or gated) by dopamine (see vertical bars). The wait states are internal states for which there is no external cue, and allow the model to represent the delay between CS onset and the shock. The safety state is a terminal state at which the trial is ended. Taken from Smith et al. (2004).

Attributing values to each state enabled the consequences of taking an action to be calculated by hypothetically playing through the options available to the animal. For example, for an animal at the CS state, the model could be used to motivate behaviour by giving an activation of 1 on presentation of the CS and generating the expected future reward of the two options available: (i) run to safety or (ii) do nothing. If the rat chooses (i) to run to safety, it is necessary to propagate the activation value to the safety state and this value is directly proportional to the strength of the transition connection between the CS and safety state, which is modulated by dopamine. This would result in a future reward of 0, the result of arriving in the terminal safety state. If the rat chooses (ii) to do nothing, it is necessary to propagate the activation value through the two wait states pending the shock, each also modulated by dopamine. This would result in a future reward of -1, the result of waiting for the shock with a negative reward of -1. It is assumed that animals are motivated towards reward and away from punishment, and so the option to run will be selected over the option to wait. Following an action the change in the animal's environment will be reflected in the model by a change in state.

The model was later adapted to a more generalised form so that choices could be made in a similar way from any state in the model by assuming a distinct delay state for each second of time between the CS and US and adding probability to the value of acting in a state. In order to interpret expected future reward as salient, the CS was ignored with increasing probability as the expected future reward associated with doing nothing became closer to zero. This resulted in activation only to the shock and not in the internal delay states, thus avoiding keeping track of non-salient stimuli.

The effect of antipsychotics, or dopamine blockade, on the choices made was modelled by assigning values to the transition connections, simulating the modulating action of dopamine (D). If D = 1, there are adequate levels of dopamine available for an optimal assessment to be made. If D = 0, there is no dopamine available due to a complete blockade of dopamine receptors, while dopamine values between 0 and 1 constitute partial dopamine blockade. The activations of the safety and wait states are affected by the activation of the CS multiplied by the probability of safety or doing nothing respectively, multiplied by the global availability of dopamine. If D = 1, the choice between safety and shock is 0 or -1 and so the choice is obvious and therefore the animal is more likely to make a good decision to avoid the shock. However, with reduced levels of dopamine, for example D = ½, the choice for safety is still 0 (½ × 0), but the choice for shock propagated back through two wait states becomes $-^{1}/_{8}$ (-1 × ½, then -1 × ¼), resulting in a less obvious decision and an increased probability of making the bad decision to wait, resulting in a shock. If dopamine is depleted, D = 0, the rat will never jump onto the pole to avoid the shock (-1 × 0 = 0) and (0 × 0 = 0).

Results

Using the ideas of dopamine as a measure of incentive salience, it was possible to model the qualitative findings of Maffii (1959) that lower doses of antipsychotic drugs were needed to disrupt secondary avoidance (being put in the box, the environmental cue) than primary avoidance (the auditory stimulus) and in the same way lower doses of antipsychotics were needed to disrupt primary avoidance than escape only (the shock state) (Figure 7.2). Results are provided for primary, secondary and escape responses following increasing doses of the antipsychotic (chlorpromazine), as a percentage of the number of responses made without the drug.

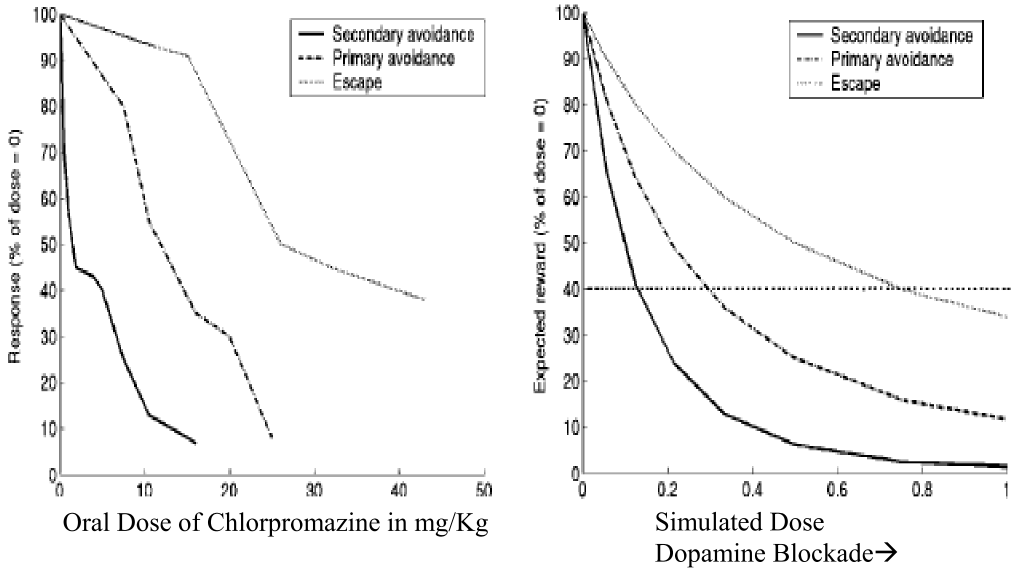


Figure 7.2 (Left) Number of secondary avoidance, primary avoidance and escape responses under increasing doses of antipsychotic as a percentage of the number of responses without the drug from Maffii (1959) (Right) Simulated results for increasing values of D (0 to 1). Taken from Smith et al. (2004).

The authors stressed that it is the qualitative and not the quantitative performance of the model that is of interest as the model does not attempt to address the underlying neurochemical processes. This effect was modelled in the internal representation by the respective distance of each CS from the shock, and achieved by implementing a larger number of internal delay states, or distance, between the secondary avoidance and the shock state, than for the primary avoidance and the shock state. In this manner, dopamine blockade, affected the previous allocation of incentive salience to the conditioned stimuli, having the greatest effect on stimuli more distal to reward (negative reward/punishment). Furthermore, the model was able to predict both the increase in latency, and the change in pattern of both avoidance and escape (not shown). These predictions were validated by animal models which confirmed the dose-dependant effect of four different antipsychotics, haloperidol, chlorpromazine, risperidone and clozapine, on peak latency to avoid (want) without affecting escape (dislike).

Conclusions

McClure et al. (2003) and Smith et al. (2004) are two different approaches, one model free and the other model based, that used the concepts of incentive salience and expected future reward in a computational model to unite psychological and pharmacological theories, grounding those psychological theories by providing testable predictions which were validated by animal experiments. McClure and colleagues modelled the acquisition process in reinforcement learning and linked the ideas of incentive salience to reward prediction via TD learning; while Smith and colleagues modelled the generation of expected reward on the expression of previously acquired behaviours, independently of the acquisition process, linking incentive salience to reward prediction via a gating role for dopamine.

In Chapter 5.2.3 I referred to a hypothesis by Ikemoto and Panksepp (1999) that suggested that dopamine may underlie appetitive approach behaviours but not consummatory behaviours such as licking. This theory is in line with other areas of research that suggest that dopamine in the nucleus accumbens/ventral striatum is important for responding to conditioned stimuli and stimuli that are spatially and temporally distant (distal) rather than proximal to the organism (e.g., Daw, Niv & Dayan 2005; Maffii 1959; Salamone, Cousins & Snyder 1997; Smith, Becker & Kapur 2005; Yin, Knowlton & Balleine 2004). As already mentioned, these results pose a problem for the basic TD framework, which does not distinguish between the two motor actions of running and licking. However, as discussed above, the model based account was able to distinguish between two different conditioned stimuli: the distal environmental box cue and the proximal auditory stimulus cue. Therefore, this model based account of tonic dopamine function has been able to demonstrate: (i) that dopamine manipulation can affect the expression of behaviours independently of their acquisition; and (ii) sensitivity to the relationship between the CS action and the US outcome, as it can distinguish between the effects of dopamine manipulation on distal rather than proximal rewards.

It is clear that both methods of modelling have something to offer: the model free gives an account of the role of dopamine in learning, but is insensitive to the difference between proximal and distal rewards; while this model based account is sensitive to the relationship between the CS and US, but does not model the learning process. Smith et al. conclude that both prediction error hypothesis and the gating role of dopamine have a role to play in the future as a combination of the two approaches may provide a wider range of behavioural data for both acquisition and expression. In the next section I look at a later model based account by Smith and colleagues that also addresses the learning process.

7.1.2 A Model Based Account of Phasic and Tonic Dopamine Function (Smith, Li, Becker & Kapur 2006)

Smith et al. (2006) modelled both the phasic dopamine prediction error signal which drives learning and the tonic background firing rate of dopamine, which could account for the expression of previously acquired behaviours (Parkinson et al. 2002). They demonstrated that their model based adaptation could account for dopamine neuron firing patterns and associative learning paradigms such as latent inhibition, Kamin blocking and overshadowing, as easily as TD. However, this model had the added advantages of offering a parsimonious distinction between phasic and tonic dopamine function as well as being able to distinguish between the effects of dopamine manipulation (via antipsychotic action) on distal rather than proximal rewards in animal studies.

Modelling Phasic Dopamine in Reward Learning

This model based approach assumed that an internal model of the environment was created and stored during learning. This involved learning the transitions between states and used different representational techniques to TD. Instead of storing values of states like TD the model based account stored: (i) the rewarding impact of stimuli, the estimated future reward, *R*; and (ii) transition connections between states, *T*. Together these stored values were used for a systematic search of the environment to calculate future rewards (the calculated return) to a pre-specified depth, υ. This approach requires a greater computational capacity than the computationally efficient TD, which only stores the values of states. The reward values associated with each state, *R*, and the estimate of future reward were speculated to be represented in the orbitofrontal cortex and the basolateral amygdala.

Another benefit of this model over TD is that it introduced the concept of surprise, which is the degree to which a current state is unpredicted. The dopamine phasic response was governed by both *surprise* and *significance* (the calculated return), where significance corresponded to the estimated future reward and was interpreted as corresponding to the degree to which an animal is motivated to achieve reward (avoid punishment) based on the expected future reward of a CS. A phasic response was given only if both surprise and significance were recorded, that is, if the currently active state was not predicted by the previously active state AND if the currently active state was rewarding or predicted reward:

$$\mathrm{DA}_{phasic} = Surprise \times Significance$$

The resulting prediction error was similar to TD (although TD only operates on significance). In the model based approach, future reward is dynamically recalculated each time a CS is encountered, giving the motivational value of the CS based on the current motivational state of the system, but in TD this value has already been learned and is stored, ready for immediate use, so TD depends on the motivational state during learning only. As already mentioned, TD is unable to distinguish between different reward types of a similar value as no representations of the underlying cause/effect of the CS-stimulus relationship are stored. This is not a problem for a model based account, which looks at the action-outcome relationships that are recalculated on every stimulus presentation and will find a different US surprising. Furthermore, as the model based account separates reward from outcome, it can distinguish between different aversive and appetitive CS. For example, if a CS precedes both a negative and a positive reward, then the model based approach will predict the phasic response, while the positive and negative predictions will cancel each other out in a model free TD account.

Modelling Tonic Dopamine in Motivational Processes

In order to model the tonic role of dopamine Smith and colleagues incorporated an aspect of their earlier model (Smith et al. 2004) into the look-ahead process described above, where the previously acquired transition strengths were multiplied by a value representing tonic dopamine of between zero and one. If $D_{tonic} = 1$, there were adequate levels of dopamine available for the look-ahead process to proceed. However, lower levels of tonic dopamine will affect each new cycle of the look-ahead process, and will temporarily reduce the future reward or incentive salience of a CS, as generation of future reward is dependent on the activity of the states while looking-ahead. Here, tonic dopamine in the ventral striatum is acting as an online discount factor and reductions in tonic dopamine will have a greater effect on distal than for proximal rewards, in a similar manner to that demonstrated in Smith et al. (2004).

Conclusions

This model based account successfully incorporated both phasic and tonic roles of dopamine into one model of reward learning and incentive salience. As with the earlier model, this model was able to distinguish between the effects of dopamine manipulation on distal rather than proximal rewards, but this updated version had the advantage of being able to model the learning process as well as modulating the expression of behaviours independently of their acquisition. However, the model did not address action selection or model dopamine neuron firing below baseline.

It should be noted that there are weaknesses associated with using a model based account that occur: (a) as a result of the large search space that arises when associating specific actions with specific outcomes, and (b) from the methods used of reducing the search space to manageable levels, which give rise to errors as a result of the short cuts taken to prevent a combinatorial explosion (Daw et al. 2005). The trade-off between using model free and model based accounts is discussed in Section 7.4.

As well as addressing some of the limitations of TD, the model based accounts of Smith and colleagues, described above, have been motivated by the goal of a better understanding of schizophrenia, and I develop these arguments further in the next section.

7.1.3 How Does This Relate to Schizophrenia?

In view of the correlation between the ability of antipsychotic drugs to block dopamine D2 receptors and the effectiveness of those drugs in the treatment of psychosis, it is posited that psychosis results from a dysregulation of the dopamine mesolimbic system (Chapter 2.1.3). Conditioned avoidance is a preclinical drug test for antipsychotic action and low doses of antipsychotic drugs are known to disrupt avoidance behaviour leaving the escape response intact (e.g., Alder & Clink 1957; Cook & Weidley 1957), but the avoidance response is restored once the drug wears off (Smith et al. 2004). As the drug has an immediate effect on behaviour, the effects are not thought to be due to unlearning, which would take time; so it is believed that blockade of the dopamine D2 receptor is the neurochemical link between conditioned avoidance in rats and antipsychotic action in people (Wadenberg et al. 2001; Smith et al. 2004).

With the primary aim of understanding the effect of dopamine dysfunction in schizophrenia and, in particular psychosis, Smith et al. (2004; 2006) modelled the functional significance of the striatal D2 receptor on behaviour in a conditioned avoidance paradigm. Unlike TD, their model based account was able to distinguish between the effects of dopamine manipulation on distal rather than proximal rewards and identified a tonic role for dopamine in the generation of reward, which was independent of the acquisition process.

In their later model of both tonic and phasic dopamine function, Smith et al. (2006) claimed that their implementation provided a formal interpretation of a process where aberrant phasic dopamine responses could label both internal and external events inappropriately as being surprising or significant, leading to delusions and possibly hallucinations associated with thought disturbance in schizophrenia, outlined in a framework by Kapur (2003) and detailed in Chapter 4.4.1. A clearer understanding of the role of dopamine and the best methods of modelling those functions will help in the quest for the understanding of schizophrenia and they posit that an aberrant internal model of the environment, such as that constructed in a model based approach, may provide some of the answers. The action of antipsychotic drugs may protect against the formation of the aberrant internal representations by both attenuating aberrant incentive salience via phasic dopamine signals, and by dampening the motivational efficacy of existing associations via tonic dopamine function (Smith et al. 2006).

7.2 The Primary Value and Learned Value Pavlovian Learning Algorithm (O’Reilly, Frank, Hazy & Watz 2007)

Randall O’Reilly and colleagues have abandoned the TD algorithm in favour of a primary value learned value (PVLV) Pavlovian learning algorithm to understand the function of dopamine in reward prediction (O’Reilly, Frank, Hazy & Watz 2007), where dopamine signals reward association and not reward prediction. The model contains two separate systems, which are further subdivided into excitatory and inhibitory subcomponents: (i) the PV is engaged by primary reward and learns to expect an unconditioned stimulus, thereby inhibiting the associated dopamine phasic burst at the time of receipt of the reward, over time; (ii) while the LV learns about stimuli that are reliably associated with primary rewards and drives the dopamine phasic burst when the conditioned stimulus is presented. The authors claim that these two systems provide a more direct mapping onto the underlying neural substrates than TD. In addition these two systems are more robust to variability in the environment as they do not rely on the prediction chains used by TD over successive time-steps, which they claim break down when the CS-US relationship is unreliable, such as in the complicated 1-2-AX working memory task, which involves maintaining both sub-goals and higher order goals. This promising avenue of research is in its infancy and a considerable amount of work remains to be done before it can compete with TD as the most popular computational account of conditioning and dopamine firing.

One of the long-term aims of O'Reilly and colleagues is to model the complex interactions between the basal ganglia and prefrontal cortex, where actively maintained representations in the prefrontal cortex are dynamically updated/gated by the basal ganglia (Chapter 3.2). In an attempt to deconstruct the homunculus and understand the mechanisms of control between the two systems, the PVLV algorithm has been applied in their Prefrontal Cortex, Basal Ganglia Working Memory (PBWM) model (Hazy, Frank & O'Reilly 2006; 2007; O'Reilly & Frank 2006) and has been successful in performing a number of tasks including Stroop, AX-CPT, 1-2-AX and the Wisconsin Card Test.

The PVLV algorithm was implemented in the *PDP++* software package using the sophisticated *Leabra* framework (O'Reilly and Munakata, 2000). An investigation of this complex system would be extremely time consuming and is beyond the scope of this thesis; although a simplified version of *PDP++* was used in Chapter 3.1 for the simulation of the speech perception network. However, I will refer briefly again to the attempts by O'Reilly and colleagues to deconstruct the homunculus in Chapter 8. In the next section I look at an approach to reinforcement learning that combines the benefits of model free and model based learning.

7.3 Differentiating Between Pavlovian and Instrumental Conditioning: Advantage Learning (Dayan & Balleine 2002)

Dayan and Balleine (2002) evaluated the actor-critic model of the dopamine system and claimed that the basic actor-critic model failed to take into account a large amount of data from psychological and neurobiological data on motivation. They referred to the many psychological differences between Pavlovian and instrumental conditioning and suggested a model with separate systems for Pavlovian and instrumental learning, with different neural underpinnings. The psychological argument for the differences between the two systems is beyond the scope of this thesis, but for a full review see Berridge and Robinson (1998) and Dickinson and Balleine (2002). Dayan and Balleine claimed that action choice and motivation separated the two systems: with the Pavlovian system being rigid in action selection, but flexible in response to motivation; and the instrumental system giving greater flexibility in action selection, but lacking sensitivity to motivational state. They equated the hard-wired, stimulus substitution-sensitive route of Pavlovian motivation for the control of habits, as acting in line with the classic TD prediction error, via the shell and possibly the core of the nucleus accumbens; while they considered that the plastic motivational route operated via the amygdala and the orbitofrontal cortex.

The simple actor-critic is insensitive to motivational state and Dayan and Balleine suggested that Q learning (Watkins & Dayan 1992), a variant of TD, can allow for choosing different actions, from a selection of actions, according to the current motivational state. Their model considered future reward by following a policy, in a similar manner to the actor-critic, but instead evaluated actions according to their advantages (a quantity of the difference between two long term rewards: a Q value relating to the action; and a value averaging over all actions). Each state had a Q value and the action chosen was the best action with the highest Q value. This was an effective way of making an appropriate action selection from a choice of possible actions according to the relevant motivational state, as the better actions were chosen as learning progressed. Advantage learning contains the benefits of model free learning, but has the added bonus of representing the advantages of moving to a state, in a model based manner. This method of learning offers a solution for the transition from the selection between many appropriate actions available during instrumental conditioning to the single action taken during Pavlovian conditioning.

Advantage learning will not be developed further in this thesis, but in the next section I look at a later model by Daw, Niv and Dayan (2005) that continues the argument for a dual system approach and demonstrates the advantages of using a model that combines both model free and model based reinforcement learning.

7.4 Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control (Daw, Niv & Dayan 2005)

Daw, Niv and Dayan (2005) referred to the competition between multiple systems for behavioural choice in the brain, and the problem of arbitration between the systems when they disagreed. They suggested a model of dual action choice, where the systems operated separately and in parallel, governed by a Bayesian principal of arbitration.

In order to illustrate the problem they identified two systems: (i) a habit system including reflex responses, associated with the dorsolateral striatum and its afferents, and (ii) a system for goal directed (reflective or cognitive action planning) actions, associated with prefrontal regions. The assumed dissociation between the two systems of control was based on electrophysiological (Jog, et al. 1999; Holland & Gallagher 2004; Pasupathy & Miller 2005), fMRI (McClure et al. 2004; O'Doherty et al. 2004) and lesion studies, where the values of rewards are re-valuated (unexpectedly changed during conditioning), and have a different effect on each system (Balleine & Dickinson 1998; Killcross & Coutureau 2003; Yin, Knowlton & Balleine 2004).

7.4.1 The Trade Off Between Model Free and Model Based Reinforcement Learning

In order to model the two systems, Daw and colleagues claimed that model free reinforcement learning, such as TD, lends itself to inflexible stimulus-response actions that take place quickly in a habitual manner, but require extensive relearning; while a model based account equates better with flexible, often one shot, goal directed actions, which require little relearning. They interpreted a model free TD controller and a model based controller, using a tree-search algorithm, as two opposite extremes in a trade-off between computational efficiency and the statistically efficient use of experience; although as already mentioned at the beginning of this chapter, there is a less clear-cut distinction between the two in reality. The benefits and weaknesses of the model free habit system and the model based system for goal directed actions are summarised in Table 7.1. The model free system is computationally simple (Chapter 4.3), yet suffers from inflexibility. Conversely, the model based system, can be expensive in terms of memory and time, but is more flexible.

Daw and colleagues claimed that accuracy was the key to arbitration between the two systems, which will have different benefits and weaknesses in different situations. The system that dominates will depend on inexperience and task complexity versus search depth. The Model based system should dominate early in training, and during complex tasks, where data is scarce and stored TD values are unreliable; but model free systems should dominate during a well learned task, when habits prevail for actions more distal from reward. Here, the deeper search is costly and more error prone due to the necessity of pruning and exploring a limited set of paths in an effort to cut down on the extensive search space in a tree search.

Table 7.1 contrasting the benefits and weaknesses of model free and model based learning

Model Free Habit System	**Model Based System for Goal Directed Actions**
-Computational simplicity	-Greater computational load, but utilises flexible, statistical use of experience
-No cognitive map/search tree	-Builds cognitive map/search tree
-Quick response	-Expensive timewise
-Cannot distinguish between different motivational outcomes	-Can distinguish between different motivational outcomes
-Inflexible and insensitive to relearning	-Flexible and sensitive to relearning
-Error prone when learning and during complex tasks	-Error prone during a well learnt task that relies on habit due to necessary approximations

7.4.2 Modelling an Outcome Devaluation Task

Daw and colleagues claimed that outcome re-valuation, where unexpected shifts in circumstances occurred that required immediate changes in behaviour, was the key to understanding the workings and neural underpinnings of each system. A typical re-valuation experiment (Figure 7.3) involves the following three stages: (i) hungry rats are trained to receive food as a reward following a sequence of actions, e.g., a lever press (distal action) followed by entry into a food magazine (proximal action); (ii) the value of the food is reduced by either pre-feeding or pairing with illness; (iii) the rats are reintroduced to the food, without the outcomes in (i) or (ii).

It was predicted that model based and model free systems would have different devaluation profiles. The model based habit system of caching values is not immediately sensitive to the specific outcome information associated with the devaluation, and it will take time for a change in behaviour to occur following relearning of the values, while the flexible model based system that is outcome sensitive and goal directed, will show an immediate behavioural change.

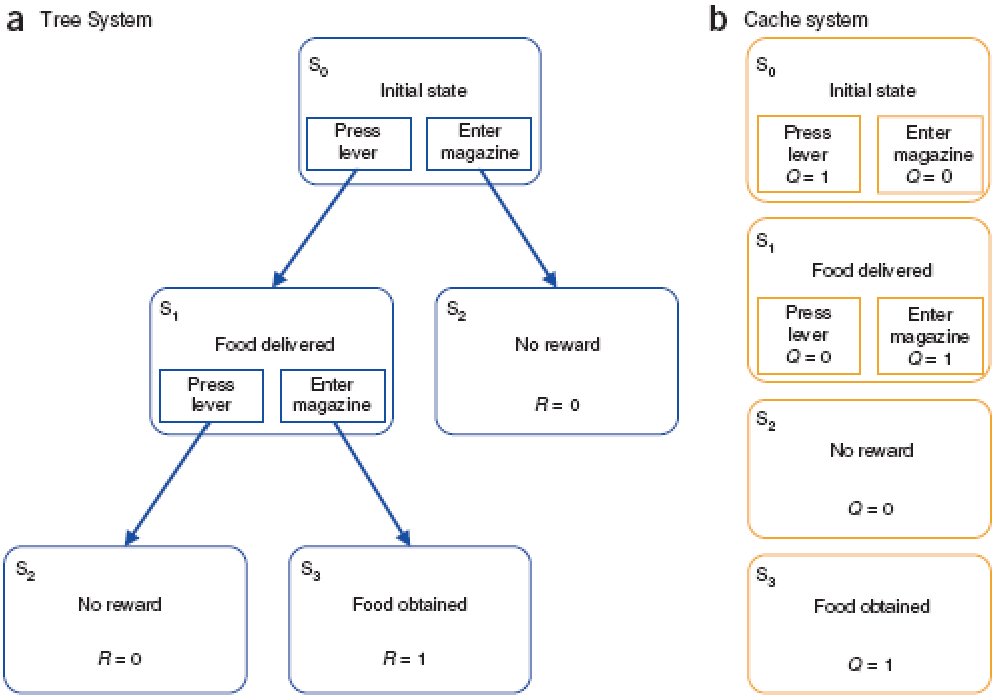


Figure 7.3 Task representations for a typical animal devaluation experiment used by (a) a model based tree-search system and (b) a model free caching system. Taken from Daw et al. (2005).

Furthermore, moderately trained lever presses are sensitive to outcome devaluation, while extensively trained lever presses are not, suggesting a transition from model based to model free systems with extensive training, i.e., after a habit has been formed. This transfer is affected by (a) the complexity of action choice, where in complex tasks extensively trained actions remain sensitive to outcome devaluation, and (b) the proximity of the action to reward, where actions closer to the reward are more sensitive to devaluation.

In their framework Daw and colleagues hypothesised that the brain might estimate the accuracy for arbitration between the model free and model based systems by tracking the relative uncertainty or expected inaccuracy of predictions made by each of the two separate systems. They implemented a model with both control systems: a TD value-caching system; and a tree-search system capable of chaining together short-tern predictions about the immediate consequences of each action in a sequence. Each system produced different and differentially accurate estimates of the value function, and was capable of estimating the uncertainty of its action predictions. They suggested a Bayesian principle of arbitration according to uncertainty, so that the system dominant at any one time was the one giving the highest level of accuracy.

As well as operating in parallel, it has also been shown that the two systems are able to operate separately. Chemical manipulations in lesion studies have revealed that the transfer to a model free caching system can be blocked by dopamine lesions or depletions to the striatum (Yin et al. 2004); while devaluation sensitivity, associated with model based systems, is eliminated by lesioning a wide range of structures, including prefrontal-associated regions of dorsomedial striatum (Yin et al. 2005) and orbitofrontal cortex (Izquierdo, Suda & Murray 2004).

7.4.3 Results

Daw and colleagues simulated the action choices in the typical animal devaluation experiments described above (Figure 7.3) and demonstrated the different devaluation profiles expected for (i) the estimated uncertainties for each system and (ii) the value predictions (Figure 7.4). As predicted, the model based system dominated at all times during early learning as it always had the advantage of experience (however small) of actions and outcomes. As predicted, the only time the model free cache system had less uncertainty than the model based system, was for the distal lever press action, where the model based system required deeper iteration into the search tree. This is reflected in the top graph and the bar chart (gold bars) of Figure 7.4a, where model free caching dominated during the later runs and the distal lever press was less sensitive to devaluation than the proximal action of magazine entry.

The model also introduced increased task complexity by simulating a task with two actions for two outcomes (not shown) and this had the effect of removing the dominance of the model free caching system seen with distal reward over time. Instead both the uncertainty advantage in early training of the model based tree system and sensitivity for devaluation were retained, even for the distal lever press. The results of all the simulations confirmed their predictions that the transfer from model based to model free learning, seen during extensive or over-training, is affected not only by the proximity of action to reward, but also by task complexity. Thus, in support of their hypothesis that each of the two systems in the brain dominates under the conditions in which they are most accurate, the authors claimed that their simulations had reproduced patterns of behaviour seen in animal devaluation experiments, which could be explained by the different devaluation profiles of model based and model free reinforcement learning systems.

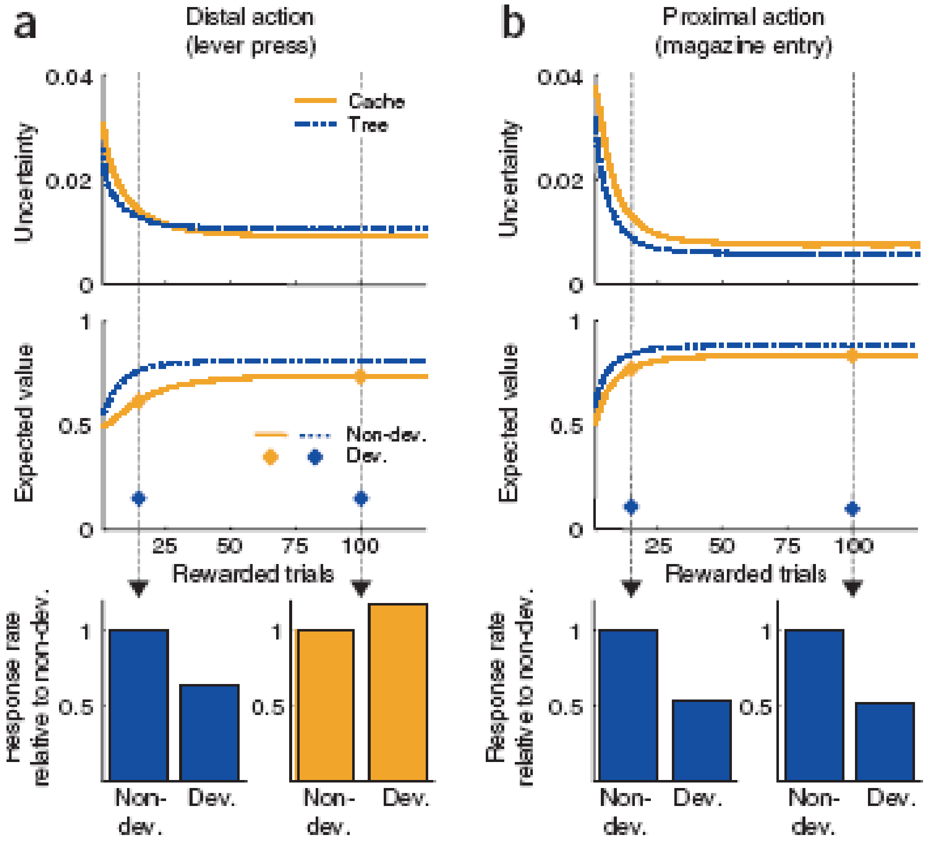


Figure 7.4 Simulation of the dual-controller reinforcement learning model in the task of Figure 7.3 (a) distal action (lever press); (b) proximal action (magazine entry). The top two graphs show uncertainties in the value estimates for different actions according to model free (blue line) and model based (gold line), as a function of the number of rewarded training trials. The middle two graphs show the value estimates; diamonds indicate the value estimates that would result after reward devaluation at various stages of training. Beneath the graphs are bar plots comparing the probability of choosing actions before and after their consequences were devalued, normalized to the non-devalued level. Bar colour denotes which system controlled the action in a majority of the 250 runs. All data reported are means over 250 runs; error bars are negligible. Taken from Daw et al. (2005).

7.4.4 Conclusions

Daw and colleagues successfully modelled controller competition between two systems simulating results from animal devaluation experiments. One of the basic assumptions of the model by Daw et al., following on from the results of lesion studies, was that the model free and model based controllers operated separately via the dorsolateral striatum and the prefrontal cortex, respectively. However, the authors pointed out that as both the striatum and cortex are innervated by dopamine and the two are interconnected via cortical/subcortical circuit loops in the brain (Chapter 4.3), a better interpretation could be that the two systems are associated with interacting dorsolateral and dorsomedial cortostriatal loops, respectively. This would involve additional interactions between model based and model free systems. Furthermore, there is limited evidence for the neural substrate of the uncertainty-based arbitration, but the anterior cingulated cortex, associated with monitoring and the resolution of response error and conflict is a possible candidate.

This was a symbolic solution, and no attempt was made to address how this may be implemented in the neurons in the brain. In Chapter 8 I present my connectionist approach to modelling the arbitration between two systems of control, which attempts to capture the distributed nature of the brain (Chapter 2.3.1).

Another limitation of the model is that Daw et al. were unable to account for the role of dopamine in the model based control system. The model based account of the role of tonic and phasic dopamine function by Smith et al. (2006), discussed in Section 7.1, has addressed this issue by modelling a tonic role for dopamine in the generation of reward, on the expression of previously acquired behaviour, independent of the acquisition process.

In the next section I describe a framework by Yin and Knowlton (2006) for the role of the basal ganglia in habit formation. Unlike the work in the previous sections, it is not a model addressing the limitations of a model free TD approach; however, it does distinguish between the goal directed actions and habit response systems referred to by Daw and colleagues, and elaborates further on the neural underpinnings of a dual system approach.

7.5 A Framework for the Role of the Basal Ganglia in Habit Formation (Yin & Knowlton 2006)

Cortico-basal ganglia networks are believed to be the fundamental basis of cerebral organisation and Yin & Knowlton (2006) posit three networks as fundamental units of function: (i) an *associative network*: for instrumental conditioning, including working memory and goal directed (GD), rewarded actions; (ii) a *sensorimotor network*: a stimulus-response (SR) system, also for instrumental conditioning, but this time for habit formation; and (iii) a *limbic network*, involved in appetitive Pavlovian learning.

The associative and sensorimotor networks differentiate between two different types of instrumental conditioning in a similar manner to Daw et al. (2005) (Section 7.4); the former for GD actions, which are explicit and controlled by consequences (which corresponds to the GD system of Daw et al.); and the latter for implicit habit formation, which is a simple SR controlled by antecedent stimuli and not guided by outcome expectancy (which corresponds to the habit system of Daw et al). In this framework Yin and Knowlton posit that the GD system operates via the hippocampus and medial temporal lobe, results from minimal training, can be applied to new situations, and can even result from a single trial. Alternatively, the inflexible SR system activates the dorsal striatum and results from overtraining.

In addition to the distinction between GD actions and SR habits, Yin and Knowlton present evidence that suggests that behavioural learning can be mapped onto distinct regions of the dorsal striatum. The dorsal striatum has long been associated with habit learning but some studies have revealed a functional dissociation between the dorsomedial striatum (DMS) and the dorsolateral striatum (DLS) (e.g. Devan & White 1999; Joel & Weiner 2000). The DMS/associative striatum/caudate appears to be associated with GD actions, while the DLS/sensorimotor striatum/putamen involves SR habits. In addition, the DMS is associated with the formation of long-term synaptic plasticity using D1 dopamine receptors; while the DLS uses D2 dopamine receptors and is associated with long-term depression of synaptic plasticity.

The roles of the PFC, the thalamus and the dopamine system must also be taken into account, and Yin and Knowlton place emphasis on the cortical/subcortical circuit loops, specifically the interactions between them (Joel & Weiner 2000) that are necessary for transforming actions into habits. The parallel circuit loop networks have strong recurrent properties and strong re-entrant projections where dopamine neurons project back to the same striatal zones that send them input enabling dopamine neurons to signal progressively earlier predictions of reinforcement, as seen in actor/critic architectures of temporal difference learning. However, as well as closed circuit network loops, there are open striatonigral projections within these loops to other nigral regions, which project to other striatal areas, resulting in transfers between the networks (Redgrave, Prescott & Gurney 1999). They posit a hierarchical organisation of the three networks where Pavlovian learning precedes instrumental learning, starting in the limbic network and spreading to the sensorimotor network via the associative network.

Yin and Knowlton refer to the shift in behavioural control from GD actions to SR habits seen in animal devaluation experiments, described in the previous section and posit that during early learning, the GD system dominates using prefrontal and parietal association cortices, DMS, associative pallidum and mediodorsal/ventral thalamus. This system is not effector specific, that is, performance does not depend on the effector (for example, a hand) with which it is originally trained. However, when full learning has taken place, a shift is made to the SR system that is more effector-specific, involving sensorimotor cortices, DLS, motor pallidum and ventral thalamus. Yin and Knowlton explain that the degree of effector-specificity reflects the level of functional integration in the hierarchical organisation of cortico-basal ganglia networks, where the associative network, which is not effector-specific, has a higher level of functional integration, with a wider range of motor programmes available for selection in order to reach a goal. Alternatively the effector-specific sensorimotor network has a lower level of functional integration. This shift from one system to the other is the process modelled in the previous section by Daw et al. (2005).

It should be noted that extensive damage to one of the networks could result in domination by the other. For example, Yin and Knowlton give an example where fMRI studies revealed that patients with Parkinson's disease, who have difficulty with SR, due to depleted dopamine stores in the striatum, are able to compensate by using their declarative memory, dependent more on the medial temporal lobe (Moody, Booenheimer, Vanek & Knowlton 2004).

Alternatively, positive symptoms in schizophrenia are believed to result from an imbalance between cortical and subcortical microcircuits leading to a dysfunction of the dopamine system (Carlsson et al. 2001; Abi-Dhargham 2004; Winterer & Weinberger 2004). The resulting imbalance may result from an insufficient inhibitory brake system leading to either a hypostimulation in the cortex of D1 receptors and a hyperstimulation in the subcortex of D2 receptors (Abi-Dhargham, 2004), or a reduced prefrontal dopamine D1/D2 receptor activation ratio, in which D2 receptors dominate, which leads primarily to a lower cortical SNR. Normally D1 receptors dominate (Winterer & Weinberger, 2004), namely the associative network. In view of the framework, the latter example would imply that schizophrenics with D2 receptor domination would rely more on the stimulus response system, the sensorimotor network, rather than the association network associated with prefrontal cortex.

7.6 Chapter Conclusions: Towards a Dual Systems Approach

7.6.1 Chapter Conclusions

In spite of the fact that model free TD reinforcement learning offers a very good account of the firing patterns of dopamine neurons (Chapters 5 and 6), it is clear that it offers an incomplete picture of the reinforcement learning account of action control in the brain for the following reasons:

- The basic TD model is unable to distinguish between different rewards with a similar value that are preceded by an appropriate CS, and between the effects of dopamine manipulation on distal rather than proximal rewards (Section 7.1).
- TD can be brittle as the prediction chains used by TD over successive time-steps break down when the CS-US relationship is unreliable (Section 7.2).
- The simple actor-critic is insensitive to motivational state and fails to take into account some of the psychological differences between Pavlovian and instrumental conditioning (Section 7.3).
- Any change in task will have to be relearned explicitly, which will take time. In reality, relearning often needs to take place quickly, so current TD models do not account for all types of learning (Section 7.4).

Having identified some of the drawbacks of using TD as the only explanation for all the behaviours seen during conditioning, the research community is currently seeking alternative approaches to compliment and extend the paradigm to account for these factors. A model based approach can address some of the problems associated with TD, but carries its own limitations (Section 7.4). An alternative is to combine the benefits of both model free and model based approaches, which offer different benefits and weaknesses in the quest for an understanding of how learning and relearning occur in the brain (Section 7.4).

In this chapter I have identified a number of models using algorithms that may be described loosely as being either model free or model based approaches. As previously mentioned in the introduction of this chapter, there is often a less clear cut distinction between them, and in reality they may be placed in a continuum between two extremes. The model of McClure et al. (2003) is the closest to being model free, although it does store a rudimentary model of the environment as a series of states and their associated values. The other models fall under the broad umbrella of being model based approaches, where tree search forms the most comprehensive model.

The model of Smith et al (2006) performs a tree search to a limited depth, Advantage (Dayan & Balleine 2002) and Q Learning (Watkins & Dayan 1992), loosely described in Sections 8.3 and 8.4, respectively, have the least representations of the environment of the models discussed in this chapter.

There is currently some debate in the literature over which algorithm better represents appetitive choice, which I touched upon in Chapter 6.6 (Morris et al. 2006; Niv, Daw & Dayan 2006). While the results of Morris et al. favour the SARSA algorithms that assign a separate value to each possible behavioural choice in every state as a better alternative in decision-making than using an actor-critic algorithm, further work is needed to find an improved algorithm that will match the anatomical data relating to the dopamine and striatal systems as well as the actor-critic (Niv et al. 2006). In the meantime, I leave this debate and instead I develop an alternative model, which is capable of incorporating both model free and model based approaches.

7.6.2 A Dual Systems Approach

In this chapter I have detailed several lines of evidence that point toward two systems of control operating together, which are summarised in Table 7.2.

Table 7.2. Studies detailed in this section offering a dual system approach. The different systems are listed (in bold), together with the possible associated neural substrate (in red).

STUDY	SYSTEM 1	SYSTEM 2
Smith, Li, Becker & Kapur (2006) (Section 7.1)	**Phasic Dopamine** for Learning Orbitofrontal Cortex, Basolateral Amygdala. OFC→Striatal→OFC Loop	**Tonic Dopamine** for Acquired Behaviour Ventral Striatum (D2 Receptors)
O'Reilly & Frank (2006) Hazy et al (2006; 2007) – PBWM Model (Section 7.2)	**Prefrontal Cortex** (Using PVLV algorithm)	**Basal Ganglia** (Using PVLV algorithm)
Dayan & Balleine (2002) (Section 7.3)	**Pavlovian Conditioning** Nucleus Accumbens	**Instrumental Conditioning** Amygdala and Orbitofrontal Cortex
Daw, Niv & Dayan (2005) (Section 7.4)	**Habit Response** and Model Free RL Dorsolateral Striatum, (Dorsolateral Cortostriatal Loop)	**Goal Directed Actions** and Model Based RL Prefrontal Cortex, (Dorsomedial Cortostriatal Loop)
Yin & Knowlton (2006) (Section 7.5)	**Habit Responses** (Sensorimotor Network) Dorsolateral Striatum	**Goal Directed Actions** (Associative Network) Hippocampus, Medial Temporal Lobe, Dorsomedial Striatum

In Section 7.1 I described how Smith et al. (2006) modelled both phasic dopamine in the learning process, and tonic dopamine in the expression of previously acquired behaviour; and in Section 7.2 I introduced the PVLV algorithm by O'Reilly et al. (2007), where the PV system controls performance and learning during primary rewards and the LV system learns about conditioned stimuli. I explained how Dayan and Balleine (2002) distinguished between modelling Pavlovian and instrumental conditioning in Section 7.3, and In Sections 7.4 and 7.5, I described accounts of the distinction between two different networks for goal directed actions and stimulus response habits by Daw et al. (2005) and Yin and Knowlton (2006).

In Chapter 8 I develop a connectionist model with dual weights which is capable potentially of implementing a dual system of control, without the need for an arbiter or a homunculus. It is important to note that while the studies in Table 7.2 all posit dual systems, they are not necessarily referring to the same systems; although there would appear to be considerable overlap between the systems of Daw et al. and Yin and Knowlton, who both refer to habit responses and goal directed actions, with similar neural underpinnings. It is not the system being modelled that is of prime importance, but the fact that the connectionist model can implement a dual system of control.

Chapter 8

A Connectionist Model of Dual System Control

In Chapter 7 I described the limitations of using model free TD to account for a reinforcement learning paradigm of action control in the brain, and referred to several lines of evidence that pointed to the concept of a dual system of control. In this chapter I develop a connectionist model with dual weights which is potentially capable of implementing two separate systems of control, without the need for an arbiter or homunculus. However, as already mentioned, it is important to note that while the studies discussed in Chapter 7 all posit dual systems; they are not necessarily referring to the same two systems (Table 7.2). With regard to the modelling in this chapter, it is not the system being modelled that is of prime importance, but the fact that a connectionist model can implement a dual system of control. However, I use the distinction between different networks for goal directed actions and stimulus response habits (Yin and Knowlton 2006) that can be modelled effectively using model based and model free learning, respectively (Daw et al. 2005) discussed in Chapter 7, Sections 7.5 and 7.4, respectively, to illustrate the workings of the model.

In this chapter I return to the connectionist approach developed in Chapter 3 to address the problem of arbitration between two systems of control. The benefits of connectionism in addressing biological systems were addressed in Chapter 2 and arise from the biologically inspired architecture which distributes knowledge across many connections between different layers of neurons. Connectionist models possess some of the advantages of the human brain, namely:

- They are distributed in nature and process information in parallel.
- They have the ability to learn from experience.
- They are able to generalise to new situations by applying information from past experience.
- They are fault tolerant and therefore resistant to damage.

Before introducing the model, I discuss in Section 8.1 a fundamental question to be addressed by any model incorporating two or more systems of control: the problem of arbitration between the different systems. In Section 8.2 I describe the artificial neural network model (ANN) by Hinton and Plaut (1987) that inspired the current model, originally designed to address the problem of catastrophic forgetting in connectionist models. An alternative use for this model is developed in Section 8.3, where my aim is to describe how a modified version of the dual weights model could be capable of implementing two systems of control, without the need for an arbiter or homunculus.

8.1 The Problem of Arbitration between Two Systems of Control

Two important questions arise from models incorporating more than one system: Which system dominates at any one time, and how is the transfer between the two systems arbitrated? In this section I refer to three bodies of research that attempt to answer this question, and show how my model can also help to address this issue.

Daw et al. (2005) (Chapter 7.4) suggested that the brain might estimate the accuracy for arbitration between goal directed and habit response systems by tracking the relative uncertainty, or expected inaccuracy of predictions, made by each of the systems. They hypothesised that each system dominated under the conditions in which they were most accurate. They suggested a Bayesian principle of arbitration according to uncertainty, so that the system dominant at any one time was the one giving the highest level of accuracy.

Yin and Knowlton (2006) (Chapter 7.5) also referred to the goal directed and habit systems for instrumental learning, which were shown to be dissociable, as extensive damage to either network could result in domination of the other. However, the two systems were also able to interact resulting in transfers between the networks. In contrast to Daw et al., they suggested a hierarchical organisation of three networks where Pavlovian learning precedes instrumental learning, starting in the *limbic network,* which plays a key role in appetitive Pavlovian learning, and spreading to the *sensorimotor network* associated with a habit response system, via the *associative network* associated with a system for goal directed actions.

In a dual system model of control Daw and colleagues reproduced patterns of behaviour seen in animal devaluation experiments, which could be explained by the different devaluation profiles of model based and model free reinforcement learning systems. However, a limitation of the model was that it assumed that the model free and model based controllers operated separately via the dorsolateral striatum and the prefrontal cortex, respectively, and did not account for interconnections between the two systems. Furthermore, there is limited evidence for the neural substrate of the uncertainty-based arbitration.

In Chapter 7.2 I referred to one of the long-term aims of Randall O'Reilly and colleagues to model the complex interactions between the basal ganglia and prefrontal cortex, where actively maintained representations in the prefrontal cortex are dynamically updated/gated by the basal ganglia (Chapter 3.2). Their PBWM model (Hazy, Frank & O'Reilly 2006; 2007; O'Reilly & Frank 2006; O'Reilly, Frank, Hazy & Watz 2007) attempts to deconstruct the homunculus by modelling the mechanisms of control between the two systems. This is a biophysically detailed connectionist model incorporating the electrophysiology of a neuron. However, as already mentioned this is a complex approach involving a sophisticated software package with complicated underlying algorithms.

The connectionist model developed in Section 8.3 is a simple model, with no need for sophisticated algorithms or software packages, which suggests a self-organising solution to the question of arbitration between two systems of control. By assuming dual weights associated with each neuron (or collection of neurons) in the network, this approach offers an online method of arbitration, without the need for an arbitrator or homunculus. This model was inspired by a connectionist network by Hinton and Plaut (1987), which is described in the next section.

8.2 Using Fast Weights to Deblur Old Memories (Hinton & Plaut 1987)

Catastrophic forgetting, or interference (McCloskey & Cohen 1989), is a fundamental limitation of a multilayer connectionist architecture, where old information, held in the weights, is forgotten or overwritten when new information is presented (Atkins 2001; Atkins & Murre 1998). This effect is not normally seen in the human brain, which typically exhibits sequential learning, where forgetting occurs more gradually over time. A major challenge for models of learning is to produce a network with the advantages of a distributed system that can accommodate sequential learning, without prior learning being disrupted by new input. Various solutions to the problem of catastrophic forgetting have been suggested, largely either by reducing the amount of overlap between input representations of the new patterns to be learned and previously learned patterns so as to effect minimal disruption to the network, or by separating new learning from old with dual-net architectures (French, 1999).

While the subject of catastrophic forgetting is not explored any further in this thesis, of particular interest to the current argument concerning dual control is a connectionist model by Hinton & Plaut (1987) who attempted to 'deblur' a set of previously learned associations, when they had been 'blurred' as a result of the subsequent learning of a second set of associations. The model was based on the idea that changes in synaptic efficacy at a single synapse occur at many different timescales, some rapid and some more slowly (Kupferman 1979; Hartzell 1981), and can be represented in an ANN by giving each connection several different weights that change at different speeds.
In this ANN each artificial neuron had two weights for each connection that summed together to form the total weight on each connection: one set of fast, elastic weights, and another set of slow, plastic weights (Figure 8.1). The set of fast weights took on values that temporarily cancelled out the adverse effects of new learning on old and rehearsing a subset of old information was shown to be sufficient to achieve this. The fast weights were the novelty; with a higher learning rate they were quicker to reflect change, and constituted the more recent past due to the fact that their weights rapidly decayed towards zero by some fraction, *h*, after each weight change. Essentially the fast weights, when significant at times of surprise or novelty, were seen to provide a temporary context or associative memory in addition to the knowledge in the slow weights. The slow weights were typical of the normal single weights usually seen in ANNs; changing slowly and holding the long term knowledge of the network. This method exploited the distributed nature of an ANN to recall previously stored patterns.

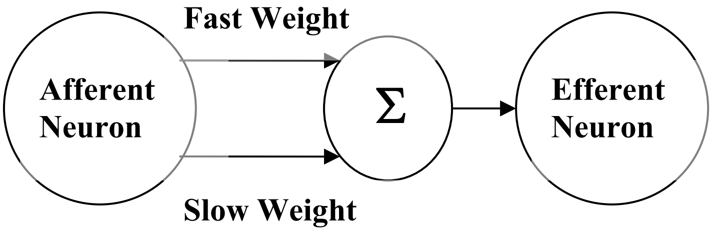


Figure 8.1. Each artificial neuron in the ANN of Hinton and Plaut (1987) had two weights for each connection, one fast and one slow, that summed together to form the total weight on each connection.

By implementing a dual weight system, Hinton and Plaut demonstrated that an additional set of fast weights were temporarily able to cancel out the interference in a set of old associations caused in more recent learning and it was possible to quickly restore a whole set of old associations by rehearsing on just a subset of them (Figure 8.2). Other studies have found a similar trajectory of recovery for unlearned items by rehearsing on just a few, using a variety of multilayer connectionist models of memory (e.g. Atkins 2001; Atkins & Murre 1998; Hinton & Sejnowski 1986a; Plaut 1996; Hinton & Shallice 1991), but the dual weight model of Hinton and Plaut is of particular interest in this thesis because of the interactions seen between the fast and slow weights during learning and relearning.

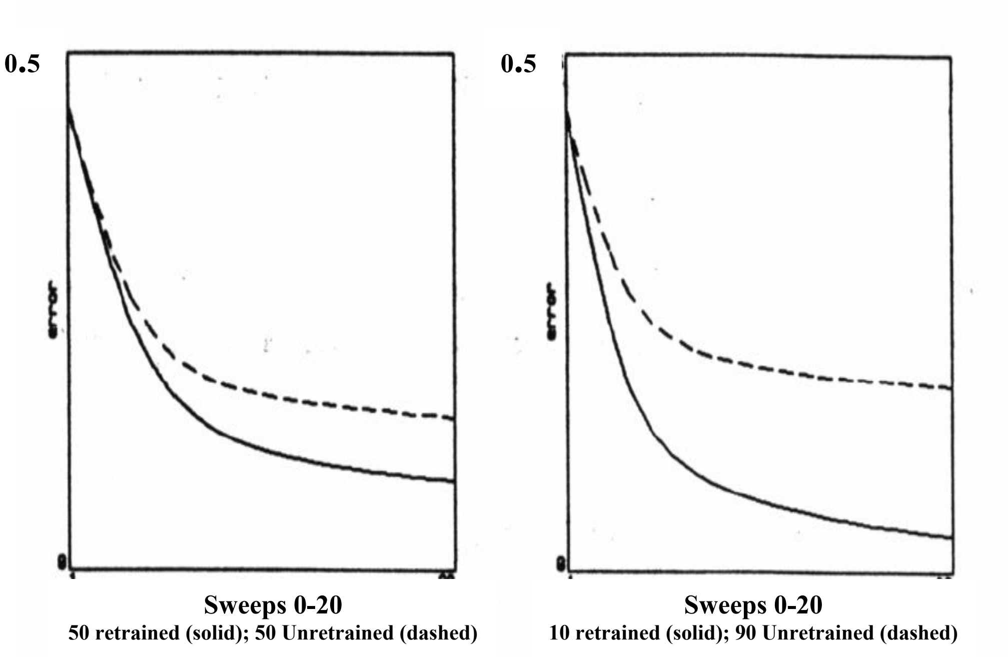


Figure 8.2 Figure taken from Hinton & Plaut (1987) comparing the errors of both retrained and unretrained data for the first 20 sweeps. When the network was retrained on 50% and 10% of the old associations (solid lines), it was found that in the early stages of retraining the improvements in the associations that were not retrained (dashed line) were nearly as good as the associations that were explicitly retrained.

Hinton and Plaut suggested that the temporary memory could be used in the following ways:

- Rapid temporary learning: It is even capable of storing a new association in one trial. For example, one-shot goal directed learning referred to in Daw et al. (2005).
- Creating temporary bindings between features.
- Recursive processing: by temporarily storing local variables in the fast weights and calling on sub-procedures. The local variables can then be restored from this temporary associative memory (McClelland & Kawamoto 1986).
- To implement the 'shortest descent' learning method i.e., minimising the amount of interference caused by new learning.

In the next section I describe my implementation of a similar model to Hinton and Plaut using dual weights to investigate the first of these suggested uses: rapid temporary learning.

8.3 A Connectionist Model of Dual System Control

While Hinton and Plaut originally designed the dual weights model to investigate the problem of interference of new information on old, in this Chapter I aim to use a modified version of this model to address rapid temporary learning and the problem of arbitration between two systems of control.

I posit that the model can be thought of as consisting of two systems that can either operate independently or interact: one system for the fast weights, and the other for the slow weights. Following this line of reasoning, Hinton and Plaut used the fast 'elastic' weight system to temporarily cancel out the interference in a set of old associations caused by more recent learning. The slow 'plastic' system was dominant when learning had taken place and the whole system was stable, but the fast weights came into operation when errors occurred as a result of introducing a change in the data set. It is the interaction between the two systems that is of interest in this thesis, which may suggest a solution to the problem of arbitration between any two systems of control.

The model described in this Chapter could refer to control between any number of dual systems, and, as already mentioned at the beginning of this chapter, it is not the system being modelled that is of prime importance, but the fact that this connectionist model can implement a dual system of control. However, in order to illustrate the use of this model I refer to the model encompassing model free and model based reinforcement learning by Daw et al. (2005), discussed in Chapter 7.4 as an example of two systems of control that could be equated by the model, where:

- Fast weights could equate with the tree-search model based system that is capable of flexible, often one shot goal directed actions.
- Slow weights could equate with the model free TD-like system associated with inflexible stimulus response actions that require extensive relearning.

A dual weighted system with fast weights to highlight past experience and identify context offers a novel approach, with the possibility of combining both goal directed and stimulus response systems in one simple model.

As discussed in Section 8.1 this is a simple model, with no need for the sophisticated algorithms or software packages used by O'Reilly and colleagues. Like Daw et al. (2005) it models two different systems of control, but unlike Daw et al. it allows for interactions between the two systems. By assuming dual weights associated with each neuron (or collection of neurons) in the network, this approach offers an online method of arbitration, without the need for arbitration (Daw et al. 2005) for which there is limited neural evidence.

Hinton and Plaut used a large training set of 100 random associations and had a large hidden layer of 100 units in the network, resulting in many modifiable connections. One unit per association would allow for perfect learning very easily by assigning one unit to each of the associations, but this approach would result in a network with a limited ability to generalise. This practice was typical of early ANNs, where models tended to be over resourced with too many free parameters, and a similar situation was encountered with the model of Hoffman and McGlashan (1987) described in the discussion of Chapter 3.1.1.

The aim of this research was not just to perform a simple replication of Hinton and Plaut, but to constrain the network in order to investigate both the contribution of the fast weights, and the interaction between fast and slow weights, in greater detail. In accordance with current standard practice when investigating the performance of an ANN I sought the lowest number of hidden layer units that would train the network. A smaller network than the one used by Hinton and Plaut would still learn the associations with minimal error, but a solution would not be reached as readily as a network with one hidden unit per association; resulting in the knowledge of the network being more widely distributed and a repeatable experiment, with a greater ability to generalise. In order to achieve this I performed a pilot study to determine the minimum number of hidden units necessary to train the associations in Phase 1. I used a smaller data set in Phase 1 than Hinton and Plaut of 20 as opposed to 100 associations, while the Phase 2 data set contained the same 5 new random associations as the original study, and none of the original 20 associations presented in Phase 1. The effect of the smaller data set in Phase 1 and of keeping the data set in Phase 2 constant will result in a greater amount of perturbation, or disturbance to the Phase 1 associations than seen in Hinton and Plaut, i.e., 25% perturbation (5 to disrupt 20) in this study, compared to 5% perturbation (5 to disrupt 100) in the original study. I was then in a position to investigate the parameters of the model: the learning rates and momentum for both the fast and slow weights, as well as the rate of decay for the fast weights

Such constraints to the model will allow a detailed investigation of the contribution of the fast weights. In addition, it will reveal how the two sets of weights interact in the model and under what conditions the interactions occur. If the fast and slow weights can be thought of as two separate systems of control capable of operating separately and interacting, as suggested at the beginning of this section, then the interaction may provide an insight into how other systems of control may coexist and interact, without the need for an explicit arbiter or homunculus.

8.3.1 Methods

The Model

The model was implemented in a version of a standard multilayer backpropagation network with momentum (Rumelhart et al. 1986), originally written in C Program by Pao (Pao 1989) and modified by a member of the research team to include the addition of dual weights, where each artificial neuron in the ANN had two weights for each connection that summed together to form the total weight on each connection as detailed in Figure 8.1.

The Network Architecture

The network consisted of 10 outputs, 10 inputs and a hidden layer, in accordance with Hinton and Plaut. The network of Hinton and Plaut contained 100 hidden units; however, in this study the number of hidden units was reduced to 7 in order to constrain the network. This value was determined by a pilot study (see Appendix VI A).

Decay Rate for fast Weights

The fast weights were the novelty:

- With a higher learning rate they were quicker to reflect change and the more recent past.
- They were also quicker to forget due to the fact that their weights rapidly decayed towards zero by some fraction, *h*, after each weight change.
- The default rate for the decay was 0.999 where the fast weights decayed at a rate of 0.1% with percentage retention of 99.9%, following every weight change. Details of the effects of using different rates of decay on learning and relearning can be found in Appendix VI C.

NB The slow weights were typical of the weights usually found in an ANN. Having a lower learning rate they were slower to change, and because they did not decay (forget), they held the long-term knowledge of the network.

Training and Testing

As in Hinton and Plaut, the training sets were selected at random (without replacement) from a possible set of 2^{10} ten bit binary vectors (1024 possibilities), but instead of using values of zero and one, I used 0.1 and 0.9, respectively. Using these values for the outputs will allow for better convergence of a network using gradient descent methods, where the error feedback is derived from a sigmoid curve that never actually reaches zero or 1.

I used a smaller data set in Phase 1 than Hinton and Plaut of 20 as opposed to 100 associations, while the Phase 2 data set contained the same 5 new random associations as the original study, and none of the original 20 associations presented in Phase 1. The effect of the smaller data set in Phase 1 and of keeping the data set in Phase 2 constant will result in a greater amount of perturbation, or disturbance to the Phase 1 associations than seen in Hinton and Plaut, i.e., 25% perturbation (5 to disrupt 20) in this study, compared to 5% perturbation (5 to disrupt 100) in the original study.

As in the original study training was carried out in three phases:
Phase 1: 20 random input vectors were associated with 20 random output vectors.
Phase 2: Five new random associations were presented to the network without rehearsing on the original 20.
Phase 3 (testing phase): The network was retrained on 10 of the original data (50%) and the improvement in performance of the retrained subset was compared to the 10 associations in Phase 1 that had not been retrained. As in Hinton and Plaut the model was able to perform a test on the unretrained associations by plotting the total error for all the output units for all the associations, sweep by sweep, and comparing them to the total errors on the retrained associations.

Simulations
I began with a pilot study where I used the slow weights only to find the optimum number of hidden layer units that would still train the 20 associations during Phase 1 in order to constrain the network and maximise generalisation. I also sought maximum and minimum learning rates for the slow weights in order to find the optimum learning rate parameters for future simulation using both fast and slow weights. I was then in a position to investigate the findings of Hinton and Plaut (1987) using a constrained version of the original model where fast weights were utilised to temporarily capture old learning. However, analysis of the interactions between fast and slow weights revealed that the value of decay rate selected for the fast weights had important bearings on the transfer of knowledge from fast to slow weights and I sought an optimal combination of rates of decay over the three phases in order to investigate the contribution of the decay rate parameter. I also investigated the contribution of the fast weights by running a series of control experiment using only the slow weights.

My investigations into the findings of Hinton and Plaut (1987) using a constrained version of the original model where fast weights were utilised to temporarily capture old learning are detailed in Appendix VI, together with details of the pilot study, the contribution of the decay rate parameter, and the contribution of the fast weights. I summarise these results in the next section.

8.3.2 Summary of Results Detailed in Appendix VI

Using Fast Weights to Temporarily Capture Old Learning (Appendix VI B)
Using a constrained dual weighted network with a reduced hidden layer, together with a data set that caused a greater amount of perturbation during Phase 2, I was able to recreate a result comparable to that of Hinton and Plaut (1987), where an improvement was seen in the unretrained associations from Phase 1 when retraining on only 50% of the Phase 1 associations during Phase 3 (Figure 8.3). NB error (*y*-axis) is the root means square error of difference between the actual response and the target response for each neuron, hereafter referred to as error.

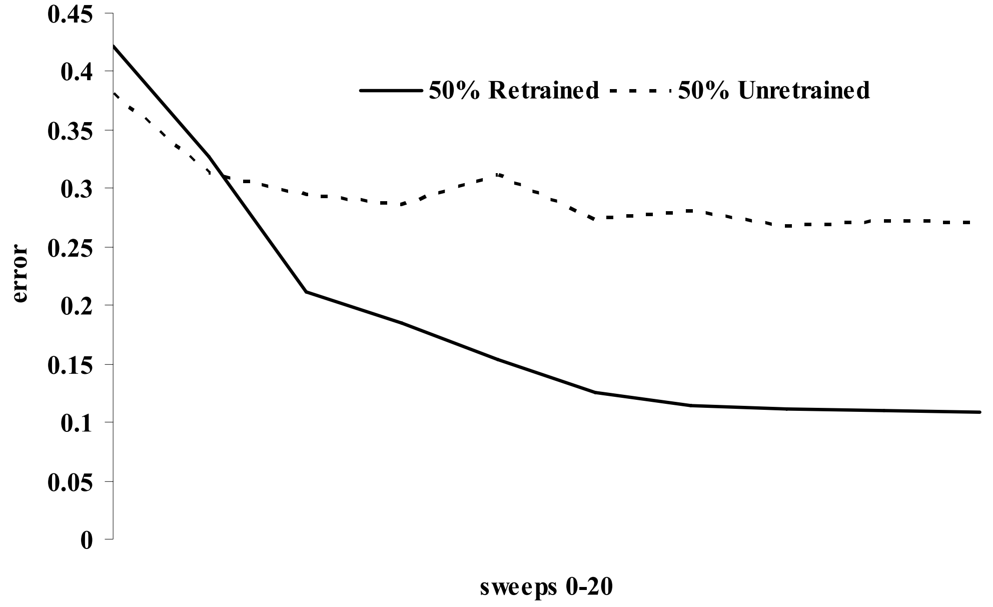


Figure 8.3 When the constrained network was retrained on 50% of the old associations (solid line) during Phase 3, it was found that in the early stages of retraining (first 20 sweeps) there was an improvement in the associations that were not retrained (dashed line).

This was not a direct replication of Hinton and Plaut and so there was less of an improvement on the 50% unretrained associations than was seen in the original study (Figure 8.2). However, my simulations captured the effect identified by Hinton and Plaut and I demonstrated that when the network was retrained on a subset of the original data it was found that in the early stages of retraining improvements were seen in the associations that were not retrained. This was because the knowledge of the original associations was distributed over many connections and retraining some of the associations pushed back the weights of the others to the point in time before the perturbation occurred. The fast weights were able to cancel out the interference in a set of old associations caused in more recent learning and it was possible to quickly restore a whole set of old associations by rehearsing on just a subset of them. The fast weights created a context in which the old associations were present again, without permanently interfering with the new associations, as the new knowledge was restored when the fast weights decayed back to zero.

The Contribution of the Decay Rate Parameter (Appendix VI C)

The simulations in this experiment demonstrated that the interaction between the fast and slow weights is enabled by the decay rate for the fast weights where the weights rapidly decay towards zero by some fraction, *h*, after each weight change. It was clear that tasks with different learning complexities will require different decay rates for the fast weights in order for the knowledge of the task to be transferred to the slow weights by the time the stopping criterion has been reached. It would appear that the decay rate of the fast weights is a critical factor in the interaction between the fast and slow weights in this model: A fast decay rate (0.99) is appropriate for tasks with a lower complexity that train in a minimal number of sweeps where the contribution of the fast weights is minimal, such as in Phase 2 where the network has to learn only 5 new associations; and a slower decay rate (0.999) is more appropriate for tasks with a higher complexity that take longer to train and are facilitated by the fast weights, such as in Phases 1 and 3. Accordingly the optimal combination for decay rates over Phases 1, 2 and 3 was found to be 0.999-0.99-0.999 for this particular combination of tasks.

The Contribution of the Fast Weights (Appendix VI D)

In order to reveal the contribution of the fast weights in a dual weighted system two control conditions were implemented, where progression was made through the three phases using slow weights only using learning rates of either 0.001 or 0.41, being the optimum learning rate values for slow and fast weights determined in the pilot study. By comparing the two control conditions to a simulation using fast and slow weights with the optimal combination of decay rates of 0.999-0.99-0.999 determined in Section B above, it was found that the fast weights in all three phases make a substantial contribution to the early stages of learning compared to a network using slow weights only. However, while the addition of the fast weights provides rapid learning during the initial stages of training, it does so at the expense of the number of sweeps, or time taken to complete the task. A dual weighted network is trying to converge on a solution using two sets of weights, which is a much harder problem than a standard network with a single set of weights. Thus, the price to pay for early rapid learning is the time taken for the network to converge on a solution.

As an example of the contribution of the fast weights during the initial stages of learning Figure 8.4 shows the effect of plotting the total error across all output units and patterns against number of sweeps through the Phase 1 data set for:

- A dual weighted network:
    - (i) Fast weights error (green line)
    - (ii) Slow weights error (pink line)
    - (iii) Overall error of the system (dark blue line)
- A network with just slow weights:
    - (iv) Control condition error using learning rate of 0.001 (red line)
    - (v) Control condition error using learning rate 0.41 (light blue line)

Similar effects were seen during Phases 2 and 3 (Appendix VI C).

Firstly, with regard to the dual weighted network, it can be seen from Figure 8.4 that learning in the early stages is rapid in the fast weights, reflected as a dramatic decrease in the fast weights error during the first 5,000 sweeps (green line). Learning is much slower initially in the slow weights (pink line), but as the overall error (dark blue line) declines with learning and the fast weights decay, knowledge is transferred to the slow weights, which is reflected in the error for the slow weights that begins to decline more rapidly than in the initial 5,000 sweeps. By around 20,000 sweeps through the data set the slow weights begin to dominate, as the error falls below that of the fast weights. Changes to the slow weights are slow and steady as a result of the low learning rate, while changes to the fast weights are larger and more volatile as a result of the higher learning rate. The volatility of the fast weights is reflected in the overall error.

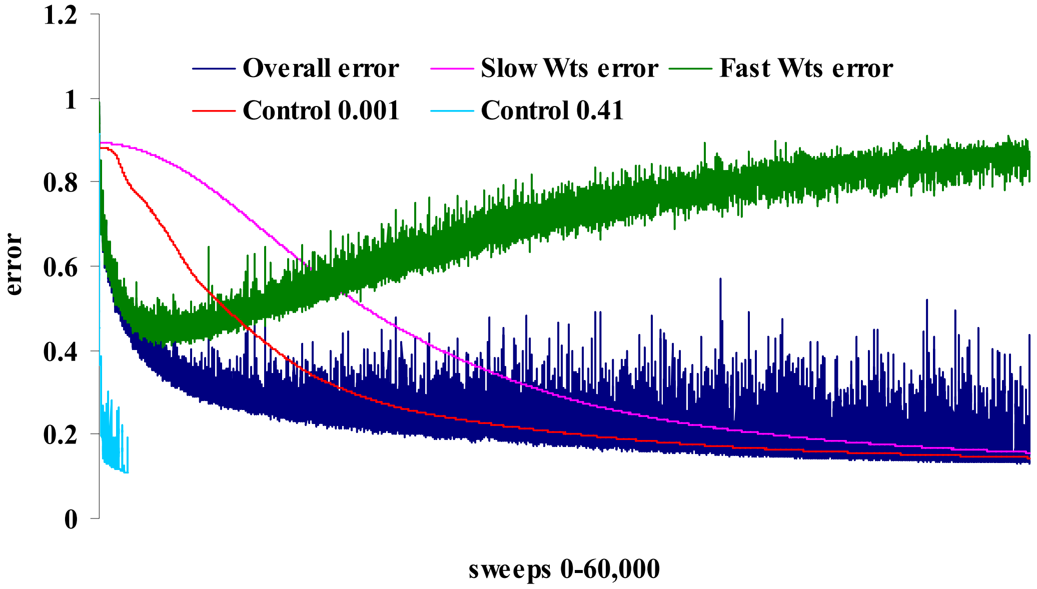


Figure 8.4 An example of the contribution of the fast weights during learning - Phase 1: The errors of the fast (green line), slow (pink line) and total weights (dark blue line) for a dual weighted network compared to the total weights of two control conditions using slow weights only with learning rates of 0.001 (red line) and 0.41 (light blue line). The fast weights make a substantial contribution during the early stages of learning.

By comparing the overall error of a dual weighted network (dark blue line) to the two control conditions using slow weights only, it can be seen that Control 0.001 (red line) is slower to learn during the early stages, while Control 0.41 (light blue line) learns almost immediately. A closer inspection of the errors for the fast and slow weights in the dual weighted network reveals that it is the fast weights that contribute to the enhanced learning during the initial stages using a dual weighted network. While the combination of fast and slow weights does not learn as fast as using a learning rate of 0.41 alone, such a high training rate often leads to error and will not always find a solution for the network, unlike a very slow learning rate, and so the combination of fast and slow weights offers an alternative to using an error-prone high learning rate only.

Conclusions

While I have substantiated the claims of Hinton and Plaut, additional analysis of the contribution of the fast weights and of the interactions seen between fast and slow weights has enabled me to suggest an alternative use of the dual weighted ANN: as a dual system of control, capable of addressing the problem of arbitration between two systems of control. In addition the fast weights can be seen as a temporary memory capable of rapid temporary learning. My arguments are detailed in Section 8.4.

8.3.3 Discussion

Using an ANN with both fast and slow weights, it was found that when 50% of a set of the Phase 1 associations were retrained, during Phase 3, improvements were seen in the other 50% of the Phase 1 associations that were not retrained, during the early stages of training. This was in accordance with the findings of Hinton and Plaut (1987).

However, this was not a direct replication of Hinton and Plaut as I used a constrained model with a reduced hidden layer. I also used a smaller data set in Phase 1 than Hinton and Plaut of 20 as opposed to 100 associations, while the Phase 2 data set contained the same 5 new random associations as the original study, resulting in a greater amount of perturbation, or disturbance to the Phase 1 associations than seen in Hinton and Plaut, i.e., 25% perturbation (5 to disrupt 20) in this study, compared to 5% perturbation (5 to disrupt 100) in the original study. The results were achieved due to rapid learning through the addition of fast weights and, as well as supporting the findings of Hinton and Plaut, they provide the following additional analysis:

- The network was a constrained version of that used by Hinton and Plaut, with a reduced hidden layer, where 20 associations were learned using 7 hidden units during Phase 1, as opposed to 100 associations using 100 hidden layer units. This will result in a more distributed network than Hinton and Plaut, who provided enough resources for one hidden unit per association.
- I also used a smaller data set in Phase 1 than Hinton and Plaut of 20 as opposed to 100 associations, while the Phase 2 data set contained the same 5 new random associations as the original study, and none of the original 20 associations presented in Phase 1. The effect of the smaller data set in Phase 1 and of keeping the data set in Phase 2 constant will represent a greater amount of perturbation, or disturbance to the Phase 1 associations than seen in Hinton and Plaut.
- This study also provided more control condition analysis than the original study which showed the contribution of the fast weights during early learning/relearning. However, this was at the expense of the number of sweeps, or time taken to complete the task. Thus, the price to pay for early rapid learning is the time taken for the network to converge on a solution.

Hinton and Plaut explained that the effects seen were achieved as the knowledge of the original associations was distributed over many connections. Retraining on some of those associations pushed back the weights to the point in time before the disturbance to the weight values occurred, which was due to the learning of the new associations in Phase 2. The fast weights were able to cancel out the interference in a set of old associations caused in more recent learning and it was possible to quickly restore a whole set of old associations by rehearsing on just a subset of them. The fast weights created a context in which the old associations were present again, without permanently interfering with the new associations, as the new knowledge was restored when the fast weights decayed back to zero.

In the next section I detail my arguments for an alternative use for a model with dual weights: As a dual system of control, capable of addressing the problem of arbitration between two systems of control, where the fast weights can be seen as a temporary memory capable of rapid temporary learning.

8.4 The Mechanisms behind the Dual System of Control

Hinton and Plaut used the dual weights model to address the problem of catastrophic forgetting in multilayer ANNs. However, I suggest an alternative use for a model with fast and slow weights to address the problem of arbitration between two systems of control through rapid learning in the fast weights which will allow for more rapid changes in the environment than in a standard network with one set of weights.

The knowledge of an ANN is held in the weights, which is analogous to the synapses in a biological system, and learning is achieved by changing the strengths of the weights in the system. If the response of an output unit is incorrect then the network can be changed so that it is more likely to produce the correct response in the future. This is achieved by changing the weight of each connection by a proportion of the error arising from the difference between the actual and desired output. The proportion of change to the weights is determined by the learning rate and the higher the learning rate, the greater the amount of change. In the dual weights model both fast and slow weights sum together to form the total weight on each connection, and as both are affected by the same network error, the knowledge of the system at any one time is determined by both sets of weights.

In a well learned situation the body of knowledge normally resides in the slow weights in the dual weights model. Here, the error in the system is low and any errors in the fast weights quickly decay to zero. As there is no decay rate associated with the slow weights the knowledge of the system is maintained within these connections. However, when the error in the system is high the fast weights will dominate, for example as a result of the presentation of the previously unseen associations during Phase 2, or the reintroduction of some of the associations from Phase 1 during Phase 3. Here, the higher learning rate of the fast weights will result in a larger update to the fast weights than to the slow weights with a lower learning rate and a smaller weight update, thus more of the new knowledge will be held in the fast weights. The overall effect will be that the fast weights learn the new associations more quickly than the slow weights, which are slower to learn and effectively hold on to the old associations for longer. The larger the values of the fast weights, the longer the fast weights will take to decay, resulting in the temporary domination of the fast weights during periods of learning/surprise in the system. Thus the fast weights can be seen to be acting as a temporary memory capable of rapid temporary learning, such as in the goal directed learning referred to by Daw et al. (2005).

As the network continues to learn and errors in the system become lower, the updates to the fast weights are lower and the rate of decay of the fast weights begins to dominate over this update. The slow weights are gradually adjusting to the new associations and an interaction is seen between fast and slow weights as the slow weights begin to dominate by holding the knowledge of the system. By the time the stopping criterion has been reached and learning is complete, effectively knowledge has been transferred to the slow weights, as the small errors in the fast weights are dominated by the tendency of the fast weights to decay towards zero.

The rate of decay for the fast weights will have an important bearing on the temporary domination of the fast weights as it is a critical factor in the interaction between the fast and slow weights in the model: the quicker the rate of decay, the lower the contribution of the fast weights; while a slower rate of decay will allow the fast weights to dominate for a longer period and this is particularly beneficial during tasks of a higher complexity requiring a higher number of sweeps (Appendix VI B).

Hinton and Plaut used the fast weights to cancel out the interference in a set of old associations caused in more recent learning, where the fast weights created a context in which the old associations were present again, without permanently interfering with the new associations. However, I suggest a wider application for a dual weighted network:

- In a dual weighted network a set of fast weights will allow for more rapid changes in the environment than in a standard network with one set of weights, by reacting quickly to new information or change/surprise when the error in the system is high. This will provide a temporary overlay of the new knowledge to the existing body of knowledge built up and stored in the slow weights, and provide time for the new knowledge to be assimilated into the old.
- But in a fully learned situation with minimum error or change/surprise the slow weights bearing the existing body of knowledge will dominate as the fast weights decay towards zero.

This suggestion has similarities to that of Daw et al. (2005) to explain the different devaluation profiles of model based and model free reinforcement learning systems, described in Chapter 7.4. Daw and colleagues demonstrated that a model free habit system of caching values is not immediately sensitive to the specific outcome information associated with the devaluation, and it will take time for a change in behaviour to occur following relearning of the values. Alternatively, a flexible model based system that is outcome sensitive and goal directed, will show an immediate behavioural change.

Accordingly, to illustrate the use of such a dual weights model, I suggest that the fast and slow weights could be mapped on to the stimulus response and goal directed systems identified by Daw et al. as follows:

- **Fast weights** could equate with the tree-search model based system that is capable of flexible, often one shot goal directed actions. However, this is at the expense of being expensive in terms of memory and time.
- **Slow weights** could equate with the model free TD-like system associated with inflexible stimulus response actions that require extensive relearning, but suffers from inflexibility.

In the functional specialisation of the striatum both Daw et al. (2005) and Yin and Knowlton (2006) suggest that goal directed actions are associated with cortical/subcortical loops involving dorsomedial striatum; while habit responses involve loops via the dorsolateral striatum. Thus the fast and slow weights could be seen to represent areas of the dorsomedial and dorsolateral striatum, respectively.

Daw and colleagues claimed that accuracy was the key to arbitration between the two systems, which would have different benefits and weaknesses in different situations. The system that dominates will depend on inexperience and task complexity versus search depth. The Model based system should dominate early in training, and during complex tasks, where data is scarce and stored TD values are unreliable; but model free systems should dominate during a well learned task, when habits prevail for actions more distal from reward. Here, the deeper search is costly and more error prone due to the necessity of pruning and exploring a limited set of paths in an effort to cut down on the extensive search space in a tree search.

This is precisely what has been demonstrated using the dual weights model: the fast weights dominated when the error in the system was high, early in training and when new associations were introduced to the network during Phase 2; while the slow weights dominated when the error in the system was low and the task was well learnt and more habitual.

Behavioural studies and evidence from fMRI point to two systems of control operating from different areas of the striatum (dorsomedial versus dorsolateral) (Daw et al. 2005; Yin & Knowlton 2006). However, this dual weighted model has shown emergent properties from a single system with dual weights representing two subsystems, via fast and slow weights that are contained within the same set of neurons. This suggests the possibility of two or more interacting systems incorporated into the same neural material, permitting transfers between the two systems.

8.5 Advantages of the Dual System of Control

While Daw and colleagues successfully modelled controller competition between two systems by simulating results from animal devaluation experiments, I identified four criticisms of the model, previously detailed in Chapter 7.4.4, namely: (i) the model did not account for interactions between the two model free and model based controllers; (ii) control between the two systems involved uncertainty-based arbitration for which there is limited neural evidence; (iii) this was a symbolic solution, and no attempt was made to address how control could be implemented in the neurons in the brain; and (iv) they were unable to account for the role of dopamine in the model based control system.

The dual weights model attempts to address the first three criticisms of the model by Daw and colleagues. Firstly, it allows for an interaction between the two systems via the decay rate parameter. Yin and Knowlton (2006) place emphasis on the cortical/subcortical circuit loops in the brain and in particular the interactions between them necessary for transforming actions into habits. The dual weights model could represent the shift in behavioural control from goal directed actions (using fast weights) to habit responses (slow weights) seen in animal devaluation experiments with overtraining, posited by Yin and Knowlton via the *associative network* involving dorsomedial striatum to the *sensorimotor network* involving dorsolateral striatum (Chapter 7.5). Common to both loops are the open striatonigral projections to other nigral regions, which project to other striatal areas, resulting in transfers between the networks (Redgrave, Prescott & Gurney 1999).

Secondly, the self-organising nature of the dual weights model does away with the need for the Bayesian arbitrator (Daw et al. 2005) or homunculus (but see Section 8.6 for future directions and a discussion of the decay rate parameter). Like Yin and Knowlton (2006) the dual weights model suggests a hierarchical transfer between two systems of control, where the associative network for goal directed actions dominates early in learning, with a higher level of functional integration and a wider range of motor programmes available for selection in order to reach a goal (represented by the fast weights). However, as learning progresses a shift is seen to a lower level of functional integration associated with habit formation and the effector-specific sensorimotor network (represented by the slow weights).

Thirdly it is a biologically inspired connectionist application possessing some of the advantages of the human brain and, as a more emergent model of two systems of control; it is in a better position to suggest how control may be exerted in the brain. Furthermore it does so without need for the sophisticated algorithms or software packages of O'Reilly and colleagues (e.g. Chapter 7.2, O'Reilly, Frank, Hazy & Watz 2007; O'Reilly and Munakata, 2000).

To summarise, a dual weighted system with fast weights to react quickly to change/surprise in the environment (this study), capable also of highlighting past experience and identifying context (Hinton & Plaut 1987), offers a novel approach and may have important implications in addressing how both goal directed and stimulus response learning may coexist.

8.6 Future Directions

The model in its present form is a very simple ANN with dual weights capable of implementing a dual system of control without the need for an arbiter or a homunculus, and as such possesses a number of limitations. However, there are many potential applications of the model and in this section I outline a possible solution to the limitations and give an indication of some of those applications.

The detailed investigations in Section 8.3 and Appendix VI have shown that although the model is capable of self arbitration by allowing interactions between the two systems (modelled by fast and slow weights) there are three main parameters of the model which require special consideration: the learning rates for each of the fast and slow weights (coupled with momentum); and the decay rate for the fast weights. These parameters have been shown to be task dependent and it will be necessary to find optimal values for each of these prior to any detailed investigation pertaining to the task, calling into question the self organising ability of the model.

In particular, it was clear that tasks with different learning complexities will require different decay rates for the fast weights in order for the knowledge of the task to be transferred to the slow weights by the time the stopping criterion has been reached (Appendix VI B). It would appear that the decay rate for the fast weights is a critical factor in the interaction between the fast and slow weights in this model:

- The faster the decay rate, the smaller the temporary window of opportunity provided by the fast weights to provide a temporary context or associative memory for the recent past (Hinton & Plaut 1987) or provide a temporary overlay of new learning, where the fast weights learn the new associations more quickly than the slow weights, which are slower to learn and effectively hold on to the old associations for longer (this study). A faster rate of decay is appropriate for tasks with a lower learning complexity that train in a minimal number of sweeps.
- The slower the decay rate, the longer the temporary window of opportunity provided by the fast weights. This is more appropriate for tasks with a higher learning complexity that take a long time to train.

Introducing dopamine into future versions of this model may obviate the need to set the parameters of the model prior to each task. Other computational models have made use of the phasic dopamine burst to gate salient or significant information into working memory, for example Braver et al. (1999) (Chapter 3.2) and earlier models by O'Reilly and colleagues (Rougier et al. 2005). A phasic dopamine reward prediction error, modelled as a TD error, representing significance/surprise (Chapter 4.2.1, Figure 4.2) could affect the modifiable parameters of the model in the following ways: (i) by changing learning rates; (ii) by changing the decay rate for the fast weights, or (iii) a combination of (i) and (ii). This could be achieved as follows:

- A large dopamine phasic burst above baseline firing rate (Figure 4.2A), resulting from unexpected surprise/significance, signifying that things are better than expected, could: (i) switch to a higher learning rate in order to allow for faster learning in the system; (ii) switch to a slower decay rate for the fast weights in order to provide a longer temporary window of opportunity for the fast weights to bear the new information, giving the slower weights the opportunity to assimilate the new information; or (iii) a combination of (i) and (ii).
- In the absence of surprise/significance, where things are just as expected and dopamine is firing at baseline levels (Figure 4.2B), there will be no dopamine phasic burst and no new learning. Here the default parameters of the model would dominate, namely: low learning rates and a high rate of decay for the fast weights.

However, this explanation currently does not account for when dopamine neurons are firing below baseline, such as when an expected reward fails to arrive (Figure 4.2C).

While the addition of the fast weights provides rapid learning during the initial stages of training, it does so at the expense of the number of sweeps, or time taken to complete the task (Appendix VI C), so the price to pay for early rapid learning is the time taken for the network to converge on a solution. A dual weighted network is trying to converge on a solution using two sets of weights and therefore two sets of free variables, which is a much harder problem than a standard network with a single set of weights and only one set of free variables. A dopamine controlled system will allow salient information into the system via the fast weights following a switch to either a higher learning rate, or a slower decay rate, or a combination of the two. This will result in a system with two sets of free variables (fast and slow weights) sensitive to rapid learning that will take longer to converge on a solution. However, non salient information will not produce a phasic dopamine burst and there will be no switch of parameters. Here, the fast weights will be close to zero as a result of their decay and only one set of free variables will dominate (slow weights), resulting in a more efficient system.

Other applications could be to apply the model to some of the other dual systems identified in Chapter 7, Table 7.2. In particular, the phasic dopamine neuron firing seen in learning (System 1) and the tonic dopamine relating to the expression of previously acquired behaviour (System 2) identified by Smith et al. (2006) and described in Chapter 7.1. A dual weighted ANN with fast weights representing phasic dopamine and slow weights representing tonic dopamine could potentially accommodate both systems within one model to address one of the limitations of the basic TD model, the inability to distinguish between the effects of dopamine manipulation on distal rather than proximal rewards (Daw et al. 2005; Smith et al. 2004; 2006).

In this chapter I have provided a theoretical argument for the implementation of an ANN with dual weights. However, in order to demonstrate the value of the model it will be necessary to test the theory using psychological tasks of sequential learning that will exploit two systems of control. Furthermore, there are clearly many systems of control operating in parallel in the brain and future versions of the model could accommodate multiple systems by having multiple weights associated with each artificial neuron (or set of neurons).

Finally, I would reiterate the limitations of modelling originally addressed in Chapter 2. The model is informative rather than definitive; qualitative rather than quantitative, and as such cannot be predictive. However, the usefulness of this model lies in its relative transparency and the ability to suggest what might be reasonable lines for research. Therefore I am not claiming that this is actually what happens in the brain, but simply showing emergent properties from a single system with dual weights representing two subsystems, via fast and slow weights. Further testing will determine the limitations of a dual weighted system of control.

8.7 Chapter Conclusions

Hinton & Plaut (1987) originally devised a dual weighted model for the purpose of addressing catastrophic forgetting, using a set of fast weights to cancel out the interference in a set of old associations caused in more recent learning, where the fast weights created a context in which the old associations were present again without permanently interfering with the new associations. While I have substantiated the claims of Hinton and Plaut using a constrained version of the model and a greater degree of perturbation to the old associations (Appendix VI B), an analysis of the contribution of the fast weights (Appendix VI D) and of the interaction between fast and slow weights (Appendix VI C) has allowed me to suggest an alternative use for a dual weighted ANN (Section 8.4).

I posit that the model can be thought of as consisting of two systems that can either operate independently or interact: one system for the fast weights, and the other for the slow weights. I have shown that the fast weights are capable of rapid temporary learning, acting as a temporary memory. Furthermore, an ANN with two sets of weights associated with each artificial neuron can act as a dual system of control and is capable of addressing the problem of arbitration between two systems of control.

In a dual weighted network a set of fast weights will allow for more rapid changes in the environment than in a standard network with one set of weights, by reacting quickly to new information or change/surprise when the error in the system is high. This will provide a temporary overlay of the new knowledge to the existing body of knowledge built up and stored in the slow weights, and provide time for the new knowledge to be assimilated into the old. But in a fully learned situation with minimum error or change/surprise the slow weights bearing the existing body of knowledge will dominate as the fast weights decay towards zero.

In order to illustrate the potential usefulness of the model I described a similar model by Daw et al. (2005), who referred to the competition between multiple systems for behavioural choice in the brain, and the problem of arbitration between the systems when they disagreed. They suggested a model of dual action choice, where the systems operated separately and in parallel, governed by a Bayesian principal of arbitration and used model based and model free reinforcement learning to represent two systems of control in the brain: goal directed actions and stimulus responses, respectively (Chapter 7.4). In this chapter I have suggested that the current dual weights model could map onto the stimulus response and goal directed systems identified by Daw et al. as follows:

- **Fast weights** could equate with the tree-search model based system that is capable of flexible, often one shot goal directed actions. However, this is at the expense of being expensive in terms of memory and time.
- **Slow weights** could equate with the model free TD-like system associated with inflexible stimulus response actions that require extensive relearning, but suffers from inflexibility.

While Daw and colleagues successfully modelled controller competition between two systems by simulating results from animal devaluation experiments, I identified four criticisms of the model (Chapter 7.4.4) three of which could be addressed by the dual weighted ANN (Section 8.5):

- The model by Daw and colleagues did not account for interactions between the two model free and model based controllers. The dual weights model could represent the shift in behavioural control from goal directed actions (using fast weights) to habit responses (slow weights) seen in animal devaluation experiments with overtraining, posited by Yin and Knowlton (2006) via the associative network involving dorsomedial striatum to the sensorimotor network involving dorsolateral striatum (Chapter 7.5).
- The self-organising nature of the dual weights model does away with the need for the Bayesian arbitrator (Daw et al. 2005) or homunculus for which there is limited neural evidence. Like Yin and Knowlton (2006) the dual weights model suggests a hierarchical transfer between two systems of control, where the associative network for goal directed actions dominates

early in learning (represented by the fast weights), but as learning progresses a shift is seen to habit formation and the sensorimotor network (represented by the slow weights).

- Daw and colleagues offered a symbolic solution to the problem of arbitration and no attempt was made to address how control could be implemented in the neurons in the brain. However, the dual weighted ANN is a biologically inspired connectionist application and, as a more emergent model of two systems of control; it is in a better position to suggest how control may be exerted in the brain.

To summarise, a dual weighted system with fast weights to react quickly to change/surprise in the environment (this study), capable also of highlighting past experience and identifying context (Hinton & Plaut 1987), offers a novel approach with the afore-mentioned advantages over an existing model of control by Daw et al. It produces emergent properties from a single system with dual weights representing two subsystems, via fast and slow weights that are contained within the same set of neurons and suggests the possibility of two or more interacting systems incorporated into the same neural material. Future improvements to the model, suggested in Section 8.6 may have important implications in addressing how both goal directed and stimulus response learning may coexist. Furthermore it does so without need for the sophisticated algorithms or software packages of O'Reilly and colleagues (e.g. O'Reilly, Frank, Hazy & Watz 2007; O'Reilly and Munakata, 2000).

This chapter is the last in a series of chapters exploring computational modelling of the neural systems involved in schizophrenia. In the final chapter of this thesis I will draw together all of the conclusions reached from the modelling experienced in this and previous chapters in an attempt to reach an answer to the longer term aims of this research, namely: *How can computational models of the neural systems involved in schizophrenia help to improve our understanding of the symptoms and cognitive deficits associated with the disorder?*

Chapter 9

Conclusions and Future Work

In this thesis I have sought to improve our understanding of the neural systems involved in schizophrenia by suggesting possible avenues for future computational modelling in an attempt to make sense of the vast number of studies relating to the symptoms and cognitive deficits associated with this disorder. Modelling in this thesis has comprised of three major themes. I started by looking at abnormalities in the microscopic brain structure, possibly due to excessive synaptic pruning during adolescence. However, it was hard to ignore the considerable evidence for dopamine dysfunction, and the focus of the thesis narrowed to the neurochemistry of the brain and modelling dopamine as a reward prediction error using TD. Finally, taking a wider perspective, I looked at the interactions between cortical and subcortical brain areas, connected by cortico-basal ganglia circuit loops.

In this final chapter I will summarise my contributions to the research area (Section 9.1), and in Section 9.2 I will draw together all of the conclusions from the modelling experienced in previous chapters in an attempt to reach an answer to the longer term aims of this research, namely: *How can computational models of the neural systems involved in schizophrenia help to improve our understanding of the symptoms and cognitive deficits associated with the disorder?* Finally in Section 9.3 I will suggest possible avenues for future research.

9.1 Summary of Contributions

9.1.1 Modelling Contributions

This research has culminated in the production of five models that provide useful clarification in this difficult field:

Model 1 - Simulation of a Speech Perception Neural Network (Chapter 3.1)
Inspired by Hoffman & McGlashan (1997) I implemented a speech perception network which aimed to test the hypothesis that schizophrenia was associated with reduced cortico-cortical connectivity and therefore may arise from excessive synaptic pruning during adolescence. However, I identified problems with the methodology of this early experiment which may explain why I was unable to replicate their findings:

- The training set used was very small compared to the size of the network. It is possible that Hoffman had a non repeatable experiment, and it is probably the case that Hoffman's original experiment had some flaws and was under-constrained.
- No mention was made in the original paper of replication in order to seek a number of solutions to the problem. A statistical analysis would have

strengthened their argument, providing robustness and validity to their findings.

I decided to abandon this modelling technique, and as I began to look at other connectionist models, other questions began to arise that were not addressed by the Hoffman and McGlashan model:

- Their model focused on connectivity in the frontal lobes, but subsequent connectivity models pointed to reduced parahippocampal connectivity in the temporal lobes as an explanation of schizophrenia-like episodic memory deficits (Talamini et al. 2005).
- They only modelled the end point of the neurodevelopmental process and not the formation and progression, which is thought to be pre-programmed and to begin in early life.
- No mention was made of the considerable evidence of dysfunction of the dopamine system and associated areas, such as the basal ganglia.
- While they addressed one of the symptoms, hallucinations, they ignored other symptoms and cognitive deficits and therefore a degree of biological plausibility.

Model 2 - Simulation of the AX-Continuous Performance Task (Chapter 3.2)
I implemented a simplified feed forward connectionist implementation that was able to learn the AX-CPT, which could have been developed further to include the gating system included in Braver et al. (1999). However, although the model made a valuable contribution to schizophrenia research at that time, like with Hoffman & McGlashan, the model appeared to have major shortcomings for use in current schizophrenia research:

- The model focused on the direct dopamine pathway from the ventral tegmental area, which delivers a homogeneous signal to prefrontal cortex, and did not include the basal ganglia and the cortico-basal ganglia circuit loops, which may be paramount in a model addressing both the symptoms and cognitive deficits of schizophrenia. The indirect pathway, via the basal ganglia is better equipped to address the fundamental issue of selective updating, where higher order goals are actively maintained, while updating lower order sub-goals (Cohen et al. 2002).
- As the AX-CPT task used does not apply uniquely to schizophrenia, I question the validity of such models as an explanation of schizophrenia.

Model 3 - An Investigation of the Effects of Dopamine Receptor Antagonism on Running Speed in a Maze (Chapter 5)
My version of the Computational Substrate for Incentive Salience, uniting psychological and formal computational theories, and interpreting expected future reward as incentive salience, has provided additional weight to the claims of McClure et al. (2003) that this single model can capture the ideas of dopamine:

- As a reward prediction error: Using TD and an actor-critic architecture.
- As a purveyor of incentive salience: The effect of dopamine receptor antagonism, modelled as a constant decrease in the TD error signal, resulted in slower running speeds in the maze, severely disrupting the approach

(wanting), providing a similar pattern of results to the original animal experiment by Ikemoto and Panksepp (1996).

The differences in the patterns of behaviour seen in my simulations of animal experiments resulting from high (Ikemoto and Panksepp 1996) and low (Wise et al. 1978) dopamine receptor antagonism, reflected in the timing of the changes to running speed, revealed the dual function of dopamine:

- As a learning signal.
- In the bias of action selection.

The same dopamine reward prediction error was seen to be acting in two different parts of the model:

- *Indirect* effects of dopamine were seen from its role in learning the estimated values that underlay the actions (the critic).
- D*irect* effects of dopamine were seen on action selection (the actor).

An analysis of the difference between high and low dopamine receptor antagonism revealed:

- The immediate effect of the reduction in running speed seen with high dopamine receptor antagonism in Ikemoto and Panksepp (1996) was shown to result from both the *direct* and *indirect* effects of dopamine.
- The delayed, progressive effect of low concentrations of dopamine receptor antagonism in Wise et al. (1978) arose mainly as a result of the *indirect* effects of dopamine, through the slow unlearning of the value estimates, characteristic of the effects of experience-dependent extinction.

By exploring the changing parameters of the model I have been able to answer a number of interesting research questions that warranted further investigation:

High dopamine receptor antagonism (Ikemoto and Panksepp (1996)

- The research in this thesis has demonstrated the *direct* effect of dopamine on action selection, by showing the resulting impact of the lower TD error on the sigmoid decision curve.
- The research in this thesis has demonstrated the *indirect* effect of dopamine, from its role in learning the estimated values that underlay the actions, on the update of the values of states in the maze.
- My simulations found an exponential decrease in running speed in a maze with increasing levels of bias.
- As well as modelling dopamine receptor antagonism using the bias, *b*, in Equation 5.2, I demonstrated that a similar effect could be produced by increasing the scaling constant, *m*, in the same equation. I showed the effect of increasing values of *m* on the sigmoid decision curve.

Low dopamine receptor antagonism (Wise et al. 1978)

- The simulations in this thesis captured the delayed, progressive effect seen with low concentrations of antagonists that emerged through repeated exposure. My comparison between the effects of higher and lower dopamine receptor antagonism showed that the effects of the lower levels of dopamine blockade took longer to develop and were not so pronounced as with a high level of bias.

- Evidence from simulations in this thesis has shown a greater reliance of lower dopamine receptor antagonism on the *indirect* effect on the update of the values of states in the maze, than on the *direct* effect of action selection seen for higher levels of dopamine receptor antagonism.
- Lower levels of dopamine receptor antagonism produced smaller, delayed TD errors, which, in turn, amounted to less of a reliance on the *direct* effects of action selection on running speed, and more reliance on the *indirect* build up of the new values for states through relearning of the value weights.
- Lower dopamine receptor antagonism resulted in a similar pattern for values of states in the maze as extinction, which was in line with the claims of McClure and colleagues, and suggested a similar pattern of unlearning.

Model 4 - An Analysis of the Relationship between Temporal Difference Learning and Uncertainty Coding in a Computational Model of Dopaminergic Signalling: (Chapter 6)

The TD model developed in Chapter 5 was extended to address the ongoing debate as to whether or not dopamine encodes uncertainty in the delay period between presentation of a conditioned stimulus and receipt of a reward, as demonstrated by sustained activation seen in single dopamine neuron recordings (Fiorillo et al. 2003).

By introducing uncertainty and scaling the negative TD errors by a factor of one sixth in order to compensate for the asymmetric coding scaling of dopamine neuron firing, the novel use of this model captured the following effects recorded by Fiorillo and colleagues, and demonstrated in the simulations of Niv et al (2005):

- The phasic activations at the expected time of reward.
- The sustained increase in activity from the onset of the conditioned stimulus until the expected time of reward.
- The sustained activation increasing with increasing reward magnitude.

What was new about this study is that I determined criteria for sustained activation (Fiorillo et al 2003; 2005) and ramping (Niv et al. 2005) that distinguished between the conflicting terminologies, which permitted analysis of single trials in my simulations. Single trial analysis has allowed me to address the following two points raised by Fiorillo et al. (2005) in response to the criticisms of Niv et al. (2005) of their original paper:

- In the simulations both sustained activation and ramping were common in single trials during uncertainty, but as neither sustained activation nor ramping was greater with maximum uncertainty the simulations did not support the claims by Fiorillo et al. (2003; 2005) that dopamine is encoding uncertainty during the delay period between CS and receipt of reward.
- It was demonstrated that activity in the last part of the delay period does not always reflect the reward outcome that followed the last exposure to that same CS. Specifically; the history of consecutive trials should be taken into consideration when analysing reward prediction errors and not just the last trial. In the presence of uncertainty, the particular course taken through a series of trials is different in each simulation, as it depends on the exact order

of rewarded and non-rewarded runs, which are delivered randomly. It is important that these factors should be taken into consideration when interpreting data from peri-stimulus-time-histograms of activity over different trials and inter-trial averaging, such as that in Fiorillo et al. (2003).

My simulations supported the claims of Niv et al. (2005) and provided predictions in an ongoing debate over whether or not dopamine encodes uncertainty that could be verified by experimental data. Capturing such detailed physiological recordings in an alternative model to Niv et al. strengthens the use of TD as a valid method of modelling and quantifying the dopamine reward prediction error.

To date this model has resulted in the following disseminations:

- Abstract published in proceedings of *International Conference on Schizophrenia Research*, Colorado Springs, Colorado, March 28 – April 1 2007.
- A Model of Dopamine and Uncertainty Using Temporal Difference. Six page paper published in the proceedings of *XXV111 Annual Conference of the Cognitive Science Society*, Vancouver, Canada. 2006b.
- Poster presented at the *10th International Conference on Cognitive and Neural Systems*, Boston, Massachusetts, USA, May 17-20, 2006.

Model 5 – A Connectionist Model of Dual System Control: (Chapter 8)
Based on a model by Hinton and Plaut (1987), originally designed to address the problem of catastrophic forgetting in multilayer ANNs, I suggested an alternative use for a model with both fast and slow weights to address the problem of arbitration between two systems of control through rapid learning in the fast weights:

- In a dual weighted network a set of fast weights allowed for more rapid changes in the environment than in a standard network with one set of weights, by reacting quickly to new information or change/surprise when the error in the system was high. This provided a temporary overlay of the new knowledge to the existing body of knowledge built up and stored in the slow weights, and provided time for the new knowledge to be assimilated into the old.

    - The fast weights dominated when the error in the system was high, early in training and when new associations were introduced to the network.

- But in a fully learned situation with minimum error or change/surprise the slow weights bearing the existing body of knowledge dominated as the fast weights decayed towards zero.

    - The slow weights dominated when the error in the system was low and the task was well learnt and more habitual.

In order to illustrate the potential of this dual weighted network I referred to the distinction between different networks for goal directed actions and stimulus response habits (Yin and Knowlton 2006) that can be modelled effectively using model based and model free learning, respectively (Daw et al. 2005):

- **Fast weights** could equate with the tree-search model based system that is capable of flexible, often one shot goal directed actions. However, this is at the cost of being expensive in terms of memory and time.

    - Associated with cortico-basal ganglia circuit loops involving dorsomedial striatum.
- **Slow weights** could equate with the model free TD-like system associated with inflexible stimulus response actions that require extensive relearning, but suffers from inflexibility.
    - Associated with cortico-basal ganglia circuit loops involving dorsolateral striatum.

This model has several advantages over the existing model of arbitration between two controllers by Daw et al. (2005):

- It allows for interactions between the two controllers.
- The self-organising nature of the dual weights model does away with the need for the Bayesian arbitrator or homunculus.
- As a biologically inspired connectionist application and an emergent model of two systems of control, it is in a better position to suggest how control may be exerted in the brain than a purely symbolic model.
- It is a simple, relatively transparent model that can offer insights into methods of arbitration between two systems without need for sophisticated algorithms or software packages.

Modelling these two systems associated with different cortical-subcortical loops offers the potential of incorporating both the symptoms and cognitive deficits associated with schizophrenia by taking into account the interactions between midbrain/striatum and cortical areas.

9.1.2 Other Contributions

The first part of the literature review in Chapter 2 contained a detailed account of the motivation behind this thesis:

- It provided a basic description of schizophrenia, how it manifests itself in the human body and its biological underpinnings.
- I outlined the difficulties in finding the ultimate cause of the disorder.
- I explained the advantages of using computational modelling for this purpose; in particular, a biologically inspired connectionist approach.

Chapter 4 contained the second part of the literature review. Having justified switching to a new line of reasoning involving the dopamine system, I looked at a body of research inspired by the physiological recordings of dopamine neurons on alert monkeys, by Wolfram Schultz and colleagues who showed that information about rewarding stimuli was encoded in dopaminergic activity. This included:

- Dopamine as a reward prediction error signal.
- TD incorporating an actor-critic architecture as an effective method of modelling the dopamine reward prediction error signal.
- The role of the basal ganglia as a contextual processor.
- The Incentive Salience Hypothesis.
- Evidence of the dopamine reward prediction error in humans from fMRI studies.

Disseminations arising from this line of reasoning include:

- How Do Computational Models of the Role of Dopamine as a Reward Prediction Error Map on to Current Dopamine Theories of Schizophrenia? Six page paper published in the proceedings of *XXV111 Annual Conference of the Cognitive Science Society*, Vancouver, Canada. 2006a. (See Appendix II).
- Abstract published in proceedings of *The International Conference on Schizophrenia Research*, Davos, Switzerland, February 2006.
- Abstract published in the proceedings of the *XXV11 Annual Conference of the Cognitive Science Society*, Stresa, Italy, 21-23 July 2005.

The sections in Chapter 4 provided the methodology behind a model by McClure et al. (2003) that incorporated the Incentive Salience Hypothesis into an Actor-Critic model of dopamine as a reward prediction error. In Chapter 5 I described and implemented a simulation of the Computational Substrate for Incentive Salience by McClure et al. and this model was expanded to look at the relationship between TD learning and uncertainty coding in Chapter 6.

Chapter 7 marked another change in the direction of modelling and contained the third part of the literature reviews. Here I began to widen the focus of this thesis, which had narrowed down to TD in Chapters 4 to 6, back to my longer-term aim of improving the understanding of the neural systems involved in schizophrenia; taking into account a wider perspective of the interactions between cortical and subcortical brain areas, connected by cortico-basal ganglia circuit loops.

In chapter 7 I drew attention to the limitations of using TD alone to account for action control in the brain:

- The basic TD model is unable to distinguish between different rewards with a similar value that are preceded by an appropriate CS (Smith et al. 2006). In particular it is unable to distinguish between the effects of dopamine manipulation on distal rather than proximal rewards (Daw et al. 2005; Smith et al. 2004; 2006).
- TD can be brittle as the prediction chains used by TD over successive time-steps break down when the CS-US relationship is unreliable, such as in the complicated 1-2-AX working memory task, which involves maintaining both subgoals and higher order goals (O'Reilly et al. 2007).
- The simple actor-critic is insensitive to motivational state and fails to take into account some of the psychological differences between Pavlovian and instrumental conditioning (Dayan & Balleine 2002).
- In TD any change in task will have to be relearned explicitly, which will take time. In reality, relearning often needs to take place quickly, so current TD models do not account for all types of learning (Daw et al. 2005).

As well as offering alternatives to TD, I noticed that all four studies pointed towards two systems of control operating together (Table 7.2). In particular, Daw et al. referred to the competition between multiple systems for behavioural choice in the brain, and the problem of arbitration between the systems when they disagreed. They suggested a model of dual action choice, where the systems operated separately and in parallel, governed by a Bayesian principal of arbitration. My alternate connectionist model of dual system control is detailed in Chapter 8.

9.2 Conclusions

In this thesis I have explored a variety of computational models to seek an answer to the question: *How can computational models of the neural systems involved in schizophrenia help to improve our understanding of the symptoms and cognitive deficits associated with the disorder?* In this section I draw together the results from all my simulations in order to specifically address that question.

Models 1 and 2
The speech perception network (Hoffman & McGlashan 1987) and the learning and gating model (Braver et al. 1999) detailed in Chapter 3 were designed specifically to address either, one of the symptoms, or one of the cognitive deficits of the disorder. However, following a detailed exploration of those models, I discovered that their power of explanation was limited. Nevertheless, the flaws of these models pointed to a new line of reasoning to account for the symptoms and cognitive deficits of schizophrenia which included dopamine dysfunction and the neuroscience of the cortico-basal ganglia circuit loops, on which the remainder of this thesis rested.

Models 3 and 4
While the TD models detailed in Chapters 5 and 6 were not specific to schizophrenia, they were models of the specific firing patterns of dopamine, a possible mechanism in the midbrain and cortex for the symptoms and cognitive deficits of schizophrenia. These computational models have suggested how dopamine firing patterns may transfer to behaviour, and attempt to answer some of the questions that are difficult to address using conventional methods. In particular, an answer to the question of whether or not dopamine is encoding uncertainty could result in a better understanding of the nature of dopamine signaling, with implications for the psychopathology of cognitive disorders, like schizophrenia, for which dopamine is commonly regarded as having a primary role.

The Computational Substrate for Incentive Salience by McClure et al. (2003) successfully united psychological and formal computational theories by interpreting expected future reward as incentive salience. This approach saw dopamine receptor antagonism, characteristic of the effects of antipsychotic drugs, as the inhibition of the ability to initiate actions necessary for gaining rewards. According to Kapur (2003) aberrant phasic dopamine responses could lead to delusions and possibly hallucinations associated with thought disturbance in schizophrenia (Chapter 4.4.1). The action of antipsychotic drugs may protect against the formation of the aberrant internal representations by attenuating aberrant incentive salience via phasic dopamine signals. A clearer understanding of the role of dopamine and the best methods of modelling those functions will help in the quest for the understanding of schizophrenia.

Much of the research into schizophrenia has been, and is still, centered on the robust finding that there is a remarkable correlation between the efficacy of antipsychotic drugs in treating psychosis and the ability of those drugs to block the dopamine D2 receptors. However, while it is posited that psychosis results from a dysregulation of the dopamine mesolimbic system (Weinberger 1987; Grace 1991; Kapur & Mamo 2003), there is still little evidence to support this hypothesis. Computational modelling can help to support this position by simulating biological data, thus providing an insight into the underlying mechanisms.

My version of the Computational Substrate for Incentive Salience by McClure et al. (2003), using TD, captured the differences in the patterns of behaviour seen in two animal experiments resulting from high and low dopamine receptor antagonism, revealing the dual function of dopamine, as a learning signal and in the bias of action selection. In addition, this model was modified to simulate the detailed electrophysiological recordings of dopamine neuron firing by Fiorillo et al (2003; 2005), strengthening the argument for TD as a valid method of modelling and quantifying the dopamine reward prediction error. The simulations detailed in these two chapters are examples of science through simulation and my model of uncertainty demonstrates how computational modelling can help to clarify a position by generating testable predictions that can be verified with behavioural data.

While I appreciate that TD must be one of many algorithms working simultaneously in the brain, I agree with Niv et al. (2005) that the ramping signal, both in single trials, and when averaged over multiple trials, is strong evidence for the nature of the learning mechanism of a shift in dopamine activity from the expected time of reward to the CS, using TD. I suggest that it is both reasonable and biologically plausible for future models of dopamine to include TD learning. However, in spite of the fact that TD offers a very good account of the firing patterns of dopamine neurons, it is clear from the studies detailed in Chapter 7 that it offers an incomplete picture of the reinforcement learning account of action control in the brain.

Model 5

My ANN containing dual weights developed in Chapter 8 demonstrated how two systems of control in the brain may coexist and interact. This dual weighted model has shown emergent properties from a single system with dual weights representing two subsystems, via fast and slow weights that are contained within the same set of neurons. This suggests the possibility of two or more interacting systems incorporated into the same neural material, permitting transfers between the two systems.

In order to illustrate the potential of this model I suggested how stimulus-response habits (associated with incentive salience and the striatum) and goal directed actions (associated with working memory and the prefrontal cortex) may relate together in different cortico-basal ganglia circuit loops; and how these two circuits may interact. Both working memory and incentive salience feature in the expression of cognitive deficits and symptoms of schizophrenia, respectively.

As mentioned in Chapter 2, it is likely that symptoms and deficits arise from different brain areas and this adds to the general difficulty of finding the cause or causes of schizophrenia. Modelling cortical-subcortical loops offers the potential of incorporating both the symptoms and cognitive deficits associated with schizophrenia by taking into account the interactions between midbrain/striatum and cortical areas. Future improvements to the dual weights model, suggested in Chapter 8.6 and discussed further in Section 9.3 may have important implications in addressing how both goal directed and stimulus response learning may coexist and how a dysfunction of these systems could underlie disorders such as schizophrenia. It may be possible to explain one in terms of the other, or it may be that the two cannot be equated.

Final Thoughts
I reiterate the limitations of modelling originally addressed in Chapter 2 that the models in this thesis are informative rather than definitive; qualitative rather than quantitative, and as such cannot be predictive. However, the usefulness of these models lie in their relative transparency and the ability to suggest what might be reasonable lines for research. Therefore, I am not claiming that the brain is actually working in the same way as the algorithms underlying these models, but simply suggesting how these processes can be modelled and what *might* occur. Such computational insights can be used to generate quantitative findings, providing avenues for further empirical study or treatment strategies, and may contribute to biological theory.

Schizophrenia has a varied evidence base with different methodologies and applications, ranging from pharmacological and neurophysiological studies to brain imaging. Ultimately the results from all these studies should be integrated to form a cohesive whole for an understanding of the disorder. However, we are still far from reaching this point and cognitive modelling will continue to try to unite biological and psychological theories in an attempt to answer the challenging question of whether or not there is a simple mechanism on which higher level cognition can be built.

9.3 Future Directions

Model 4 - An Analysis of the Relationship between Temporal Difference Learning and Uncertainty Coding in a Computational Model of Dopaminergic Signalling: (Chapter 6)

- The current model is parameter dependent and discrete, containing a set number of states. In reality neuron firing is noisy and therefore less predictable and a spiking form of this model could contain more realistic noise and more closely resemble dopamine neuron firing in vivo.

- The current model only allows for one CS in any one trial. Future versions could provide multiple CS, allowing me to address a further point by Fiorillo et al. (2005) concerning a dissociation between the size of the ramp and the sustained activation at the estimated time of reward, identified in Tobler, Fiorillo and Schultz (2005).

- Apart from modelling uncertainty in the receipt of reward, a modified version of this model could be used to look at an alternative, under researched form of uncertainty: the uncertainty in the timing of our own intrinsic internal clocks.

- Some evidence suggests that the persistent reward responses of dopamine cells during conditioning are only accurately replicated by a TD model with long-lasting eligibility traces, such as TD(λ) (Pan et al. 2005). It would be interesting to implement the model using this version of the TD algorithm to see the effect of different strengths of eligibility trace.

Model 5 – A Connectionist Model of Dual System Control: (Chapter 8)

- Introducing dopamine driven learning into future versions of the model will enhance biological plausibility and may obviate the need to set the parameters of the model prior to each task. A phasic dopamine reward prediction error, modelled as a TD error, representing significance/surprise could affect the modifiable parameters of the model in the following ways:
    - By changing learning rates.
    - By changing the decay rate for the fast weights.
    - A combination of the above.

- Other applications could be to apply the model to some of the other dual systems identified in Chapter 7, Table 7.2. In particular, modelling the phasic dopamine neuron firing seen in learning and the tonic dopamine relating to the expression of previously acquired behaviour could potentially address one of the limitations of the basic TD model; the inability to distinguish between the effects of dopamine manipulation on distal rather than proximal rewards (Daw et al. 2005; Smith et al. 2004; 2006).

- I have provided a theoretical argument for the implementation of an ANN with dual weights. However, in order to demonstrate the value of the model it will be necessary to test the theory using psychological tasks, possibly of sequential learning, that will exploit two systems of control. It will then be possible to explore how a dysfunction of these systems could influence the two routes to action.

  Ideally, the task should involve learning using a supervised data set, where the difference between the actual and target output can be backpropagated through the network. In addition, in order to demonstrate flexible cognitive control and broad training experience there should be subtasks. Rougier et al. (2005) provided an interesting sequential learning task bearing these characteristics that could be adapted to test this model.

- There are clearly many systems of control operating in parallel in the brain and future versions of the model could accommodate multiple systems by having multiple weights associated with each artificial neuron (or set of neurons).

References

Aakerlund, L., & Hemmingsen, R., (1998) Neural networks as models of psychopathology. *Soc. Of Biological Psychiatry,* **43**, 471-482

Abi-Dhargham, A., (2004) Do we still believe in the dopamine hypothesis? New data bring new evidence *Neuropsychopharmacology,* **7**(supplement 1), S1-S5

Alder, R., & Clink, D.W., (1957) Effects of chlorpromazine on the acquisition of an avoidance response in the rat. *J. Pharmacol. Exp. Ther.,* **131**, 144-148

Alexander, G. E., & Delong, M. R. (1985) Microstimulation of the primate neostriatum. Somatotopic organization of striatal microexcitable zones and their relation to neuronal response properties. J. Neurophysiol. **53**, 1417–1430

Atallah, H.E., Lopez-Paniagua, D., Rudy, J.W., & O'Reilly, R.C., (2007) Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nature Neuroscience,* **10(1)**, 126-131

Atkins, P.W.B., (2001) What happens when we relearn part of what we previously knew? Predictions and constraints for models of long term memory. *Psychological Research,* **65**, 202-215

Atkins, P.W.B., & Murre, J.M.J., (1998) Recovery of unrehearsed items in connectionist models. *Connection Science,* **10(2),** 99-119

Balleine, B.W., & Dickinson, A., (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacoloy,* **37**, 407-419

Barto, A.G., (1995) Adaptive critics and the basal ganglia. In *Models of Information Processing in the Basal Ganglia.* Edited by Houk, J.C., Davis, J.L., & Beiser, D.G. Cambridge, MA: MIT Press; 215-232

Berridge, K.C., (2001) In *The Psychology of Learning and Motivation: Advances in Research and Theory,* (ed. Medin, D.L.) vol 40:223-278 (Academic, San Diego)

Berridge, K.C., & Robinson, T.E., (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* **28**, 309-369

Bilder, R.M., Goldman, R.S., et al (2000) Neuropsychology of first-episode schizophrenia: Initial characterisation and clinical correlates. *American Journal of Psychiatry* **157(4)**, 549-559

Bindra, D., (1974) A motivational view of learning, performance, and behaviour modification. *Psychological Review,* **81(3)**, 99-213

Bolles, R.C., (1972) Reinforcement, expectancy and Learning. *Psychological Review,* **79**, 94-409

Braver, T.S., (1997) *Mechanisms of Cognitive Control: A Neurocomputational Model*. PhD Thesis, Carnegie Mellon University

Braver, T.S., Barch, D.M., & Cohen, J.D., (1999) Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry.* **46**, 312-328

Brunel, N., & Wang, X-J., (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience,* **11**, 63-85

Carlsson, A., Waters, N., et al (2001) Interactions between monoamines, glutamate and GABA in Schizophrenia. *Annu. Rev.Pharmacol. Toxicol,* **41(237)**, 237-60

Chapman, L.J., & Chapman, J.P., (1978) The measurement of differential deficit. *J. Psychiatric res.* **14**, 303-311

Cohen, J.D., Braver, T.S., & Brown, J.W., (2002) Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology,* **12**, 223-229

Cohen, J.D., & Servan-Schreiber, D., (1992) Context, Cortex, and Dopamine: A Connectionist Approach to Behavior and Biology in Schizophrenia. *Psychological Review,* **1**, 45-77

Cohen, J.D., & Servan-Schreiber, D., (1993) A theory of dopamine function and its role in cognitive deficits in schizophrenia. *Schizophrenia Bulletin,* **19**, 85-104

Cook, L., & Weidley, E., (1957) Behavioral effects of some pharmacological agents. *Ann. NY Acad Sci.,* **66**, 740-752

Cousins, M.S., Atherton, A., Turner, L & Salamone, J.D., (1996) Nucleus accumbens dopamine depletion after relative response allocation in a T-maze cost/benefit task. *Behavioural Brain Research,* **74**, 189-197

Daw, N.D., Niv, Y., & Dayan, P., (2005) Actions, policies, values and the Basal Ganglia. *In Bezard, Editor, Recent Breakthroughs in Basal Ganglia Research.* New York, NY: Nova Science Publishers

Daw, N.D., Niv, Y., & Dayan, P., (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience,* **8(12)**, 1704-1711

Dayan P., & Balleine, B.W., (2002) Reward, Motivation and Reinforcement learning. *Neuron,* **36**, 285-298

Devan, B.D., & White, N.M., (1999) Parallel information processing in the dorsal striatum: relation to hippocampal function. *Journal of Neuroscience,* **19**, 2789-2798

Dickinson, A., & Balleine, B.W., (2002) The role of learning in motivation. *In Learning, Motivation and Emotion, Volume 3 of Steven's Handbook of Experimental Psychology, Third Edition,* C.R. Gallistel, ed. New York: John Wiley & Sons

Doya, K., (2007) Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal,* **1(1)**, 30-40

Durstewitz, D., Kelc, M., & Gunturkun, O., (1999) A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience,* **19(7)**, 2807-2822

Egan, M.F., & Weinberger, D.R., (1997) Neurobiology of schizophrenia *Current Opinion in Neurobiology,* **7**, 701-707

Egelman, D.M., et al. (1998) A computational role for dopamine delivery in human decision-making. *J. Cognitive Neuroscience,* **10**, 623-630

Elman, J.L., (1990) Finding structure in time. *Cognitive Science,* **14(2)**, 179-211

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, **299**, 1989-1902

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behavioural and brain Functions*, **1**, 7

Frank, M.J., (2005) Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cognitive Deficits in Medicated and Non-medicated Parkinsonism. *Journal of Cognitive Neuroscience*, **17**, 51-72

Freeman, W.J., (1979) Nonlinear gain mediating cortical stimulus-response relations. *Biological Cybernetics,* **33**, 243-247

French, R.M., (1999) Catastrophic forgetting in connectionist networks: causes, consequences and solutions. *Trends in Cognitive sciences,* **3(4)**, 128-135

Gottesman, I. I., (1991) *Schizophrenia Genesis: The Origins of Madness.* Freeman, New York

Grace, A. A., (1991) Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. *Neuroscience*, **41**, 1–24

Guarracui, F.A., & Kapp, B.S., (1999) An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential Pavlovian fear conditioning in the awake rabbit. *Behav. Brain Res.,* **99**, 69-179

Harrison, P.J., (1997) Schizophrenia: a disorder of neurodevelopment? *Current Opinion in Neurobiology,* **7**, 285-289

Harvey, P.D., Koren, D., Reichenberg, A., & Bowie, C.R. (2005) Negative Symptoms and Cognitive Deficits: What is the nature of their Relationship? *Schizophrenia Bulletin,* **32(2)**, 250-258

Hartzell, H.C., (1981) Mechanisms of slow synaptic potentials. *Nature,* **291**, 539-543

Hazy, T.E., Frank, M.J., & O'Reilly, R.C., (2006) Banishing the homunculus: Making working memory work. *Neuroscience,* **139**, 105-118

Hills, T.T., (2006) Animal foraging and the evolution of goal-directed cognition. *Cognitive Science* **30**, 3-41

Hinton, G.E., & Plaut, D.C., (1987) Using fast weights to deblur old memories. *In proceedings of the 9th Annual Conference of the Cognitive Science Society,* pp177-186. Hillsdale, NJ. Erlbaum

Hinton, G.E., & Sejnowski, T.J., (1986a) Learning and relearning in Boltzmann machines. In D.E. Rumelhart, J.L. McClelland and the PDP Research Group (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition, Vol 1: Foundations,* Cambridge, MA, MIT Press

Hinton, G.E., & Shallice, T., (1991) Lesioning and attractor network: investigations of acquired dyslexia. *Psychological Review,* **98**, 74-95

Hoffman, R.E., & McGlashan, T.H., (1997) Synaptic elimination, neurodevelopment, and the mechanism of hallucinated 'voices' in schizophrenia. *American Journal of Psychiatry*. **154(12)**, 1683-1689

Hoffman, R.E., & McGlashan, T.H., (2001) Neural network models of schizophrenia. *The Neuroscientist.* **7(5)**, 441-454

Holland, P.C., & Gallagher, M., (2004) Amygdala-frontal interactions and reward expectancy. *Curr. Opinion Neurobiol.,* **14**, 148-155

Hollerman, J.R., & Schultz, W., (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience,* **1,** 304-309

Horvitz, J.C., et al (1997) Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Res.,* **759**, 51-258

Houk, J.C., (1995) Information processing in modular circuits linking basal ganglia and cerebral cortex. In *Models of Information Processing in the Basal Ganglia.* Edited by Houk, J.C., Davis, J.L., & Beiser, D.G. Cambridge, MA: MIT Press; 3-10

Houk, J.C., Adams, J.L., & Barto, A.G., (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of Information Processing in the Basal Ganglia.* Edited by Houk, J.C., Davis, J.L., & Beiser, D.G. Cambridge, MA: MIT Press; 249-270

Houk, J.C., Davis, J.L., & Beiser, D.G., (1995) *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press

Huttenlocher, P.R., (1979) Synaptic density in the human frontal cortex-developmental changes and effects of aging. *Brain Research,* **163**, 195-205

Ikemoto, S., & Panksepp, J., (1996) Dissociations between appetitive and consummatory responses by pharmacological manipulations of reward-relevant brain regions. *Behavioral Neuroscience,* **110**, 331-345

Ikemoto, S., & Panksepp, J., (1999) The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews,* **31(1)**, 6-41

Izquierdo, A., Suda, R.K., & Murray, E.A., (2004) Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *Journal of Neuroscience,* **24**, 7540-7548

Joel, D., & Weiner, I., (2000) The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience,* **96**, 451-474

Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V., & Graybiel, A.M., (1999) Building neural representations of habits. *Science,* **286**, 1745-1749

Johnstone, E.C., Crow, T.J., Frith, C.D., Husband, J., & Kreel, L. (1976) Cerebral ventricular size and cognitive impairment in chronic schizophrenia. *Lancet,* **I**, 848-851

Kapur, S., (2003) Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry,* **160(1)**, 13-23

Kapur, S., (2004) How antipsychotics become anti-'psychotic' – from dopamine to salience to psychosis. *Trends in Pharmacological Sciences*. **25(8)**, 402-406

Kapur, S., & Mamo, D., (2003) Half a century of antipsychotics and still a central role for dopamine D2 receptors. *Prog. Neuropsychopharmacology Biol. Psychiatry,* **27**, 1081-1090

Kapur, S., Mizrahi, R., & Li, M., (2005) From dopamine to salience to psychosis – linking biology, pharmacology and phenomenology of psychosis. *Schizophrenia Research*. **79**, 59-68

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

Kupferman, I., (1979) Modulatory actions of neurotransmitters. *Annual Review of neuroscience,* **2**, 447-465

Laurelle, M., (1998) Imaging dopamine transmission in schizophrenia. A review and meta-analysis. *Q J Nucl. Med.* **42**, 211-221

Liddle, P.F., (1996) Functional imaging--schizophrenia. *British Medical Bulletin.* **52(3)**, 486-94

Lim, K.O., Tew, W., Kushner, M., Chow, K., Matsumoto, B, DeLisi, L.E., (1996) Cortical gray matter volume deficit in patients with first-episode schizophrenia. *Am. Journal of Psychiatry,* **153**, 1548-1553

Ljungberg, T., et al (1992) Responses of monkey dopamine neurons during learning of behavioural reactions. *Journal of neurophysiology,* **67**, 45-163

Ljunberg, T., Apicella, P., & Schultz, W., (1992) Responses of monkey dopamine neurons during learning of behavioural reactions. *Journal of Neurophysiology,* **67**, 145-163

McClelland, J.L., & Kawamoto, A.H., (1986) Mechanisms of sentence processing: Assigning roles to constituents of sentences. In J.L. McClelland, D.E. Rumelhart and the PDP Research Group (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition, Volume II: Psychological and biological models,* Cambridge, MA, MIT Press

McCloskey, M., & Cohen, N., (1989) Catastrophic interference in connectionist networks: the sequential learning problem, in *The Psychology of Learning and motivation,* (Vol 24) (Bower, G.H., ed.), pp109-164, Academic press

McClure, S.M., Berns, G.S., & Montague, P.R., (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron,* **38**, 339-346

McClure, S.M., Daw, N.D., & Montague, P.R., (2003) A computational substrate for incentive salience. *Trends in Neurosciences* **26(8)**, 423-428

McClure, S.M., Laibson, D.I., Lowenstein, G., & Cohen, J.D., (2004) Separate neural systems value intermediate and delayed monetary rewards. *Science,* **306**, 503-507

McKenna, P.J., (1997) *Schizophrenia and Related Syndromes.* Psychology Press Ltd, Hove, East Sussex

McLeod, P., Plunkett, K., & Rolls, E.T., (1998) *Introduction to Connectionist Modelling of Cognitive processes.* Oxford University Press

Maffii, G., (1959) The secondary conditioned response of rats and effects of some psychopharmalogical agents. *J. Pharmacy Pharmacol.,* **11**, 129-139

Maher, B.A. (1988) Anomalous experience and delusional thinking: The logic of explanations. In T.F.Oltmanns and B.A. Maher (Eds.), *Delusional beliefs* (pp. 15-33), New York: Wiley

Montague, P.R., Dayan, P., & Sejnowski, T.J., (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian Learning. *Journal of neuroscience,* **16(5)**, 1936-1947

Montague, P.R., Hyman, S.E., & Cohen, J.D., (2004) Computational roles for dopamine in behavioral control. *Nature* **431***,* 760-767

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

Moody, T.D., Bookheimer, S.Y., Vanek, Z., & Knowlton, B.J., (2004) An implicit learning task activates medial temporal lobe in patients with Parkinson's disease. *Behavioral Neuroscience,* **118**, 438-442

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H., (2006) Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience,* **9(8)**, 1057-1063

Niv, Y., Daw, N.D., Joel, D., & Dayan, P., (2007) Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology,* DOI: 10.1007/s00213-006-0502-4

Niv, Y., Daw, N.D., & Dayan, P., (2006) Choice Values. *Nature Neuroscience,* **9(8)**, 987-988

Niv,Y., Duff, M.O., & Dayan, P. (2005). Dopamine, uncertainty and TD Learning. *Behavioural and brain Functions*, **1**, 6 1-9

O'Doherty, J.P., Dayan, P., et al (2003) Temporal difference models and reward-related learning in the human brain. *Neuron,* **38,** 329-337

O'Doherty, J.P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R.J (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science,* **304** no5669, 452-454

O'Reilly, R.C., (2006) Biologically based computational models of high-level cognition. *Science,* **314** no5796, 91-94

O'Reilly, R.C., & Frank, M.J., (2006) Making Working Memory Work: A Computational Model of Learning in the Frontal Cortex and Basal Ganglia. *Neural Computation* **18**, 283-328

O'Reilly, R.C., Frank, M.J., Hazy, T.E., & Watz, B., (2007) PVLV: The primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience,* **121(1)**, 31-49

O'Reilly, R.C., & Munakata, Y., (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the mind by simulating the brain.* MIT Press

Pao, Y-H., (1989) *Adaptive Pattern Recognition and Neural Networks.* Addison-Wesley Publications Co. Inc: USA

Pan, W-X., Schmidt, R., Wickens, J.R., & Hyland, B.I., (2005) Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience,* **25(26)**, 6235-6242

Parkinson, J.A., Dalley, J.W., Cardinal, R.N., Bamford, A., Fehnert, B., Lachenal, G., Rudarakanchana, N., Halkerston, K.M., Robbins, T.W., & Everitt, B.J., (2002) Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behavioural Brain Research,* **137**, 149-163

Pasupathy, A., & Miller, E.K., (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature,* **433**, 873-876

Pavlov, I.P., (1927) *Conditional Reflexes.* Oxford Univ. Press, London

Pearce, J.M., & Hall, G.A., (1980) *Psychological Review,* **87**, 532

Plaut, D.C., (1996) Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language,* **52**, 25-82

Plunkett, K., & Elman, J.L., (1997) *Exercises in rethinking innateness: A handbook for connectionist simulations.* MIT Press, Cambridge, MA

Redgrave, P., Prescott, T.J., & Gurney, K., (1999) The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience,* **89**, 1009-1023

Rescorla, R.A., & Wagner, A.R., (1972) A theory of Pavlovian conditioning: the effectiveness of reinforcement and non-reinforcement. In: *Classical Conditioning. 2. Current Research and Theory* (Black, A.H., Prokast, W.F., eds.) 64-69. New York: Appleton Century-Crofts

Reynolds, S. M., & Berridge, K.C., (2000) Positive and negative motivation in nucleus accumbens shell: Bivalent rostrocaudal gradients for GABA-elicited eating, taste, 'liking'/'disliking' reactions, place preference/Avoidance and fear. *Journal of Neuroscience,* **22(16)**, 7308-7320

Romo, R., & Schultz, W. (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiol.* **63(6)**, 7-624

Rougier, N.P., Noelle, D.C., Braver, T.S., Cohen, J.D., & O'Reilly, R.C. (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. *PNAS*, **102(20)**, 7738-7343

Rumelhart, D.E., Hinton, G.E., & Williams, R.J., (1986) Learning internal representations by error propagation. In D.E.Rumalhart, J.L.McClelland & the PDP Research Group(Eds), *Parallel Distributed Processing,* 1: *Foundations* pp319-362. Cambridge, MA: MIT Press

Schultz, W., (1986) Responses of midbrain dopamine neurons to behavioural trigger stimuli in the monkey. *J. Neurophysiol.* **56**, 1439-1462

Schultz, W., (1992) Activity of dopamine neurons in the behaving primate. *Seminars in Neurosciences,* **4**, 129-138

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

Schultz, W., Apicella, P., & Ljunberg, T., (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience,* **13(3)**, 900-913

Schultz, W., Dayan, P., Montague, P.R., (1997) A Neural substrate of prediction and reward. *Science*, **275(5306)**, 1593-1599

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

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J.R., & Dickenson, A., (1995) Reward-related signals carried by dopamine neurons. In *Models of Information Processing in the Basal Ganglia.* Edited by Houk, J.C., Davis, J.L., & Beiser, D.G. Cambridge, MA: MIT Press; 233-248

Salamone, J.D., Cousins, M.S., & Snyder, B.J., (1997) Behavioural functions of nucleus accumbens dopamine: empirical and conceptual problems with the anhedonia hypothesis. *Neurosci. Biobehav. Rev.,* **21**, 341-359

Selemon, L.D., (2004) Increased cortical neuronal density in schizophrenia. *American Journal of Psychiatry,* **161(9)**, 1564

Selemon, L.D., Rajkowski, G., & Goldman-Rakic, P.S., (1995) Abnormally high neuronal density in schizophrenic cortex: A morphometric analysis of prefrontal area 9 and occipital cortex area17. *Arch Gen Psychiatry,* **52**, 805-18

Servan-Schreiber, D., Printz, H., & Cohen, J.D., (1990) A network model of catecholamine effects: Gain, signal-to-noise ratio and behavior. *Science,* **249**, 892-895

Seymour, B., O'Doherty, J.P., et al (2004) Temporal difference models describe higher-order learning in humans. *Nature,* **429**, 664-667

Shenton, M.E., Dickey, C.C., Frumin, M., & McCarley, R.W., (2001) A review of MRI findings in schizophrenia. *Schizophrenia Research,* **49**, 1-52

Simpson J, Done D.J., Valeé-Tourangeau (1998) An Unreasoned Approach: A Critique of Research on Reasoning and Delusions. *Cog. Neuropsychiatry,* **3**, 1-20

Smith, A., Becker, S., & Kapur, S., (2005) A computational model of the functional role of the ventral-striatal D2 receptor in the expression of previously acquired behaviors. *Neural Computation,* **17**, 391-395

Smith, A., Li, M., Becker, S., & Kapur, S., (2004) A model of antipsychotic action in conditioned avoidance: a computational approach. *Neuropsychopharmacology,* **29**, 1040-1049

Smith, A., Li, M., Becker, S., & Kapur, S., (2006) A computational model of the functional role of the ventral-striatal D2 receptor in the expression of previously acquired behaviors. *Neural Computation,* **17**, 361-395

Smith, A., Li, M., Becker, S., & Kapur, S., (2007) Linking animal models of psychosis to computational models of dopamine function. *Neuropsychopharmacology,* **32**, 54-66

Steele, G.L., (1990) *Common Lisp: The Language.* Digital Press

Stone, J.M., Morrison, P.D., & Pilowsky, L.S., (2007) Glutamate and dopamine dysregulation in schizophrenia – a synthesis and selected review. *J. of Psychopharmacology,* **21(4)**, 440-452

Sutton, R.S., (1988) Learning to Predict by the Methods of Temporal Differences. *Machine Learning* **3**, 9-44

Sutton, R.S., & Barto, A.G., (1998) *Reinforcement learning,* MIT Press

Talamini, L.M., Meeter, M., Elvevag, B., Murre, J.M.J., & Goldberg, T.E., (2005) Reduced parahippocampal connectivity produces schizophrenia-like memory deficits in simulated neural circuits with reduced parahippocampal connectivity. *Arcg. Gen. Psychiatry,* **62**, 485-493

Toates, F., (1986) *Motivational Systems.* Cambridge, UK: Cambridge University Press

Toates, F., (1994) The interaction of cognitive and stimulus-response processes in the control of behaviour. *Neuroscience and Behavioural Reviews,* **22(1)**, 9-83

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, **412**, 43-48

Wadenberg, M-LG, Soliman, A Vanderspek, S.C., & Kapur, S., (2001) Dopamine D2 receptor occupancy is a common mechanism underlying animal models of antipsychotics and their clinical effects. *Neuropsychopharmacology,* **25**, 633-641

Watkins, C. J. C. H., & Dayan, P., (1992) Q-learning. *Machine Learning*, **8**, 279-292

Weinberger, D.R., (1987) Implications of normal brain development for the pathogenesis of schizophrenia. *Arch Gen Psychiatry*, **44**, 660–669

Wickens, J., & Kotter, R., (1995) Cellular models of reinforcement. In *Models of Information Processing in the Basal Ganglia.* Edited by Houk, J.C., Davis, J.L., & Beiser, D.G. Cambridge, MA: MIT Press; 187-214

Wilson, C.J., (1990) Basal Ganglia. In G.M. Shepherd (ed.), *The Synaptic Organization of the Brain.* New York: Oxford University Press, 279-316

Winterer, G., & Weinberger, D.R., (2004) Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends in Neurosciences,* **27(11)**, 683-690

Wise, R.A., et al. (1978) Neuroleptic-induced anhedonia in rats: pimozide blocks reward quality of food. *Science,* **201**, 262-264

Wyvell, C.L., & Berridge, K.C., (2000) Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: Enhancement of reward 'wanting' without response liking or response reinforcement. *Journal of Neuroscience,* **20(21)**, 9122-9130

Yin, H.H., and Knowlton, B.J., (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience,* **7**, 464-476

Yin, H.H., Knowlton, B.J., & Balleine, B.W., (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neuroscience,* **19**, 181-189

Yin, H.H., Ostlund, S.B., Knowlton, B.J., & Balleine, B.W., (2005) The role of the dorsomedial striatum in instrumental conditioning. *Eur. Journal Neuroscience,* **22**, 513-523

Zakzanis, K.K., & Hansen, K.T., (1998) Dopamine D2 densities and the schizophrenic brain. *Schizophrenia Research,* **32**, 201-206

Appendix I

Chapter 4.3.3.1 An Example of the Learning Process

Appendix I contain raw data from the simulation detailed in Chapter 4.3.3.1 for runs 1-3, 15-16 and 27-29 through the maze in Figure 4.6. The data is for actual moves taken and does not include considerations of moves not made, which are an artefact of the model. The information includes: (i) the state from which the move is made; (ii) the state moved to (iii) the TD error, δ(t); (iv) the values, $V(s_t)$, for each state in the maze; and (iii) a brief explanation of how progression through the maze occurs.

Moved from state	to	$\delta(t)$	$V(s_t)$ S1	$V(s_t)$ S2	$V(s_t)$ S3	$V(s_t)$ S4	$V(s_t)$ S5	$V(s_t)$ S6	$V(s_t)$ S7	$V(s_t)$ S8	$V(s_t)$ S0	**Reward**
Run 1												
0	1	0	0	0	0	0	0	0	0	0	0	Rat in a one way maze, beginning in S0 and thinking ahead of possible move to next state S1. Moves are made by chance according to a sigmoid decision curve, P = $(1 + e^{-m(\delta(t) - b)})^{-1}$, where moves are more likely when $\delta(t)$ is high. Moves are eventually made to subsequent states.
1	2	0	0	0	0	0	0	0	0	0	0	
2	3	0	0	0	0	0	0	0	0	0	0	
3	4	0	0	0	0	0	0	0	0	0	0	
4	5	0	0	0	0	0	0	0	0	0	0	
5	6	0	0	0	0	0	0	0	0	0	0	
6	7	0	0	0	0	0	0	0	0	0	0	Has moved to S7 and is thinking ahead of move to S8 where there is a reward of 1, but has not yet encountered reward $\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t)$ $= 0 + 0 - 0 = 0$ **R1 GIVEN**
7	8	**1**	0	0	0	0	0	0	0	0	0	Have moved to reward state S8 and is considering move to next state 0. Only gets positive $\delta(t)$ when reward is received $\boldsymbol{\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t)}$ $\boldsymbol{= 1 + 0 - 0 = 1}$
8	0	0	0	0	0	0	0	0	0	**0.5**	0	Consider move to S0 No positive $\delta(t)$, $\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t)$ ($0 + 0 - 0 = 0$) $V(s_t)$ is updated only when actual move made away from reward state. $\boldsymbol{V(s_i) \leftarrow V(s_i) + \alpha\, \delta(t)}$ **(alpha = 0.5)** **Update is 0 + (0.5 x 1) = 0.5**
Run 2												
0	1	0	0	0	0	0	0	0	0	0.5	0	$V(s_t)$ remains stored
1	2	0	0	0	0	0	0	0	0	0.5	0	
2	3	0	0	0	0	0	0	0	0	0.5	0	

Moved from	**to**	$\delta(t)$	$V(s_t)$ **S1**	$V(s_t)$ **S2**	$V(s_t)$ **S3**	$V(s_t)$ **S4**	$V(s_t)$ **S5**	$V(s_t)$ **S6**	$V(s_t)$ **S7**	$V(s_t)$ **S8**	$V(s_t)$ **S0**	**Reward**
3	4	0	0	0	0	0	0	0	0	0.5	0	
4	5	0	0	0	0	0	0	0	0	0.5	0	$V(s_t)$ remains stored
5	6	0	0	0	0	0	0	0	0	0.5	0	
6	7	**0.5**	0	0	0	0	0	0	0	0.5	0	Have moved to S7 and is thinking ahead of move to S8 where there is a reward of 1. Has encountered reward before, the effect of which is stored in $V(s_t)$ as 0.5. **$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t)$** **0 + 0.5 – 0 = 0.5** **R1 GIVEN**
7	8	**0.5**	0	0	0	0	0	0	**0.25**	0.5	0	Have moved to reward state S8 and is considering move to next state 0. Reward of 1 is received a second time in S8, which is reflected in a positive $\delta(t)$ **$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t)$** **1 + 0 – 0.5 = 0.5** Additionally, $V(s_t)$ of S7 is updated once the rat has moved away from that state **$V(s_i) \leftarrow V(s_i) + \alpha\,\delta(t)$** **Update is 0 + (0.5 x 0.5) = 0.25**
8	0	0	0	0	0	0	0	0	0.25	**0.75**	0	$V(s_t)$ of S8 is updated when rat moves away from that state **$V(s_i) \leftarrow V(s_i) + \alpha\,\delta(t)$** **Update is 0.5 + (0.5 x 0.5) = 0.75**
Run 3												
0	1	0	0	0	0	0	0	0	0.25	0.75	0	
1	2	0	0	0	0	0	0	0	0.25	0.75	0	$V(s_t)$ remains stored. Effects of δ(t) and $V(s_t)$ are seen earlier in transition from S5 to S6 and in S6, respectively
2	3	0	0	0	0	0	0	0	0.25	0.75	0	
3	4	0	0	0	0	0	0	0	0.25	0.75	0	
4	5	0	0	0	0	0	0	0	0.25	0.75	0	
5	6	0.25	0	0	0	0	0	0	0.25	0.75	0	
6	7	0.5	0	0	0	0	0	0.12	0.25	0.75	0	**R1 GIVEN**

Moved from	to	$\delta(t)$	$V(s_t)$ S1	$V(s_t)$ S2	$V(s_t)$ S3	$V(s_t)$ S4	$V(s_t)$ S5	$V(s_t)$ S6	$V(s_t)$ S7	$V(s_t)$ S8	$V(s_t)$ S0	Reward
7	8	0.25	0	0	0	0	0	0.12	0.5	0.75	0	
8	0	0	0	0	0	0	0	0.12	0.5	0.87	0	
RUNS 4-14												
Run 15												
0	1	0.21	0.4	0.6	0.79	0.91	0.97	0.99	1	1	0	
1	2	0.18	0.5	0.6	0.79	0.91	0.97	0.99	1	1	0	By run 15 δ(t) is at a maximum of 0.5 at state transition S8-S0, where it is already the dominant CS. The V(st) for each state is rising and states S6, S7 and S8 have already achieved a maximum V(st) of 1.
2	3	0.12	0.5	0.7	0.79	0.91	0.97	0.99	1	1	0	
3	4	0.06	0.5	0.7	0.85	0.91	0.97	0.99	1	1	0	
4	5	0.02	0.5	0.7	0.85	0.94	0.97	0.99	1	1	0	
5	6	0.01	0.5	0.7	0.85	0.94	0.98	0.99	1	1	0	
6	7	0	0.5	0.7	0.85	0.94	0.98	1	1	1	0	**R1 GIVEN**
7	8	0	0.5	0.7	0.85	0.94	0.98	1	1	1	0	
8	0	0.5	0.5	0.7	0.85	0.94	0.98	1	1	1	0	
Run 16												
0	1	0.2	0.5	0.7	0.85	0.94	0.98	1	1	1	0	
1	2	0.15	0.6	0.7	0.85	0.94	0.98	1	1	1	0	
2	3	0.09	0.6	0.77	0.85	0.94	0.98	1	1	1	0	
3	4	0.04	0.6	0.77	0.89	0.94	0.98	1	1	1	0	
4	5	0.01	0.6	0.77	0.89	0.96	0.98	1	1	1	0	
5	6	0	0.6	0.77	0.89	0.96	0.99	1	1	1	0	
6	7	0	0.6	0.77	0.89	0.96	0.99	1	1	1	0	**R1 GIVEN**
7	8	0	0.6	0.77	0.89	0.96	0.99	1	1	1	0	
8	0	0.6	0.6	0.77	0.89	0.96	0.99	1	1	1	0	
RUNS 17-26												
Run 27												
0	1	0.01	0.99	1	1	1	1	1	1	1	0	
1	2	0	0.99	1	1	1	1	1	1	1	0	
2	3	0	0.99	1	1	1	1	1	1	1	0	
3	4	0	0.99	1	1	1	1	1	1	1	0	
4	5	0	0.99	1	1	1	1	1	1	1	0	
5	6	0	0.99	1	1	1	1	1	1	1	0	

Moved from	to	δ(t)	V(s_t) S1	V(s_t) S2	V(s_t) S3	V(s_t) S4	V(s_t) S5	V(s_t) S6	V(s_t) S7	V(s_t) S8	V(s_t) S0	Reward
6	7	0	0.99	1	1	1	1	1	1	1	0	R1 GIVEN
7	8	0	0.99	1	1	1	1	1	1	1	0	
8	0	0.99	0.99	1	1	1	1	1	1	1	0	
Run 28												
0	1	0.01	0.99	1	1	1	1	1	1	1	0	
1	2	0	0.99	1	1	1	1	1	1	1	0	
2	3	0	0.99	1	1	1	1	1	1	1	0	
3	4	0	0.99	1	1	1	1	1	1	1	0	
4	5	0	0.99	1	1	1	1	1	1	1	0	
5	6	0	0.99	1	1	1	1	1	1	1	0	
6	7	0	0.99	1	1	1	1	1	1	1	0	R1 GIVEN
7	8	0	0.99	1	1	1	1	1	1	1	0	
8	0	0.99	0.99	1	1	1	1	1	1	1	0	
Run 29												
0	1	0	0.99	1	1	1	1	1	1	1	0	
1	2	0	1	1	1	1	1	1	1	1	0	
2	3	0	1	1	1	1	1	1	1	1	0	
3	4	0	1	1	1	1	1	1	1	1	0	
4	5	0	1	1	1	1	1	1	1	1	0	
5	6	0	1	1	1	1	1	1	1	1	0	
6	7	0	1	1	1	1	1	1	1	1	0	R1 GIVEN
7	8	0	1	1	1	1	1	1	1	1	0	
8	0	1	1	1	1	1	1	1	1	1	0	

By run 29 complete learning has taken place:
δ(t) in state transition S8-S0 = 1 and has become the CS.
While V(s_t) for each state is 1 (except for reset state S0).

Appendix II

How Do Computational Models of the Role of Dopamine as a Reward Prediction Error Map on to Current Dopamine Theories of Schizophrenia? (Chapter 4)
(Thurnham, Done, Davey & Frank 2006a)

Paper published in the proceedings of XXV111 Annual Conference of the Cognitive Science Society, Vancouver, Canada, 26-29 July 2006, pp 2263-2268, Lawrence Erlbaum Associates

How Do Computational Models of the Role of Dopamine as a Reward Prediction Error Map on to Current Dopamine Theories of Schizophrenia?

Angela J. Thurnham* (a.j.thurnham@herts.ac.uk), D. John Done (d.j.done@herts.ac.uk), Neil Davey* (n.davey@herts.ac.uk), Ray J. Frank* (r.j.frank@herts.ac.uk)**
School of Computer Science,* School of Psychology, **University of Hertfordshire,
College Lane, Hatfield, Hertfordshire. AL10 9AB United Kingdom

Abstract

A review of the current dopamine theories of schizophrenia reveals a likely imbalance between cortical and subcortical microcircuits due to an insufficient inhibitory brake, leading to a disruption of the dopamine system and the classic positive psychotic symptoms, negative symptoms and cognitive deficits associated with the disorder. Recent computational models have modelled the role of dopamine as a reward prediction error, using Temporal Difference and have successfully shown how these symptoms could arise from a disturbance to the dopamine system. We review these models in the light of dopamine theories of schizophrenia and highlight some of the major points that should be addressed by future computational models.

Keywords: Dopamine; Schizophrenia; Neurocomputational Modelling; Salience; Temporal Difference.

Theories of the role of dopamine over the last five years tend to converge on the idea that dopamine encodes a reward prediction error (RPE) of the discrepancy between actual and expected future reward. This discrepancy is used to drive learning towards actions which are necessary for survival in the real world (Schultz, 1998), and it is likely that disruption to this system gives rise to an abnormality in information processing by dopamine and some of the symptoms currently associated with schizophrenia, particularly psychosis and deficits in working memory. Temporal Difference Learning (Sutton, 1988; Sutton & Barto, 1998), a form of Reinforcement Learning Theory, provides an explicit method of modelling and quantifying the Reward Prediction, or Temporal Difference (TD), error (Schultz, Dayan & Montague, 1997; Hollerman & Schultz, 1998) and can be used as a valid computational implementation of the RPE for neural network simulations. While dopamine should not be viewed in isolation, but seen to be working in concert with other neurotransmitters, such as glutamate and GABA (Abi-Dhargham, 2004; Carlsson, Waters, Holm-Waters, Tedroff, Nilsson & Carlsson, 2001; Winterer & Weinberger, 2004), there are still attributes and deficiencies that can be strongly linked to dopamine activity.

The role of dopamine, and the possible location and nature of the dysfunction, presented in theories of schizophrenia by Carlsson et al, (2001); Kapur, (2003); Abi-Dhargham, (2004) and Winterer & Weinberger, (2004), are discussed in the first section on dopamine theories of schizophrenia below. The second section relates specifically to computational models, particularly existing connectionist models of dopamine as a reward prediction, or TD error, including evidence that the RPE model of dopamine activity applies to humans as well as primates. The biological plausibility of existing neural network models by Cohen & Servan-Schreiber, (1992); Braver Barch & Cohen, (1999); Suri & Schultz, (1999); Rougier, Noelle, Braver, Cohen & O'Reilly, (2005) and O'Reilly & Frank, (2006) are then discussed in the light of the afore-mentioned dopamine theories of schizophrenia. Finally, we conclude with four major questions arising from recent dopamine theories of schizophrenia that remain to be addressed by current computational models.

Dopamine Theories of Schizophrenia

Role of Dopamine

It is generally agreed that dopamine enables the ability to focus on task relevant information. Current theories of the effects of dopamine on behaviour focus on the role of dopamine as a neuromodulator in Reinforcement Learning, where organisms learn to organise their behaviour under the influence of goals, and expected future reward is believed to drive action selection, as seen during conditioning. Neurophysiological recordings of single dopamine neurons in primates have identified a reward prediction error signal of the discrepancy between actual and expected future reward (Schultz et al., 1997; Hollerman & Schultz, 1998). In conditioning, before learning, this phasic burst of dopamine occurs at the time an unexpected reward is encountered. As trials progress and learning continues, the reward becomes more and more predictable and the phasic burst effectively moves backwards towards the time the conditioned stimulus (CS) occurs. Eventually, when full learning has taken place, the CS will elicit the same phasic response previously associated with the unexpected reward.

In particular, evidence suggests that the dopamine system may mediate the Incentive Salience of rewards, modulating their motivational value, which is dissociable from hedonia and reward learning (Berridge & Robinson, 1998). The modern Incentive Salience Theory distinguishes ‘wanting’ from ‘liking,’ and the dopamine system is regarded as that which calculates the ‘want’ rather than the ‘act’ parts of instrumental behaviour. Kapur’s framework of psychosis builds on this hypothesis, and sees the role of dopamine as mediating the salience of both internal and environmental representations.

Abi-Dhargham refers to the dopamine hypothesis of schizophrenia and uses neuroimaging techniques such as SPECT and PET to monitor changes in synaptic dopamine levels. Using data from electrophysiological techniques on a smaller timescale, Winterer & Weinberger are more explicit and refer to the apparent ability of dopamine to optimise the signal to noise ratio (SNR) of local cortical microcircuits.

Carlsson et al. take a wider view and see dopamine as one of many possible dysfunctional neurotransmitters affected in the brain in schizophrenia. Pharmacological evidence suggests small differences in the fragile balance between multi-neurotransmitters at various points in local cortical microcircuits leads to many of both the positive and negative symptoms associated with the disorder. They posit that although there may be an elevated baseline release of dopamine in schizophrenia, it is possibly secondary to hypoglutamatergia.

Where is the Dysfunction?

One of the few biological disturbances that have been identified in schizophrenic patients is an impaired dopamine system, which traditionally has been of an increase in dopamine signaling in the striatum, leading to psychosis (Winterer & Weinberger, 2004). The Dopamine Hypothesis of Schizophrenia arose as a result of two major findings: (i) Exposure to dopamine receptor agonists, such as amphetamine, induces psychosis, and (ii) antipsychotic drugs provide an antipsychotic effect by blocking dopamine receptors (Abi-Dhargham, 2004). Current views still posit deficits due to an increase in dopamine; however it is the site of the excess that is controversial. Kapur refers to a general excess, while Abi-Dhargham refers to the traditional cortical/subcortical imbalance, with an excess in the subcortex and a deficit in the cortex. Winterer & Weinberger, on the other hand, suggest that it may be the cortical and not striatal microcircuits that give rise to abnormal dopamine signaling. Carlsson et al. also refer to possible cortical steering of subcortical systems, but by glutamate action. However, all agree that it is the resulting imbalance that leads to the problem and, overall, current opinion would imply that it is the imbalance in the dopamine circuits between cortical and striatal brain regions that leads to the dysfunction, while the actual point of the dysfunction remains controversial. Indeed it may be that disruption at different points in the circuits may lead to different symptoms or cognitive deficits and computational modelling may help us to answer these questions.

Carlsson et al. refer to a secondary general elevated baseline release of dopamine in schizophrenia, possibly due to a primary disturbance in cortical glutamate/GABA mediated steering of monoamine subcortical systems, (including dopamine). There is a direct glutamate pathway which acts as an accelerator and an indirect glutamate pathway that activates GABA and is an effective brake on the activity of monoamines. It is the balance between accelerator and brake that maintains stability and glutamatergic failure in the cerebral cortex may lead to negative symptoms, while glutamatergic failure in the basal ganglia would favour positive symptoms. These result from dysregulation of the dopamine system.

Abi-Dhargham and Winterer & Weinberger also refer to such an insufficient inhibitory brake as the possible nature of the dysfunction. Abi-Dhargham refers to a hypostimulation in the cortex of D1 receptors which causes a deficit in working memory, and a hyperstimulation in the subcortex of D2 receptors which leads to psychotic symptoms, as a result of the reduced cortical brake. Winterer & Weinberger refer principally to a reduced prefrontal dopamine D1/D2 receptor activation ratio which leads to a lower cortical SNR. They posit that normally it is the D1 receptors that dominate, but in schizophrenia D2 receptors dominate, and as a result of the primary disturbance, secondary effects will occur subcortically in the striatum leading to contextually inappropriate, inflexible and bizarre behavioural routines.

All these theories seem to point to an imbalance in the dopamine system between the cortical and subcortical areas, due to an insufficient inhibitory brake system, with negative symptoms occurring as a result of disturbance to the cortex and positive symptoms as a disturbance to subcortical areas.

What is the Dysfunction?

Kapur posits that psychosis is a state of aberrant salience, where excess levels of dopamine are no longer stimulus-linked and context-driven. Delusions (paranoia, aliens interfering with one’s brain), and hallucinations (hearing voices), may arise then as a result of the individual attempting to provide their own explanations for experiences which come out of the blue and are imbued with high importance. This is in keeping with an earlier theory of schizophrenia by Maher (1988) that patients make normal attributions, or reasoned normally to abnormal experiences, i.e., subcortical abnormality with normal cortical function. It is known that patients with schizophrenia suffer from a wide-spread cognitive dysfunction that affects memory, executive functioning and attention (Bilder et al., 2000; McKenna, 1997). However, there seems to be a dissociation between the psychotic experiences (delusions, hallucinations) and cognitive dysfunction. The latter occur well in advance of onset symptoms, and the trajectory of symptom recovery is not matched by cognitive recovery (Harvey, Koren, Reichenberg & Bowie, 2005). Traditional cognitive models of schizophrenia based on cognitive dysfunction in memory/attention/executive dysfunction have poor face validity when used to explain the spontaneous experiences (delusions/hallucinations) which are bizarre, or strange, since these are unrelated to past experience and stored memories (Simpson, Done, Valeé-Tourangeau , 2002).

The recent developments in understanding the role of dopamine in salience allocation do permit the formulation of cognitive neuroscience models which can integrate both Maher's theory with the known cognitive dysfunctions in schizophrenia. Computing the salience of stimuli (both external and internal, such as thoughts/ideas) is probably achieved by midbrain/ventral striatal dopamine systems rather than cortical ones (O'Doherty Dayan, Schultz, Deichmann, Friston & Dolan, 2004). This is the 'critic' in models of the dopamine system in the basal ganglia (Montague, Hyman & Cohen, 2004; Sutton & Barto, 1998). In schizophrenia we posit that within the critic, the signal (winner) is distinguished from the noise (losers). This signal is then transmitted to other systems (e.g., 'actor' in dorsal striatum), or cortical systems, such as the dorsolateral prefrontal cortex (DLPFC) responsible for various attributional, memory, executive and attentional processes (Durstewitz, Kelc & Gunturkun, 1999). Thus stimuli, or experienced episodes, which are unimportant, can be imbued with a high degree of salience by the critic in the ventral striatum/midbrain. This provides the spontaneous experience imbued with importance. High variance in the level of background dopamine activity would also mean that these experiences occur from time to time, but not all of the time. Dopamine abnormalities in DLPFC would not only account for the neuropsychological deficits found in schizophrenia but they could also integrate the abnormal experiences into dysfunctional attributional, executive and memory systems. We can crudely equate these dual roles as being due to dopamine abnormalities in the midbrain/striatum and cortex respectively, as outlined previously in the theories of Abi-Dhargham (2004) and Winterer and Weinberger (2005). As described previously, the interaction between these different levels means that they cannot operate independently, but in consort. This permits a more tractable model of the psychology of schizophrenia, i.e. a model of both symptoms and classical cognitive abnormalities.

Antipsychotics

The action of antipsychotic drugs can help further understand what is going wrong with the dopamine system. Kapur proposes that antipsychotics dampen 'aberrant saliences' by blocking excess dopamine, leading to an attenuation of motivational salience of ideas and perceptions. In this way antipsychotics remove the degree to which symptoms occupy the mind, but not the core content of the symptom. They simply provide a neurochemical balance where dopamine levels return to normal, new aberrant saliences are less likely to form and existing ones are more likely to stop. It is only in the weeks to come that an individual may work through and resolve their delusions in their own time. In this way the delusions and hallucinations may be deconstructed, but this is not always the case as some patients are never able to resolve their symptoms psychologically.

Abi-Dhargham does not refer to antipsychotic action, but Winterer & Weinberger deviate from the traditional view of antipsychotic action on D2 receptors in the striatum and, using evidence from imaging studies, suggest that antipsychotics may exert actions instead through D2 receptor blockade in the cortex. Carlsson et al. refer to the adverse effects of classic antipsychotics which lead to hypodopaminergia in patients in remission from their positive symptoms that cause failure of the reward system leading to dsyphoria and anhedonia; and negative effects, such as catatonia and cognitive deficits. They have developed both partial dopamine-receptor agonists, and antagonists, that act on D2 receptors, stabilising the elevated dopamine levels without causing hypodopaminergia. However, they do not refer to the exact site of those receptors.

Both Carlsson et al. and Winterer & Weinberger focus on D2 receptor blockade as means of resolving the dopamine imbalance which leads to psychotic symptoms, but the exact site of impact remains unclear.

Interim Conclusions

Dopamine provides a RPE signal of the discrepancy between actual and expected future reward and it would appear to be an imbalance between cortical and subcortical microcircuits that leads to a dysfunction of the dopamine system. However, the actual point of the dysfunction remains controversial. Recently it has been suggested that it may be cortical microcircuits that give rise to abnormal dopamine signaling, with secondary downstream subcortical deficits, instead of the traditional view of a primary subcortical disturbance (Winterer & Weinberger, 2004).

It is generally agreed that the resulting imbalance may result from an insufficient inhibitory brake system leading to either a hypostimulation in the cortex of D1 receptors and a hyperstimulation in the subcortex of D2 receptors (Abi-Dhargham, 2004), or a reduced prefrontal dopamine D1/D2 receptor activation ratio, in which D2 receptors dominate, which leads primarily to a lower cortical SNR (Winterer & Weinberger, 2004). D2 receptor blockade would appear to be important in restoring the cortical/subcortical imbalance (Carlsson et al., 2001; Winterer & Weinberger, 2004).

Furthermore, positive psychotic symptoms arise from either a primary subcortical hyperstimulation of dopamine receptors (Abi-Dhargham, 2004), or secondary effects of either reduced cortical SNR on subcortical systems (Winterer & Weinberger, 2004), or cortical gluatamate/GABA steering of subcortical systems (Carlsson et al., 2001). Negative symptoms and working memory deficits are thought to result from either hypostimulation of D1 receptors (Abi-Dhargham, 2004) or reduced prefrontal dopamine D1/D2 receptor activation ratio with D2 receptors dominating (Winterer & Weinberger, 2004).

Computational Models of Dopamine as a Reward Prediction/Temporal Difference Error Signal

Several computational models of the role of dopamine as a RPE have incorporated Temporal Difference (TD) Learning (Sutton, 1988), a form of Reinforcement Learning Theory, which provides an explicit method of modelling and quantifying the Reward Prediction error (Schultz et al, 1997; Hollerman & Schultz, 1998; Montague et al., 2004). Specifically, it provides a mathematical interpretation of how dopamine is thought to mediate reward-processing and reward-dependent learning, thus optimising behaviour in an environment. A class of TD models, known as actor-critic models (Sutton & Barto, 1998), have been adapted so that expected future reward is equivalent to incentive salience (McClure, Daw & Montague, 2003; Montague et al., 2004). Here, the error signal generated is used in two ways: (i) The 'critic' - as a prediction error or learning signal used to create better estimates of future reward. (ii) The 'actor' - to bias action selection towards situations that predict the best reward.

It is possible that the same RPE is signaled from dopamine neurons in both the ventral tegmental area (VTA) and substantia nigra (SN). The signal is used in two ways depending on the route it takes, with the projections from VTA to ventral striatum as the 'critic' in TD models, associated with reward and motivation, and projections from SN to dorsal striatum as the 'actor', associated with motor control (O'Doherty et al., 2004; Daw, Niv & Dayan, 2005). The dopamine pathways are arranged in cortical/subcortical circuit loops involving prefrontal cortex (Alexander et al., 1985), and it is in the cortical areas that dopamine dysfunction is believed to have an effect on working memory.

It has also been suggested that TD Learning can help with the dynamic choice of action selection to obtain natural rewards required for survival. As well as assisting in the learning process, it has been suggested that the dopamine signal can be used in decision-making, when full learning has taken place, to bias the choice of actions that lead to better rewards in another actor/critic model by Schultz et al., (1997). When full learning has taken place the RPE will be zero and fluctuations above and below that point will provide important ongoing evaluations in the environment of salience which can be assessed quickly according to whether the fluctuations represent potential actions that are better or worse than expected. In this way an instant comparison can be made between well-learnt possibilities; all that is required is a simple behaviour strategy, to choose those actions associated with increased dopaminergic activity and incentive salience, and avoid those of low salience where dopaminergic activity is decreased. In this way, a damaged dopamine system could explain why adults become slow to do things that they used to do so easily. Their ability to make these instant comparisons or to maintain context would become impaired, and lead to some of the cognitive deficits associated with schizophrenia, such as poor performance in the Wisconsin Card Sorting Test (WCST) or the 1-2-AX Test, where it is important to maintain context.

TD models have proved to be very successful in many behavioural tasks and are used extensively in robotics to enable learning and reacting to an environment. However, while they are often more efficient than other reinforcement learning algorithms (Suri & Schultz, 1999), complications may arise when unpredictable events occur, which break the learning chains constructed through prediction (O'Reilly & Frank, 2006), and this has led to some researchers who have previously used TD, seeking alternative combinations of algorithms as learning mechanism (Hazy, Frank & O'Reilly, In Press).

Evidence for Role of Dopamine as a Reward Prediction Error/Temporal Difference Signal

Functional imaging techniques have provided evidence that the RPE model of dopamine activity applies to human reward learning, and not just to primates, as seen in neurophysiological recordings by Schultz and colleagues mentioned above. Transient learning-related changes associated with the ‘critic’ have been identified in the brains of humans subjected to classical conditioning procedures, in the ventral striatum (putamen) (McClure, Berns & Montague, 2003; O’Doherty, Dayan, Friston, Critchley & Dolan, 2003). While O’Doherty et al, (2004) showed that activity in the dorsal striatum is associated with the ‘actor’ only, as no activity was seen in this area unless an action was required.

In addition, activation patterns consistent with predictions from a TD model of learning have also been recorded in the orbital frontal cortex (O’Doherty et al., 2003). Furthermore, Seymour et al. (2004) have used fMRI to show that neural activity in the ventral striatum and the anterior insula corresponds to the signals for sequential learning predicted by TD models, in humans in higher-order learning.

How Do Existing Computational Models Compare with the Cortical/Subcortical Debate of Theories for Schizophrenia?

The early connectionist model by Cohen & Servan-Schreiber (1992) and some biophysically detailed neural network models (Brunel & Wang, 2001; Durstewitz et al., 1999; Durstewitz, Seamans & Sejnowski, 2000) have modelled dopamine as a neuromodulator crucial for optimising the SNR thought to enhance working memory. This model is limited as it simulates only the DLPFC circuits, but not the critic in striatum and midbrain. Other models have incorporated Reinforcement Learning methods and modelled dopamine as a RPE signal, which can be effectively modelled using TD Learning (Braver et al., 1999; Suri & Schultz, 1999; Rougier et al., 2005).

As previously mentioned, it is believed that the actual point of dysfunction in subcortical/cortical microcircuits remains controversial. Cohen and colleagues have modelled working memory deficits, simulating the continuous performance test (CPT) (Cohen & Servan-Schreiber, 1992; Braver et al., 1999), with the latter model, a more powerful and complete theory of the mechanism of cognitive control, incorporating both TD learning and gating functions for dopamine, where dopamine was seen as a unitary function which enabled an organism to predict and respond appropriately to events that led to reward. In this later model schizophrenia was seen as an impaired ability to internally represent, maintain and update context relating to working memory from increased noise in the dopamine system, focusing particularly on the prefrontal cortex. The model suggested that reduced phasic activity, i.e., reduced update to active memory, led to perservatory behaviour; while increased phasic activity, i.e., increased update, led to poor interference control, and therefore distractibility. Additionally, increased tonic (or longer-term background) activity led to delay related decay of active memory, and therefore maintenance deficits. Both perseverations and distractibility are known disturbances to the prefrontal cortex and are typical symptoms of Schizophrenia, along with poor maintenance control. Perseveratory behaviour occurs when a patient becomes preoccupied with a task and is unable to change strategy or appropriately update goal representations, while distractibility is the inability to concentrate or focus on the task at hand. This model posits that both perseverations and distractibility are due to impairments in phasic dopaminergic activity which affect working memory. However, the model is of two very different systems in the brain doing different jobs and possibly coding for two different things; salience in the midbrain and how it possibly affects working memory in the prefrontal cortex. It is important, therefore, to investigate how these behaviours relate to each other and it is this interaction that will be explored in the current research.

The increased noise could be due to an imbalance between cortical and subcortical structures due to the insufficient inhibitory brake system on the dopamine system. However, the model has a simple architecture with no hidden layers and modules containing between one and four neurons. The simple task is hard-wired and it is not a cognitive model.

Using a more sophisticated architecture, a neural network model by Suri & Schultz (1999) specifically modelled Wolfram Schultz's work on the response of dopamine neurons in the striatum to reward-related stimuli using a 'critic', which computed and sent a TD error to an 'actor', which governed behaviour. The model did not refer explicitly to the prefrontal cortex, but showed that a reinforcement signal without RPE led to perseverations, and sustained reductions of reinforcement signal led to a loss of learned behaviour as seen in Parkinson's disease and lesioned animals.

O'Reilly and colleagues have produced a range of biophysically detailed cognitive connectionist models using O'Reilly and Munakata's Leabra algorithm, which combines error-driven and Hebbian learning with k-Winners-Take-All inhibitory competition (O'Reilly & Munakata, 2000). These models are capable of implementing the learning and gating ideas of Braver et al. (1999) mentioned above, incorporating a brake and accelerator system. A model of dynamic DA modulation in the basal ganglia by Frank (2005) separates out the roles of the D1 and D2 receptors applicable to Parkinson's disease, without using TD. The XT model (Rougier et al., 2005) uses an adaptive gating mechanism, based on an adaptive critic unit, driven by TD Learning and relates specifically to how the biological mechanisms of the prefrontal cortex support flexible cognitive control. Dorsal Lateral Prefrontal Cortex lesions were simulated by removing units and asymmetric training, resulting in perseverations in prefrontal cortex, as seen in the WCST and Stroop tasks. However, in this and all previous models, it was necessary for the dynamic gating of the basal ganglia to be hard-wired. The Prefrontal Basal Ganglia Working Memory model of learning (O'Reilly & Frank, 2006) incorporates the dynamic interactions between the prefrontal cortex and the basal ganglia in working memory, and in doing so, abandons the use of TD in favour of an alternate associative Pavlovian mechanism. Here dopamine signals reward association and not reward prediction. Instead of using TD prediction chains over successive time-steps, which they claim break down when modelling complicated tasks such as the 1-2-AX task, the new algorithm uses the Rescorla-Wagner/Delta-rule algorithm trained by the unconditioned stimulus for the current time-step. However, this model is of learning and has not been used to model dysfunction so far.

Conclusions

The following important questions arising from recent dopamine theories of schizophrenia that remain to be addressed by current computational models:

1. Is it the cortical microcircuits that give rise to abnormal dopamine signaling with secondary downstream subcortical deficits (Winterer & Weinberger, 2004) or the traditional view of a primary subcortical disturbance?

2. For the most part connectionist models to date do not differentiate between D1 and D2 dopamine receptors, or locate the point(s) of dysfunction in the local microcircuits that give rise to a possible cortical/subcortical imbalance. They do not distinguish between the theories of Abi-Dhargham and Winterer & Weinberger, of either: (i) A hypostimulation in the cortex of D1 receptors and a hyperstimulation in the subcortex of D2 receptors (Abi-Dhargham, 2004), or (ii) A reduced prefrontal dopamine D1/D2 receptor activation ratio, in which D2 receptors dominate, leading primarily to a lower cortical SNR (Winterer & Weinberger, 2004).

3. Do positive psychotic symptoms arise from either: (i) A primary subcortical hyperstimulation of dopamine receptors (Abi-Dhargham, 2004)? Or (ii) Secondary effects of either reduced cortical SNR on subcortical systems (Winterer & Weinberger, 2004) or cortical gluatamate/GABA steering of subcortical systems (Carlsson et al., 2001)?

4. Do negative symptoms and working memory deficits result from either: (i) Hypostimulation of D1 receptors (Abi-Dhargham, 2004)? Or (ii) Reduced prefrontal dopamine D1/D2 receptor activation ratio with D2 receptors dominating (Winterer & Weinberger, 2004)?

Furthermore, while enormous progress has been made regarding flexible, self-organising cognitive control, without the need for a homunculus (Rougier et al., 2005; O'Reilly & Frank, 2006), it remains to be seen whether it is prudent to abandon TD Learning, which has been shown to be an effective model of RPE (see above), or whether the problems in the break down of chaining can be overcome by some other means.

References

Abi-Dhargham, A. (2004) Do we still believe in the dopamine hypothesis? New data bring new evidence *Neuropsychopharmacology,* 7 (supplement 1), S1-S5.

Alexander, G. E., & Delong, M. R. (1985) Microstimulation of the primate neostriatum. Somatotopic organization of striatal microexcitable zones and their relation to neuronal response properties. *J. Neurophysiol.* 53: 1417–1430.

Berridge, K.C., & Robinson, T.E. (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 28, 309-369.

Bilder et al. (2000) Neuropsychology of first-episode schizophrenia: initial characterization and clinical correlates. *Am. J. Psychiatry*, 157(4), 549-559.

Braver, T.S., Barch, D.M., & Cohen, J.D. (1999) Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry,* 46**,** 312-328.

Brunel, N., & Wang, X. (2001) Effects of neuromodulation in a corticalnetwork model of object working memory dominated by recurrent inhibition. *J. Computational Neuroscience,* 11, 63-85.

Carlsson, A., Waters, N., Holm-Waters, S., Tedroff, J., Nilsson, M., & Carlsson, M.L. (2001) Interactions between monoamines, glutamate and GABA in Schizophrenia. *Ann. Rev.Pharm. Toxicol,*41(237), 237-60.

Cohen, J.D., & Servan-Schreiber, D. (1992) Context, Cortex, and Dopamine: A Connectionist Approach to Behavior and Biology in Schizophrenia. *Psychological Review* 1**,** 45-77.

Daw, N.D., Niv, Y., & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioural control. *Nature Neuroscience,* 8(12), 1704-1711.

Durstewitz, D., Kelc, M., & Gunturkun, O. (1999) A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience,* 19(7), 2807-2822.

Durstewitz, D., Seamans, J.K., & Sejnowski, T.J. (2000) Dopamine mediated stabilization of delay-period activity in a network model of prefrontal cortex. *J. Neurophysiology,* 83, 1733-1750.

Frank, M.J., (2005) Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cognitive Deficits in Medicated and Nonmedicated Parkinsonism. *J. of Cognitive Neuroscience,* 17, 51-72.

Harvey, P.D., Koren, D., Reichenberg, A., & Bowie, C.R. (2005) Negative Symptoms and Cognitive Deficits: What is the nature of their Relationship?, *Schizophrenia Bulletin*, (Advanced access 12 /10/05).

Hazy, T.E., Frank, M.J., & O'Reilly, R.C (In Press) Banishing the homunculus: Making working memory work. *Neuroscience.*

Hollerman, J.R., & Schultz, W. (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1,* 304-309.

Kapur, S., (2003) Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry,* 160(1), 13-23.

McClure, S.M., Berns, G.S., & Montague, P.R. (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron,* 38, 339-346.

McClure, S.M., Daw, N.D., & Montague, P.R., (2003) A computational substrate for incentive salience. *Trends in Neuroscience* 26(8)*,* 423-428.

McKenna, P.J. (1997) *Schizophrenia and related syndromes.* Hove: Psychology Press.

Maher, B.A. (1988) Anomalous experience and delusional thinking: The logic of explanations. In T.F.Oltmanns and B.A. Maher (Eds.), *Delusional beliefs* (pp. 15-33), New York: Wiley.

Montague, P.R., Hyman, S.E., & Cohen, J.D. (2004) Computational roles for dopamine in behavioral control. *Nature* 431*,* 760-767.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., & Dolan, R.J. (2003) Temporal difference models and reward-related learning in the human brain. *Neuron,* 38**,** 329-337.

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

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

O'Reilly, R.C., & Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the mind by simulating the brain.* MIT Press.

Rougier, N.P., Noelle, D.C., Braver, T.S., Cohen, J.D., & O'Reilly, R.C. (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. *PNAS,* 102(20), 7738-7343.

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

Schultz, W., Dayan, P., Montague, P.R. (1997) A Neural substrate of prediction and reward. *Science*, *275:* 5306, 1593-1599.

Seymour et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature,*429**,** 664-667.

Simpson J, Done D.J., Valeé-Tourangeau (1998) An Unreasoned Approach: A Critique of Research on Reasoning and Delusions. *Cog. Neuropsychiatry,* 3, 1-20.

Suri, R.E., & Schultz, W. (1999) A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neurosci.*, 91(3), 871-890.

Sutton, R.S. (1988) Learning to Predict by the Methods of Temporal Differences. *Machine Learning 3,* 9-44.

Sutton, R.S., & Barto, A.G. (1998) *Reinforcement learning,* MIT Press.

Winterer, G., & Weinberger, D.R. (2004) Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends in Neurosciences,* 27(11), 683-690.

Appendix III

Lisp Code for Simulation of McClure, Daw & Montague (2003) (Chapter 5)

```lisp
;set discount factor at 0.9
 (defvar *gamma* 0.9)

;this function generates the possible moves that the agent may take
(defun possible-move (state-transitions current-state)
   (let ((possible-next-states (nth current-state state-transitions)))
     (nth (random (length possible-next-states)) possible-next-states)))

;updates the value of a state following a move and outputs delta
(defun move-and-update (current-state state-transitions rewards Vs gamma bias m alpha &optional (attempts 0))
   (let* ((possible-next-state (possible-move state-transitions current-state))
          (delta (+ (nth current-state rewards) (* gamma (nth possible-next-state Vs)) (* -1 (nth current-state Vs))))
          (prob (sigma delta m bias)))
     (cond ((> prob (random 1.0))
            (format t "~,2f " delta)
            (incf (nth current-state Vs) (* alpha (- delta bias))) (list possible-next-state
                (incf attempts)))
           (T (move-and-update current-state state-transitions rewards Vs gamma bias m alpha (+ 1 attempts))))))

;this function initiates a selected number of state transitions through the maze and outputs the current state, the
;values of states and the total number of timesteps.
(defun go-all  (state state-transitions rewards Vs gamma number-of-steps &key (bias 0) (m 1) (alpha 0.5) &aux
(total 0) state-and-number)
   (dotimes (i number-of-steps)
     (setf state-and-number (move-and-update state state-transitions rewards Vs gamma bias m alpha))
     (setf state (car state-and-number))
     (incf total (cadr state-and-number))
     (format t " ~d  ~{~,2f ~}  ~d ~%" state Vs total))
   (format t "~%~%Total states examined ~d" total))

;defines sigmoid function
(defun sigma (delta m bias)
   (/ 1 (+ 1 (exp (* -1 m  (- delta bias))))))

;initializes parameters
(setf *McClure-Maze* '((1) (0 2) (1 3) (2 4) (0)))
(setf *McClure-rewards* '( 0 0 0 0 1))
(setf *McClure-Vs* '( 0 0 0 0 0))

;starts program
(go-all 0 *McClure-Maze* *McClure-rewards* *McClure-Vs* *gamma* (input number of state transitions,
default: 1000) (input m, default: 5) (input bias, default: 0))
```

Appendix IV

A Model of Dopamine and Uncertainty Using Temporal Difference (Chapter 6) (Thurnham, Done, Davey & Frank 2006b)

Paper published in the proceedings of XXV111 Annual Conference of the Cognitive Science Society, Vancouver, Canada, 26-29 July 2006, pp 2263-2268, Lawrence Erlbaum Associates

A Model of Dopamine and Uncertainty Using Temporal Difference

Angela J. Thurnham* (a.j.thurnham@herts.ac.uk), D. John Done**
(d.j.done@herts.ac.uk),
Neil Davey* (n.davey@herts.ac.uk), Ray J. Frank* (r.j.frank@herts.ac.uk)
School of Computer Science,* School of Psychology, **University of Hertfordshire,
College Lane, Hatfield, Hertfordshire. AL10 9AB United Kingdom

Abstract

Does dopamine code for uncertainty (Fiorillo, Tobler & Schultz, 2003; 2005) or is the sustained activation recorded from dopamine neurons a result of Temporal Difference (TD) backpropagating errors (Niv, Duff & Dayan, 2005)? An answer to this question could result in a better understanding of the nature of dopamine signaling, with implications for cognitive disorders, like Schizophrenia. A computer simulation of uncertainty incorporating TD Learning successfully modelled a Reinforcement Learning paradigm and the detailed effects demonstrated in single dopamine neuron recordings by Fiorillo et al. This alternate model provides further evidence that the sustained increase seen in dopamine firing, during uncertainty, is a result of averaging firing from dopamine neurons across trials, and is not normally found within individual trials, supporting the claims of Niv and colleagues.

Keywords: Dopamine; Uncertainty; Single Cell Recordings; Temporal Difference; Computer Simulation.

Dopamine and Uncertainty

Current theories of the effects of dopamine on behaviour focus on the role of dopamine in Reinforcement Learning, where organisms learn to organise their behaviour under the influence of goals, and expected future reward is believed to drive action selection (McClure, Daw & Montague, 2003; Montague, Dayan & Sejnowski, 1996; Schultz, Dayan & Montague, 1997; Suri & Schultz, 1999). Single cell recordings of dopamine neurons have identified a phasic dopamine burst of activity which is posited to be a reward prediction error (Schultz, 1998; Waelti, Dickinson & Schultz, 2001) and Temporal Difference (TD) Learning (Sutton, 1988; Sutton & Barto, 1998), a form of Reinforcement Learning, provides an explicit method of modelling and quantifying this error (Hollerman & Schultz, 1998; Schultz et al., 1997). It is likely that disruption to the dopamine system gives rise to an abnormality in information processing by dopamine and some of the symptoms currently associated with schizophrenia, particularly psychosis and deficits in working memory.

It has been posited that dopamine also codes for uncertainty (Fiorillo, Tobler & Schultz, 2003), as under conditions of maximum uncertainty, observations of single cell recordings have shown a sustained increase in activity from presentation of a conditioned stimulus (CS) to the expected time of a reward. They recorded the activity of neurons in two primates, identified as dopamine neurons from their electrophysiological characteristics, during a delay paradigm of classical conditioning to receive a fixed juice reward, while manipulating the probability of receipt of the reward. Two related but distinct parameters of reward were identified from the activation produced, after learning had taken place: (i) A phasic burst of activity, or reward prediction error, at the time of the expected reward, whose magnitude increased as probability decreased; and (ii) a new slower, sustained activity, above baseline, related to motivationally relevant stimuli, which developed with increasing levels of uncertainty, and varied with reward magnitude. Both effects were found to occur independently within a single population of dopamine neurons.

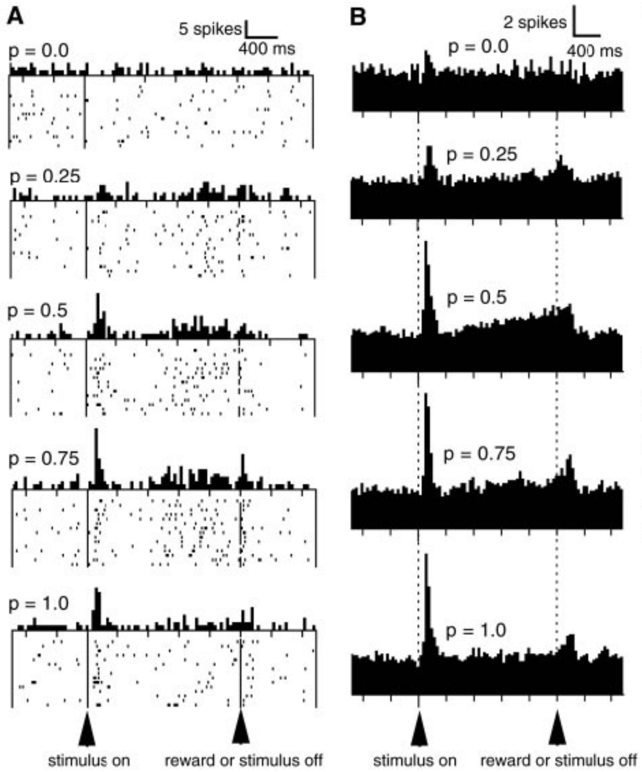


Figure 1: Sustained activation of dopamine neurons with uncertainty taken from Fiorillo et al. (2003) (A) Rasters and histograms of single cell activity (B) Population histograms

With uncertainty, the sustained activation began on presentation of a CS and increased in strength until a reward was due, at which point the activation ceased (Figure 1B, where P = 0.25, 0.5 and 0.75). This activation was greatest when uncertainty of reward was at a maximum, i.e., when the reward was received on only 50% of occasions and probability (p) was 0.5. Sustained activation was also seen at lower values of uncertainty, when probability was 25% and 75%, but to a lesser extent. No sustained activation was seen when probability was certain at either zero or 1, suggesting that the sustained activation coded for uncertainty.

However, this view is controversial as Niv, Duff and Dayan, (2005) have suggested that the sustained activation, or 'ramping' effect in the delay period, is due to backpropagating TD prediction errors, and not to uncertainty. Specifically, they suggest that it is the asymmetric coding of those prediction errors that give rise to the effects seen in time, over consecutive CS presentations, due to a low baseline rate of activity in dopamine neurons. Firing rates of positive prediction errors typically rise to about 270% above baseline, while negative errors only fall to approximately 55% below baseline (Fiorillo et al. 2003). During uncertainty, these asymmetrical positive and negative errors, when summed, will not cancel each other out, as predicted by the TD algorithm, even after extensive training periods. The overall effect, as seen in Fiorillo et al., will be of (i) a positive response across trials at the expected time of reward, and (ii) a 'ramping' effect from presentation of the CS to the expected time of reward, described by Fiorillo and colleagues as sustained activation. The resulting effects arise as a result of averaging across multiple trials and are not a within trial phenomena.

Using TD, Niv and colleagues successfully modelled both effects identified by Fiorillo et al. (2003) during uncertainty. They also showed that the shape of the ramp depended on the learning rate, and that the difference in the steepness of the ramp between delay and trace conditioning could be accounted for by the low learning rates associated with trace conditioning, resulting in a smaller or even negligible ramp.

In reply to Niv et al., Fiorillo and colleagues defend their original claim that dopamine encodes uncertainty about reward (Fiorillo, Tobler & Schultz, 2005). Three of the five points raised are of particular interest to this study. Firstly, they give two examples as evidence of sustained activation within single trials, which is contrary to the postulations of Niv et al., and secondly, they suggest that activity in the last part of the delay period should reflect the activity of the preceding trial. Finally, they suggest that other ways of using TD to model dopamine as a TD error are more biologically plausible than backpropagating TD errors. It is important, therefore, to look at a range of models in order to understand the limitations of using the TD algorithm to model the role of dopamine.

In the present study a simulation of a 'rat' in a one-armed maze was used to investigate the claims of Fiorillo and colleagues, using an alternative TD model to Niv et al. The maze modelled was similar to that used by McClure et al. (2003) linking the ideas of reward prediction error and incentive salience, but contained an additional 'satiety' state and only allowed travel in one direction. The aim of this investigation was to use TD learning to model the following effects seen in dopamine neuron firing by Fiorillo and colleagues: (a) The phasic activation at the expected time of reward that increased as probability decreased; (b) the sustained increase in activity from the onset of the CS until the expected time of reward, during uncertainty, posited either as uncertainty, or as backpropagating TD prediction errors; and (c) the sustained activation increasing with increasing reward magnitude. In addition, in the discussion an attempt is made to address three of the points raised by Fiorillo et al. (2005) in response to Niv et al. (2005).

Method

Temporal Difference

The maze incorporated an 'actor-critic' architecture (McClure et al., 2003; Montague, Hyman & Cohen, 2004; Sutton & Barto, 1998), a form of reinforcement TD learning where an 'adaptive critic' computes a reward prediction error, which is used by the 'actor' to choose those actions that lead to reward.

The Critic The TD algorithm is designed to learn an estimate of a value function V^*, representing expected total future reward, from any state, s, (Equation 1), where t represents time and subsequent time steps t = 1, t = 2 etc; E is the expected value and r represents the value of the reward. γ is a discounting parameter between 0 and 1 and has the effect of reducing previous estimates of reward exponentially with time, so that a reward of yesterday is not worth as much as a reward of today. Equation 2 is Equation 1 in a recursive form that can be used in the learning process.

$$V^*(s_t) = E[r_t + \gamma\, r_{t+1} + \gamma^2\, r_{t+2} + \gamma^3\, r_{t+3} + \ldots] \quad \text{[Eqn 1]}$$

$$V^*(s_t) = E[r_t + \gamma V^*(s_{t+1})] \quad \text{[Eqn 2]}$$

TD prediction error is a measure of the inconsistency for estimates of value at successive time steps. The error, δ(t), is derived by rearranging Equation 2 into Equation 3, which is a measure of the relationship between two successive states and the current reward. This will give estimates, V, of the value function V^*. The dopamine prediction error signal, δ(t), takes into account the current reward, plus the next prediction multiplied by the discounting parameter γ, minus the current prediction. It is the error δ(t) that is equivalent to the dopamine reward prediction error, or learning signal, to create better estimates of future reward.

$$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t) \qquad \text{[Eqn 3]}$$

The Actor An extension to the TD model has been made to include a role for dopamine in biasing action selection using the same prediction error signal, δ(t), to teach a system to take the best actions, namely those that are followed by rewards (McClure et al., 2003; Montague et al., 1996). The way an action is selected is that the actor randomly chooses a possible action, and the anticipated δ(t) is calculated using Equation 3. The probability of taking this action is then calculated from this δ(t) value using the softmax function in Equation 4 (where *m* and *b* are parameters of the softmax curve), which calculates the probability of that action occurring from the anticipated δ(t) value. If no action is selected, time is increased by one step and another random action is considered.

$$P\text{ (of taking action)} = (1 + e^{-m(\delta(t) - b)})^{-1} \qquad \text{[Eqn 4]}$$

Actions are generated with a probability of selection based on the predicted values of their successor states, preferring those actions that give a high burst of dopamine, or TD error signal. There is a greater probability of remaining at the same state and not making a move when the error signal is low as all states become increasingly probable.

Learning takes place in the model according to Equation 5, where α is a learning rate parameter.

$$V(s_i) \leftarrow V(s_i) + \alpha\, \delta(t) \text{ [Eqn 5]}$$

The Maze

A computer simulation was constructed of a ‘rat’ learning to traverse a one-arm maze to receive a reward, using the TD algorithm with an ‘actor-critic’ architecture. Figure 2 shows a maze with positions modelled as five states, starting at State 0 (the CS) and progressing through intermediate states to receive a simulated reward in State 4 (the reward state). In order to model the breaks between maze runs in real rats, it was necessary to insert a ‘satiety’ state (State 5) into the maze, between the goal (State 4) and the start (State 0), where the transition between that state and State 0 remained at zero so that no learning could take place. This had the effect of resetting the value of start State 0 to zero, acting as a ‘resting’ state and ensuring that the ‘rat’ was always surprised when starting the maze. Without this additional state, the simulated rat learnt the value of the start state, and in effect, there was no CS. Intermediate states were added and removed, as required to make mazes of different lengths.

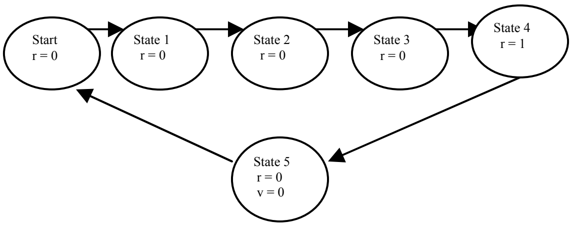


Figure 2: Maze with five states plus 'satiety' state

Simulations

Uncertainty – Degree of Probability The ranges of probabilities used for trials were 0.25, 0.5 (maximum uncertainty), or 0.75. The δ(t) values were recorded for each state transition, for a single probability in each trial. Each trial consisted of 1000 steps through a one-way maze with eight states plus a 'satiety' state, with a step being a transition from one state to the next, and a run being one complete journey through the maze, from start to finish. At the beginning of each trial the values of each state in the maze (V) were set to zero. Movement to the next state in the maze was selected according to the effect of TD learning on different probabilities of receiving a reward for each run.

In keeping with the biology of dopamine, namely the asymmetry in coding of positive and negative errors, any negative prediction errors were scaled by a factor of one sixth, the scaling factor used by Niv et al. (2005). The scaled δ(t) values were then averaged across fifty consecutive runs for each state, where $\gamma = 0.98$, and the magnitudes of the scaled values compared. This averaging corresponded to the summing of peri-stimulus-time-histograms (PSTH) of activity over different trials and inter-trial averaging used by Fiorillo et al. (2003).

Reward Magnitude Individual reward magnitudes of 0.5, 1 and 2 were compared in different trials to see the effect on the sustained activation.

An Example of Learning

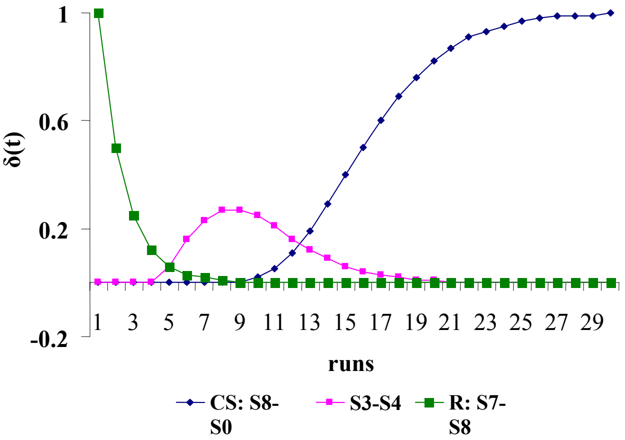


Figure 3: Delta values for each state transition over first thirty runs, p = 1, r = 1

With a probability of 1 and a maze of eight states plus a 'satiety' state, complete learning took place over the first thirty runs ($\gamma = 1$). On the first run a large prediction error, δ(t), was recorded at the expected time of the reward (S7-S8), and as runs progressed, this δ(t) was transferred back to the CS (S8-S0). When full learning had taken place only the CS elicited a reward prediction error. This effect is demonstrated in Figure 3, which shows the δ(t) at the expected time of reward beginning at 1 and reducing to zero by run 9 at which point the value of the state is learnt and the reward fully predicted. The δ(t) at the CS begins at zero and increases gradually to 1, from run 10 to run 30. An intermediate state transition S3-S4 is included which records the δ(t) backpropagated from the reward state by run 5. The error increases until run 8 and then reduces to zero by run 21.

All the following tests with uncertainty were done post training.

Results

Uncertainty – Degree of Probability

Eventually, by chance, actions were selected in trials for the entire range of probabilities (p), 0.25, 0.5 or 0.75, and the 'rat' progressed along the maze towards the reward state receiving the reward (r) of that state, r = 1. On subsequent runs, learning occurred as the value of the reward was propagated backwards, updating earlier states using a proportion of the prediction error signal, δ(t).

The patterns of data obtained show that it is necessary for the history of previous runs to be taken into consideration when analysing reward prediction errors and not just the last trial. Accordingly, consecutive runs should be selected for averaging in order to preserve the backward chaining effect of the TD algorithm. The TD algorithm uses rewards obtained in the past to make predictions about future expected reward, affecting the values of all the states in the maze, which are continually being updated as the rat progresses along the maze. With uncertainty, the particular course a rat takes on a particular trial is novel in each trial, as it depends on the exact order of rewarded and non-rewarded runs, which are delivered randomly by the computer program. The δ(t) values are then propagated backwards, in order, from later states to earlier states, as time progresses.

As the probability of obtaining a reward increased, from 25% to 50% to 75%, so did the level of phasic activation at the CS (S8-S0) (Table 1, Figure 5), with average δ(t) values of 0.23, 0.57 and 0.70 respectively.

(a) The phasic activations at the expected time of reward Without scaling the δ(t) values recorded for each state transition to compensate for the biologically asymmetric coding of positive and negative prediction errors, no average positive phasic activation was seen at the expected time of reward (Figure 4 S7-S8). However, after scaling δ(t) values by a factor of one sixth and averaging δ(t) values over consecutive trials, positive phasic activation was seen at the expected time of reward (Figure 5).

When comparing the average scaled δ(t) values across trials with probabilities of 0.25, 0.5 and 0.75, similar averaged, scaled δ(t) values were recorded of 0.16, 0.16 and 0.14 respectively. However, if averages were taken over rewarded trials only, as suggested in Figure 2A in Fiorillo et al. (2003), δ(t) values would be positive at the expected time of reward as all negative values would be removed. In addition, there would be less non-rewarded runs to be removed at higher probabilities, resulting in the phasic activation varying monotonically with reward probability. There would also be less of an effect on the phasic activation seen at the CS as the effect of reward would take longer to reach that state. As suggested above, difficulties arise when the backpropagating chain of reward prediction errors is broken and runs are taken out of context of the trial history.

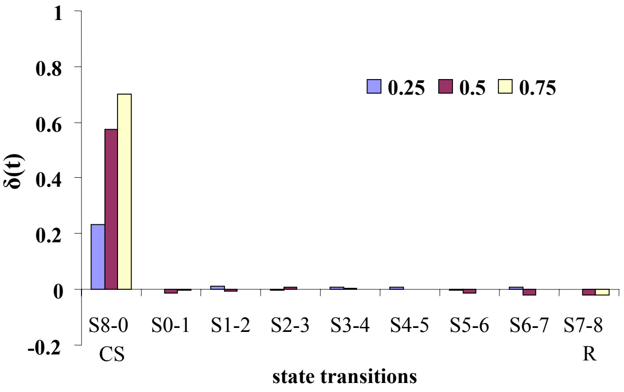


Figure 4: Average δ(t) values, **before scaling**, for each state transition over 50 runs, when p = 0.25, 0.5 and 0.75, α = 0.9

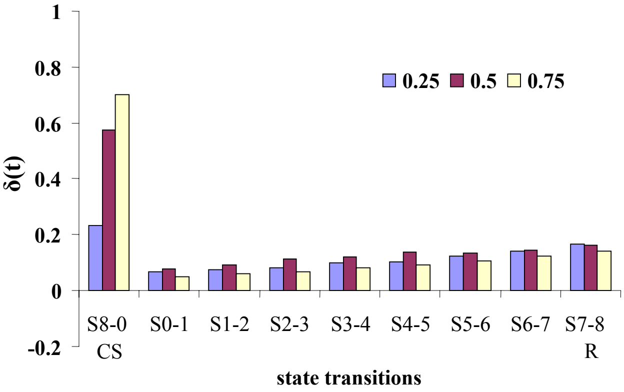


Figure 5: Average δ(t) values, **after scaling**, for each state transition over 50 runs, when p = 0.25, 0.5 and 0.75, α = 0.9

In conclusion, phasic activations were seen at the expected time of reward, in accordance with the findings of Fiorillo et al., when δ(t) values were scaled to compensate for the asymmetric coding. However, unless rewarded only trials were averaged, the phasic activation did not vary monotonically with reward probability.

(b) The sustained increase in activity Figure 4 shows that no 'ramping' effect is seen from plotting the average δ(t) values obtained for probabilities of 0.25, 0.5 and 0.75 for each state transition. Here the symmetrical positive and negative errors effectively cancel each other out, in accordance with the TD algorithm. However, when the δ(t) values were scaled by a factor of one sixth to compensate for the biological asymmetric coding of positive and negative errors, and averaged across consecutive runs, positive δ(t) values were seen that corresponded to the sustained activation and 'ramping' effects reported in Fiorillo et al. (2003) and Niv et al. (2005) respectively (Figure 5). The magnitude of the ramping effect is marginally greater for maximum probability, p = 0.5 than for the lower probabilities of p = 0.25

and p = 0.75, in accordance with the findings of Fiorillo et al. However, the difference seen between the two trials with probabilities of 0.25 and 0.75, which are comparable levels of uncertainty, could be accounted for by the different reward history for each. This difference should be negligible if more trials were taken into account.

(c) Reward Magnitude The value of the reward was manipulated across different trials, with rewards given of 0.5, 1 and 2. The size of the reward had an effect on the range of δ(t) values available for each state. With a larger reward comes a larger range of possible δ(t) values, and, accordingly, larger 'ramping' effects (Figure 6). Therefore, the sustained activation increased with increasing reward magnitude, in accordance with Fiorillo et al. (2003).

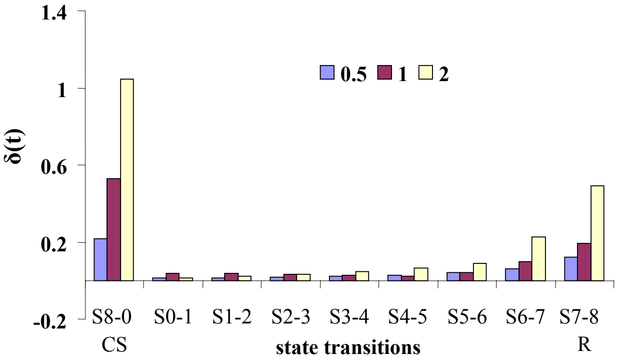


Figure 6: Scaled average δ(t) values over 30 runs for reward values of 0.5, 1 and 2, p = 0.5, α = 0.5

Discussion

A simulation of reinforcement learning, incorporating an 'actor-critic' architecture of TD learning, successfully modelled the following properties of dopamine demonstrated by Fiorillo et al. (2003): (a) The phasic activations at the expected time of reward; (b) the sustained increase in activity from the onset of the conditioned stimulus until the expected time of reward, during uncertainty; and (c) the sustained activation increasing with increasing reward magnitude. This supports the argument by Niv et al. (2005) that the ramping effect seen during uncertainty is a result of backpropagating TD errors and not a within-trial encoding of uncertainty.

In response to the claims of Niv and colleagues, Fiorillo et al. (2005) raised several points in support of their original argument, three of which are relevant to this study. Firstly, they refer to the difficulty of determining whether or not activity increases on single trials as Niv et al. (2005) did not specify what a single trial increase in delay-period activity should look like. In our simulations, Figure 7 is an example of a single trial (or a single run in this simulation), and is represented by recording the scaled prediction errors, δ(t), for each state transition, for one run through the maze. This run is analogous to the activity of a single neuron in a single trial over time and is simply a snapshot of the δ(t) values for each state, which may be either positive or negative with respect to baseline firing, depending on the history of previous runs.

The single run in Figure 7 is taken from actual run 6 in Figure 8 and represents non-rewarded run N preceded by ….RNRNR, where R is a rewarded run. The preceding RNR can be clearly identified in the δ(t) values seen for state transitions S4-S5, S5-S6 and S6-S7 respectively, but the results of earlier runs are harder to make out further back in time, as the TD algorithm ensures rewards or non-rewards in the past are not worth as much as those of the present.

Examination of many single runs through the maze did not reveal a ramping effect. Fiorillo et al. (2005) provided two examples of possible sustained activation in single trials, but these effects could have occurred quite by chance due to the order of rewarded and non-rewarded trials, as explained above, and not necessarily be examples of uncertainty. Indeed, if this within trial ramping effect were a regular occurrence then there would be many examples of single trials in support of the uncertainty hypothesis.

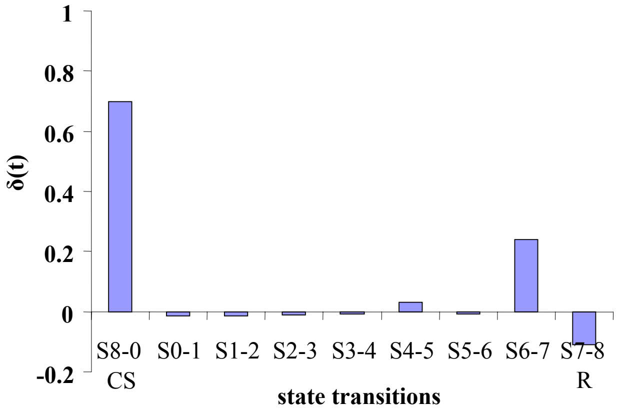


Figure 7: An example of a single trial:
Scaled δ(t) values for a single run, p = 0.5, r = 1

Secondly, Fiorillo and colleagues claimed that if activity during the delay period is due to backpropagating error signals that originated on previous trials, then the activity in the last part of the delay period of each individual trial should reflect the last reward outcome. Specifically, they suggest that if the preceding trial was rewarded, there should be more activity at the end of the delay period, and less activity if it was not rewarded, but they found no dependence of neural activity on the outcome of preceding trials.

Our results show that it is necessary for more of the history of previous runs to be taken into consideration than just the last reward outcome, when analysing reward prediction errors. For example, Figure 8 shows a history of rewarded and non-rewarded runs RNRNRNNNNN. After scaling, large δ(t) values were seen for runs 1-6 because alternate rewards and non-rewards were given, but runs 7-10 were not rewarded and, consequently, gradual extinction of the negative prediction error occurred. This example shows that it is not always the case that less activity will be seen if a trial is not rewarded (and vice versa), as runs 8-10 show an increase in firing (towards baseline) following non-rewarded runs.

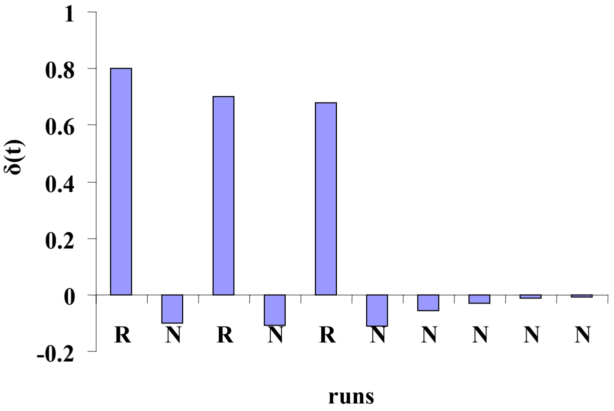


Figure 8: Scaled average δ(t) values at expected time of reward (S7-S8) recorded over 10 runs, p = 0.5, r = 1

Finally, Fiorillo et al. (2005) raise the argument that other TD models of dopamine are more biologically plausible than backpropagating TD errors, for example Suri & Schultz (1999), and it is important, therefore, to look at a range of models in order to understand the limitations of using the TD algorithm to model the role of dopamine. However, our work has shown that the predictions of Niv et al. (2005) are robust in the sense that they transfer to another type of model, albeit still using the same TD algorithm.

Conclusion

This alternate TD model to Niv et al. (2005) has effectively simulated conditioning in a Reinforcement Learning paradigm and successfully modelled the effects demonstrated in single dopamine neuron recordings, suggested to be coding for uncertainty, by Fiorillo et al. (2003). In addition, we have demonstrated what a single trial in TD Learning might look like and provide further evidence that ramping of the reward prediction error, δ(t), is not normally found within a trial of a single dopamine firing, but instead arises from averaging across trials.

Our simulations add further weight to the criticisms of Niv et al. that the effects demonstrated by Fiorillo and colleagues are due to backpropagating TD errors, and not a within-trial encoding of uncertainty. We support the claims by Niv et al. (2005) that the ramping signal is the best evidence yet for the nature of the learning mechanism of a shift in dopamine activity from expected time of reward to the CS.

References

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*, 1989-1902.

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behavioral and brain Functions*, *1:7.*

Hollerman, J.R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1,* 304-309.

McClure, S.M., Daw, N.D., & Montague, P.R. (2003). A computational substrate for incentive salience. *Trends in Neuroscience 26(8),* 423-428.

Montague, P.R., Dayan, P., & Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian Learning. *Journal of neuroscience, 16(5),* 1936-1947.

Montague, P.R., Hyman, S.E., & Cohen, J.D. (2004). Computational roles for dopamine in behavioral control. *Nature 431,* 760-767.

Niv, Y., Duff , M.O., & Dayan, P. (2005). Dopamine, uncertainty and TD Learning. *Behavioral and brain Functions, 1:6* 1-9.

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

Schultz, W., Dayan, P., Montague, P.R. (1997). A Neural substrate of prediction and reward. *Science*, *275:* 5306, 1593-1599.

Suri, R.E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91(3)*, 871-890.

Sutton, R.S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning 3,* 9-44.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement learning,* MIT Press.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*, 43-48.

Appendix V

Lisp Code for an Analysis of the Relationship between Temporal Difference Learning and Uncertainty Coding in a Computational Model of Dopaminergic Signaling (Chapter 6)

```
;set discount factor at 1
(defvar *gamma* 1)

;this function generates the possible moves that the agent may take
(defun possible-move (state-transitions current-state)
  (let ((possible-next-states (nth current-state state-transitions)))
    (nth (random (length possible-next-states)) possible-next-states)))

; updates the values of a state following a move (unless state 8, the reset state) and ;outputs current state, possible
next state, delta and values of  states
 (defun move-and-update (current-state state-transitions rewards Vs gamma bias m alpha reward-probability)
  (do* ((possible-next-state (possible-move state-transitions current-state)(possible-move state-transitions
                    current-state))
      (reward (get-reward current-state rewards reward-probability))
      (delta (+ reward (* gamma (nth possible-next-state Vs))(* -1 (nth  current-state Vs)))
           (+ reward (* gamma (nth possible-next-state Vs)) (* -1 (nth  current-state Vs))))
      (prob (sigma delta m bias)))
      ((> prob (random 1.0))
      (format t "~%~d ~d ~,2f ~{ ~,2f~}" current-state possible-next-state delta Vs)
      (unless (=  current-state 8) (setf (nth current-state Vs)(+ (nth current-state Vs)(* alpha (- delta
                    bias)))))possible-next-state)

;this function initiates a selected number of state transitions through the maze
 (defun go-all (state state-transitions rewards Vs gamma bias m alpha reward-probability number-of-steps)
  (dotimes (i number-of-steps)
    (setf state (move-and-update state state-transitions rewards Vs gamma bias m alpha reward-probability))))

;defines sigmoid function
 (defun sigma (delta m bias)
  (/ 1 (+ 1 (exp (* -1 m  (- delta bias))))))

;outputs whether run is rewarded or not
(defun get-reward (state rewards reward-probability &aux result)
  (setf result (if (< (random 1.0) reward-probability) (nth state rewards) 0))
  (when (= (nth state rewards) 1) (format t " ~d" result)) result)

;initializes parameters
 (progn
  (setf *McClure-Maze* '((1) (2) (3) (4) (5) (6) (7) (8) (0)))
  (setf *McClure-Vs* '( 0 0 0 0 0 0 0 0 0 ))
  (setf *McClure-rewards* '( 0 0 0 0 0 0 0 1 0 ))
```

Appendix VI

A Connectionist Model of Dual System Control (Chapter 8)

Section A
Pilot Study: Using Slow Weights Only

The slow weights in this ANN are typical of the weights seen in standard backpropagation ANNs and are thus capable of learning the data sets presented during Phases 1 and 2 on their own. What makes this ANN different from other ANNs capable of learning this task (e.g. Atkins 2001; Atkins & Murre 1998; Hinton & Sejnowski 1986; Plaut 1996; Hinton & Shallice 1991) is the addition of the fast weights. In order to ascertain the contribution of the fast weights, it was necessary first to train the network using only one set of weights to find the optimum parameters of the model, before introducing the fast weights. This was achieved by setting the learning rate and momentum of the fast weights to zero; although it was necessary to set the decay rate for the fast weights to a value of 0.999 in order that any random starting weights resulting from initialisation of the network should quickly decay.

To find the optimum number of hidden layer units
The hidden layer in Hinton and Plaut contained 100 units, one for each of the 100 associations in the Phase 1 data set, and training was easily achieved with only 1300 sweeps through the data set. As already mentioned, in accordance with current standard practice when investigating the performance of an ANN, I sought the lowest number of hidden layer units that would train the 20 associations in my Phase 1 data set. A smaller network will take longer to learn the associations to a minimal error, but will have a greater ability to generalise. This was achieved by starting with a localist network with 20 hidden units, one for each of the 20 associations like Hinton and Plaut, and then by training on fewer and fewer hidden units until I found the least number of units that would train to the stopping criterion where all 200 units were correct to within an error of 0.1.

Using ten different networks with ten different initialisation points, the minimum number of hidden layer units necessary for training to occur to the strict stopping criterion was found to be 7, using a low learning rate of 0.005 and a momentum of 0.9. The best of the ten networks trained in a minimum of 22,197 sweeps through the data set, giving rise to a total error of 0.013 across all of the output units and associations for the slow weights.

To find the optimum values of learning rates for fast and slow weights
With the optimum number of 7 hidden units established, it was necessary to seek optimum learning rate values for each of the fast and slow weights. This would be achieved by finding the maximum and minimum learning rates that would train the 20 associations in Phase 1, with 7 hidden layer units, using just the slow weights.

This was harder to achieve, but as I was interested in the interaction between fast and slow weights and strict convergence was not an important aspect of this investigation, I decided to relax the stopping criterion slightly for all the other experiments by accepting 199 of the possible 200 units being correct to within an error of 0.1. This allowed for one of the units to be incorrect.

Having looked at a range of learning rates from various different initialisation points of the network I found an optimum range of 0.001 to 0.41. These were the two rates I would be using to investigate the contribution of the fast weights in the remaining experiments.

Stopping Criteria

Hinton and Plaut did not give details of their training criteria only that they trained on the associations until perfect learning had occurred. In the pilot study I initially decided on a strict stopping criterion for training, where 200 of the possible 200 units (10 output units multiplied by 20 associations) for the Phase 1 associations were each correct to within an *error* of 0.1. (NB *error* is the root means square error of difference between the actual response and the target response for each neuron, hereafter referred to as error). However, as I was interested in the interaction between fast and slow weights and strict convergence was not an important aspect of this investigation, I decided to relax the stopping criterion slightly for all the other experiments by accepting 199 of the possible 200 units for the Phase 1 associations being correct to within an error of 0.1. This allowed for one of the units to be incorrect. Accordingly, the stopping criteria for Phases 2 and 3 with different numbers of associations would be 49 out of 50 units correct (10 output units multiplied by 5 associations) and 99 out of 100 units correct (10 output units multiplied by 10 associations), respectively.

In all simulations the network was trained ten times, using ten different initialisation points for starting weights in order to provide average results and to show generalisation of the network. However, some of the figures in the following sections are for one simulation only in order to demonstrate the typical interaction between fast and slow weights seen across all simulations. The simulation chosen was from an initialisation point that converged easily on a solution and was typical of a number of other simulations from different initialisation points that produced similar results.

Section B
Using Fast Weights to Temporarily Capture Old Learning

The aim of this experiment was to investigate the findings of Hinton and Plaut (1987) using a constrained version of the original model, to demonstrate that an additional set of fast weights is able to temporarily cancel out the interference in a set of old associations caused in more recent learning by rehearsing on just a subset of them.

METHODS

With the optimum network configuration of 7 hidden units and learning rate values of 0.001 and 0.41 for the slow and fast weights, respectively, training and testing were carried out in the following three phases:

Phase 1: The network was trained on 20 associations to a stopping criterion where only one of the 200 units (10 output units multiplied by 20 associations) were allowed to show an error of over 0.1. The weights were then frozen before Phase 2 commenced.

Phase 2: Training resumed on five new random associations presented to the network without rehearsing on the original 200 until the same stopping criterion had been met as in Phase 1. The weights were frozen, once again, before Phase 3 commenced.

Phase 3 (testing phase): Training resumed on a subset of the original data, 10 out of the 20 original associations from Phase 1, and the improvement in performance of the retrained 10 was compared to a test made on the unretrained 10 associations from Phase 1.

RESULTS

Results are shown for a typical simulation from one of ten different networks to show both the contribution of the fast weights and the interaction between fast and slow weights. It should be noted that while the results are comparable to Hinton and Plaut, they are not intended to be a direct replication of the original study and result from a constrained network using a data set with a greater amount of perturbation to the old associations during Phase 2.

Phase 1

The network learned the 20 associations to the stopping criterion in 83,001 sweeps, giving rise to a total error of 0.016 across all the output nodes for all of the associations. It should be noted that the number of sweeps in this constrained simulation are much higher than those seen during Phase 1 of the original study, reflecting the higher ratio of hidden layer units to associations, and hence the more distributed representation of the knowledge in this model rather than the more localist representation of Hinton and Plaut.

Plotting the total error across all output units and patterns against number of sweeps through the Phase 1 data set for (i) the fast weights, (ii) the slow weights and (iii) the total error of the system, shows the contribution of each of the fast and slow weights during learning. Figure B.1 shows the errors for the first 60,000 sweeps from which it can be seen that learning in the early stages is rapid in the fast weights, reflected as a dramatic decrease in the fast weights error during the first 5,000 sweeps (green line). Learning is much slower initially in the slow weights (pink line), but as the overall error (dark blue line) declines with learning and the fast weights decay, knowledge is transferred to the slow weights, which is reflected in the error for the slow weights that begins to decline more rapidly than in the initial 5,000 sweeps. By around 20,000 sweeps through the data set the slow weights begin to dominate, as the error falls below that of the fast weights. Changes to the slow weights are slow and steady as a result of the low learning rate, while changes to the fast weights are larger and more volatile as a result of the higher learning rate. The volatility of the fast weights is reflected in the overall error.

It is important to note at this stage for interpreting Figure B.1 and all further figures in this appendix that the figures show the root means squared error (abbreviated to error on the graph) for fast weights (fast wts error: green line), slow weights (slow wts error: pink line) and both sets of weights (overall error: blue line) of the system, over successive sweeps through the data sets. The overall error of the system should not be confused with the total weights of the system, where fast and slow weights sum together to give the total weights. Changes to fast and slow weights are derived from the same error in the system, but each will give different individual errors due to the differences in the respective learning rates, which will not sum to give the total error. The fast and slow weights are both free variables that can operate together to give many different solutions to a problem, unlike using slow weights only with a single solution.

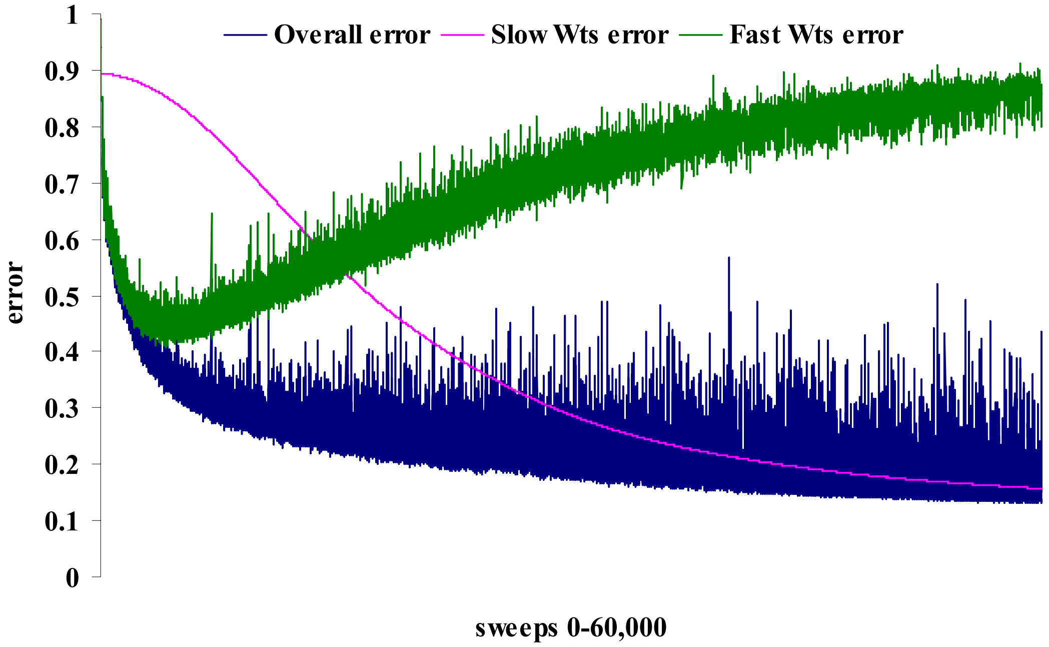


Figure B.1 During Phase 1 learning takes place initially in the fast weights (green line), but as the error declines the slow weights begin to dominate (pink line). Changes to the slow weights are slow and steady as a result of the low learning rate, while changes to the fast weights are larger and more volatile as a result of the higher learning rate. The volatility of the fast weights is reflected in the total weights error.

Phase 2

By continuing training on a set of 5 previously unseen associations only, the stopping criterion is reached in a further 21,001 sweeps (Figure B.2), giving rise to a total error of 0.028 across all the output nodes for all of the associations for the slow weights. When the new associations are initially presented the error in the system is high and the fast weights (green line) dominate during this period for the first 200 or so sweeps through the training set. As the error in the system reduces and the fast weights decay, knowledge of the new training set begins to be transferred to the slow weights (pink line) in a similar manner to that seen during Phase 1. There is a good interaction between fast and slow weights as the slow weights begin to take over from the fast weights by around 16,000 sweeps. However, it should be noted that, although the stopping criterion has been met by 20,000 sweeps, there has been insufficient time for the knowledge to be completely transferred to the slow weights and the fast weights continue to contribute to the knowledge in the entire network. I conclude from this that the decay for the fast weights, arbitrarily set at 0.999, is too slow for the less complex task in Phase 2. This aspect of the decay rate is investigated further in Section C.

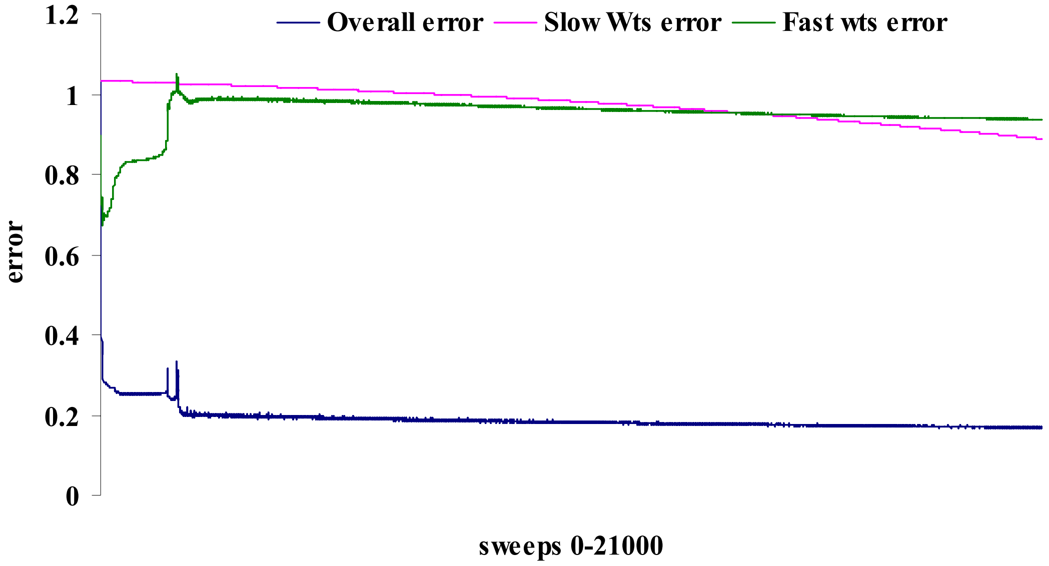


Figure B.2 During Phase 2 the fast weights (green line) become dominant for the first 200 or so sweeps through the new data set. However, as the error in the system declines with training and the fast weights decay, knowledge begins to be transferred from the fast weights to the slow weights (pink line). NB this is a sub-optimal simulation as the decay rate is too slow for the complexity of this task and there has been insufficient time for the knowledge to be transferred to the slow weights.

Phase 3

Training resumed on only 10 out of 20 of the original associations from Phase 1 and none of the associations from Phase 2. It took 446 sweeps through the Phase 3 data set to reach the stopping criterion producing a minimum error of 0.082 across all the output nodes for all of the associations for the slow weights. Figure B.3 shows the first 200 sweeps through the Phase 3 data set, where, once again, the error in the system is high as a result of reintroducing the previously learned associations, and the fast weights dominate during the initial stages. As in Phase 2 it would appear that the decay rate of 0.999 for the fast weights was too slow for the task complexity during Phase 3 as there was no interaction between the fast (green line) and slow (pink line) weights and insufficient time for the knowledge to be transferred to the slow weights (See Section C).

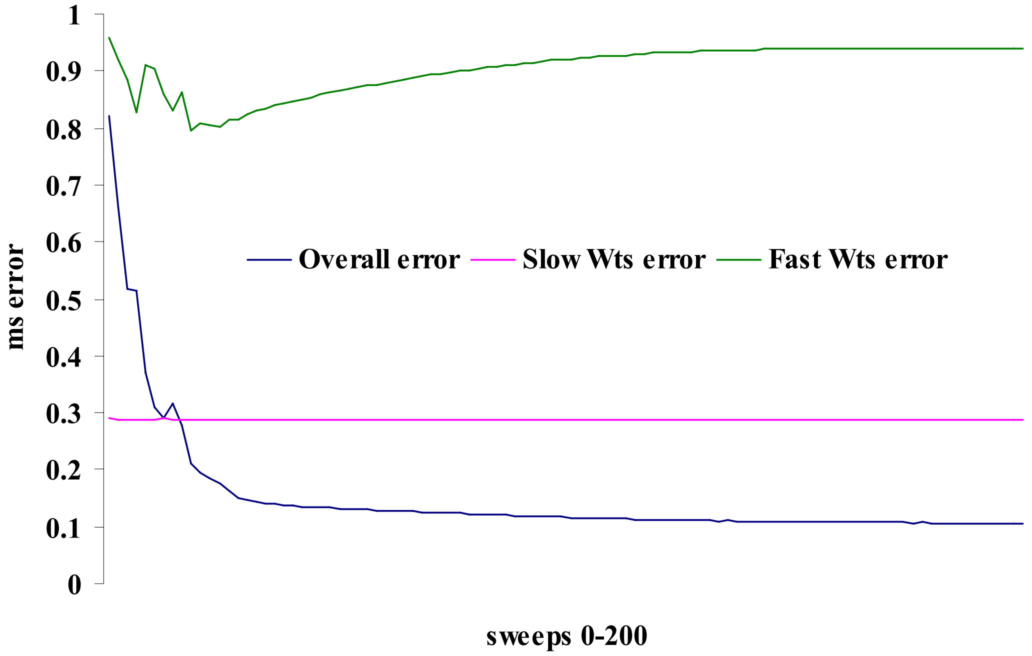


Figure B.3. During Phase 3 the fast weights dominate for the first 50 sweeps.

Phase 3 Testing

This was the testing phase of the experiment in Hinton and Plaut, where the point of interest was the first 20 sweeps through the Phase 3 data set and the effect of the initial training of the 10 retrained associations on the 10 unretrained associations. As well as keeping track of the retrained associations the model was also able to perform a test on the unretrained associations. However, this was not a direct replication of Hinton and Plaut as I used a constrained model with a reduced hidden layer. I also used a smaller data set in Phase 1 than Hinton and Plaut of 20 as opposed to 100 associations, while the Phase 2 data set contained the same 5 new random associations as the original study, resulting in a greater amount of perturbation, or disturbance to the Phase 1 associations than seen in Hinton and Plaut, i.e., 25% perturbation (5 to disrupt 20) in this study, compared to 5% perturbation (5 to disrupt 100) in the original study.

By plotting the total errors of both the retrained and the unretrained associations over the first 20 sweeps through the Phase 3 data set, it was possible to recreate a result comparable to that of Hinton and Plaut (1987). Figure B.4 is a typical example of one of the ten networks I used to capture this effect and demonstrates that the constrained model was working correctly. By using a constrained network with a greater amount of perturbation during Phase 2 there was less of an improvement on the 50% unretrained associations than was seen in the original study. However, my simulations have captured the effect identified by Hinton and Plaut and I have demonstrated that when the network was retrained on a subset of the original data it was found that in the early stages of retraining improvements were seen in the associations that were not retrained. This was because the knowledge of the original associations was distributed over many connections and retraining some of the associations pushed back the weights of the others to the point in time before the perturbation occurred. The fast weights were able to cancel out the interference in a set of old associations caused in more recent learning and it was possible to quickly restore a whole set of old associations by rehearsing on just a subset of them. The fast weights created a context in which the old associations were present again, without permanently interfering with the new associations, as the new knowledge was restored when the fast weights decayed back to zero.

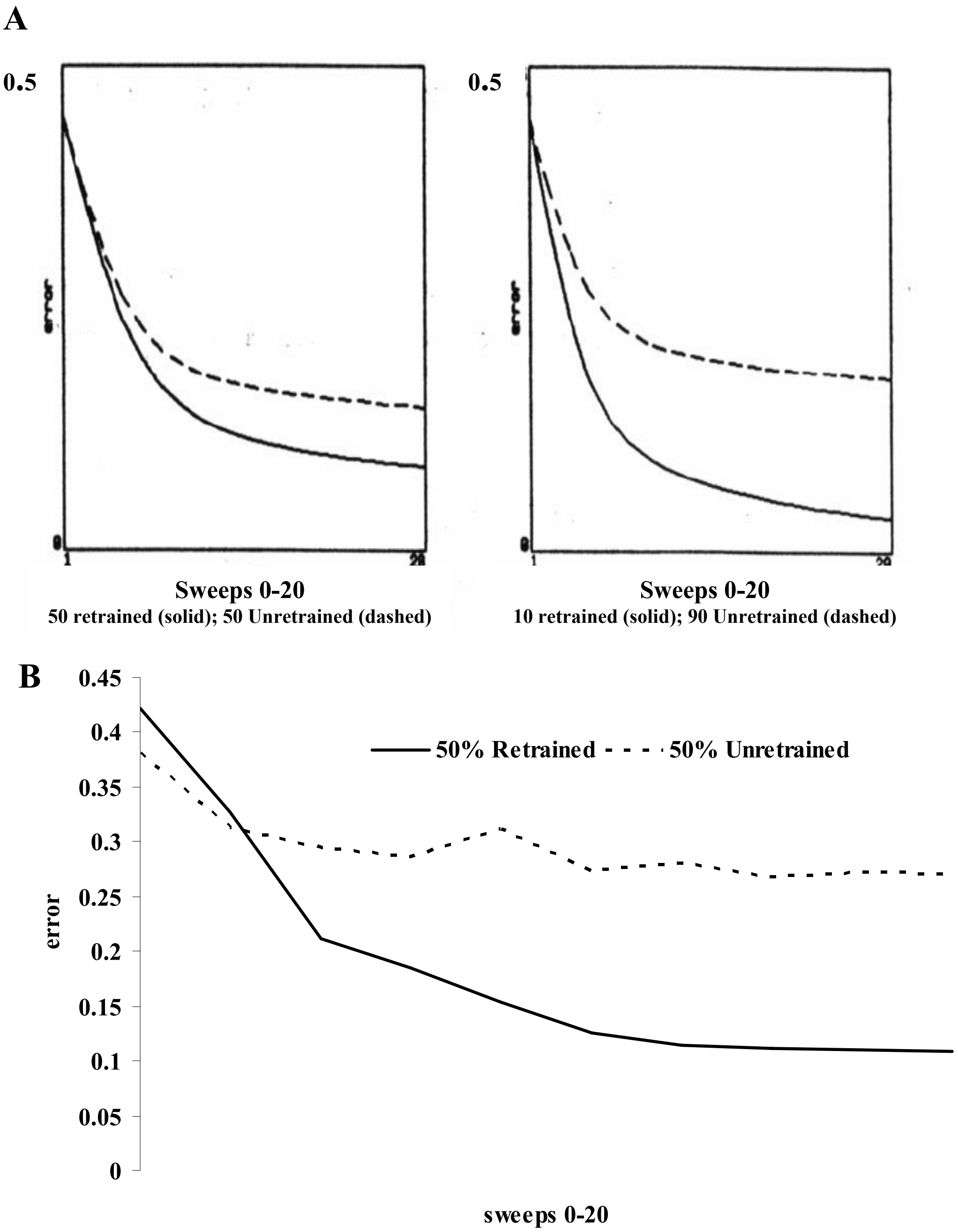


Figure B.4 Phase 3: A. Figure taken from Hinton & Plaut (1987) comparing the errors of both retrained and unretrained data for the first 20 sweeps. B. When the present network was retrained on 50% of the old associations (solid line) during Phase 3, it was found that in the early stages of retraining (first 20 sweeps) there was an improvement in the associations that were not retrained (dashed line).

CONCLUSIONS

Having established that my constrained model was performing correctly, by reproducing the effect identified by Hinton and Plaut, I was now in a position to investigate the parameters of the dual weights model.

The decay rate of 0.999 used in the above simulations, where the fast weights decayed at a rate of 0.1% with percentage retention of 99.9%, was suitable for the task complexity during Phase 1. However, it was clear that tasks with different learning complexities will require different decay rates for the fast weights in order for the knowledge of the task to be transferred to the slow weights by the time the stopping criterion has been reached. It would appear that the decay rate of the fast weights is a critical factor in the interaction between the fast and slow weights in this model, and before investigating this interaction in greater detail it will be necessary to investigate the contribution of the decay rate parameter for the fast weights and find the best possible combination of decay rates for the fast weights over Phases 1, 2 and 3.

The contributions of the decay rate parameter and the fast weights can be found in Sections C and D, respectively.

Section C
The Contribution of the Decay Rate Parameter

The results of the previous experiment revealed that a decay rate of 0.999 for the fast weights following every weight change was sufficient for the task complexity during Phase 1, as by the time the stopping criterion had been reached most of the knowledge of the system had been transferred to the slow weights. This provided a good interaction between the fast and slow weights during that phase. However, the task in Phase 2 was of a lower complexity and the stopping criterion was reached more quickly, requiring a lower number of sweeps through the data set. Therefore, it will be necessary to implement a higher rate of decay for the fast weights for Phase 2 in order for knowledge to be transferred to the slow weights by the time the stopping criterion is reached. A similar problem was seen for the task in Phase 3, although this did not present any difficulties for the current experiment as Phase 3 was a testing phase where the period of interest was the first twenty sweeps through the Phase 3 data set.

The aim of this experiment was to find the optimum combination of fast weight decay rates for the three phases for the network detailed in the previous experiment, and thus the best interactions between fast and slow weights, for each phase.

METHODS

In this experiment I looked at three rates of decay for the fast weights; the slowest being 0.999, where 99.9% of the knowledge in the system held across all the weights is retained in each sweep through the data set, and the fastest being 0.99, where only 99% of knowledge is retained. Table C.1 shows the percentage decay and percentage retention of the three different rates. These three rates were deemed sufficient for the purposes of this experiment, which was to demonstrate the effect of different rates of decay on the fast weights.

Table C.1 showing percentage decay and percentage retention for different rates of decay following every weight change.

Decay Rate	Percentage Decay	Percentage Retention
0.999	0.1	99.9
0.995	0.5	99.5
0.99	1	99

Using the optimal decay rate of 0.999 for Phase 1, already established, comparisons were made between fast weight decay rates of 0.99, 0.995 and 0.999 for Phases 2 and 3 in six simulations, according to the combinations in Table C.2 in the Results section, to find the optimum combination of decay rates for the fast weights for the three phases. The following criteria will give the best interactions between fast and slow weights:

Criteria for Phases 1 and 2 Learning:
1. An initial reduction should be seen in the error for the fast weights, which should decay before the stopping criterion has been met.

2. Initially the error for the slow weights should be high and this should gradually reduce over the course of the task to approach the total weights error by the time the stopping criterion has been reached.

3. An interaction should be seen between the errors of the fast and slow weights, where a crossover occurs as the knowledge is transferred from the fast to the slow weights.

Criterion for task 3:
Only part of Criterion 1 above is relevant in this scenario, where an initial reduction should be seen in the error for the fast weights. Criteria 2 and 3 above do not apply here as the aim of Phase 3 in Experiment 1 is to regain temporary access to old information in the fast weights and not to relearn the old information in the slow weights.

RESULTS

The numbers of sweeps taken for each phase in each of the six simulations are recorded in Table C.2. On average over a number of different trials it should be expected that the faster decay rate of 0.99 would require the highest number of sweeps to compensate for the weight decay, as more of the information in the system is forgotten than with the slower decay rates of 0.995 and 0.999. Similarly, the slowest decay rate of 0.999 should require the least amount of sweeps. While, this pattern was not always reflected in these individual results (e.g. Simulation 2, Phase 3), it should be mentioned that the program is a stochastic process that would only make such predictions on average and not in specific cases.

Table C.2 showing number of sweeps taken to fulfil relevant stopping criteria for 6 simulations using fast and slow weights, from the same initialisation point as detailed in the previous experiment. The optimum combination of decay rates for the fast weights is highlighted in red.

Sim. No.	**Phase 1** (20 associations)		**Phase 2** (5 associations)		**Phase 3** (10 associations)	
	Decay rate (% decay)	Sweeps	Decay rate (% decay)	Sweeps	Decay rate (% decay)	Sweeps
1	0.999 (0.1)	83,001	0.99 (1)	251,001	0.99 (1)	10,001
2	0.999 (0.1)	83,001	0.99 (1)	251,001	0.995 (0.5)	11,001
3	**0.999 (0.1)**	**83,001**	**0.99 (1)**	**251,001**	**0.999 (0.1)**	**1,001**
4	0.999 (0.1)	83,001	0.995 (0.5)	15,001	0.99 (1)	3,001
5	0.999 (0.1)	83,001	0.995 (0.5)	15,001	0.995 (0.5)	1,001
6	0.999 (0.1)	83,001	0.995 (0.5)	15,001	0.999 (0.1)	81

To Find the Optimum Fast Weight Decay Rate for Phase 2

The weights of Phase 1 were saved before training resumed on 5 previously unseen associations in Phase 2, without continuing training on the Phase 1 associations.

Simulations 1, 2 and 3: 0.999-0.99

The combination of fast weight decay of 0.999 for Phase 1 followed by 0.99 for Phase 2 satisfied the three criteria for Phase 2 learning, namely: (i) an initial reduction was seen in the error for the fast weights, which decayed before the stopping criteria had been met; (ii) initially the error for the slow weights was high and this gradually reduced over the course of the task to approach the total weights error by the time the stopping criterion had been reached; and (iii) an interaction was seen between the errors of the fast and slow weights, where a crossover occurred as the knowledge was transferred from the fast to the slow weights. Figure C.1 shows that all three criteria were satisfied for Phase 2 learning by the first 60,000 sweeps through the associations (although it took a further 191,000 for the stopping criterion to be reached), while the interaction between fast and slow weights had occurred by 5,000 sweeps.

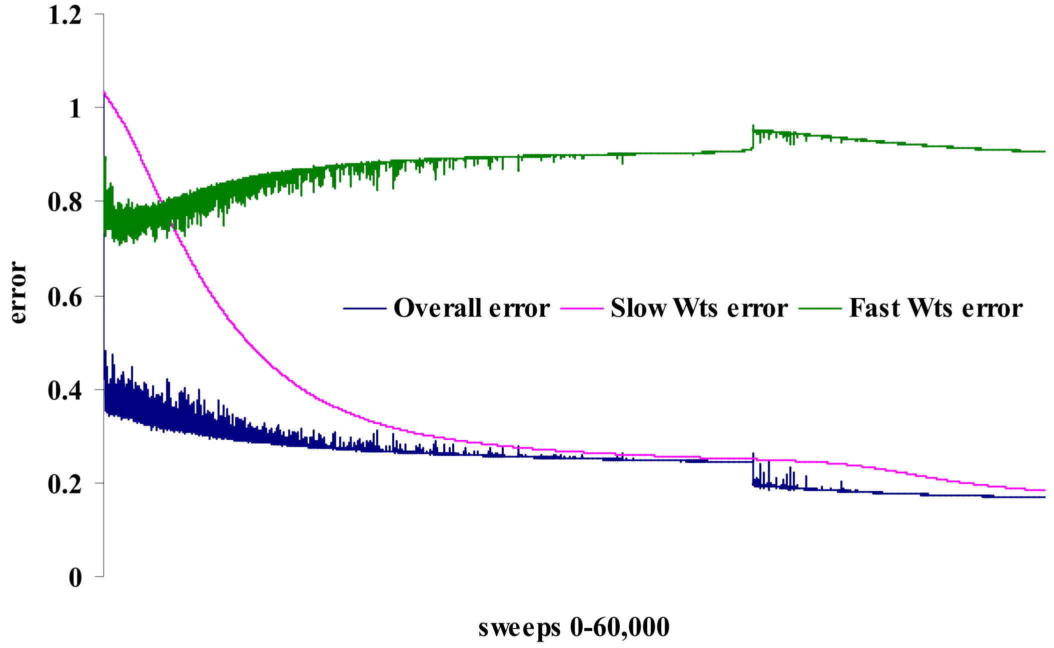


Figure C.1 Optimal combination of 0.999-0.99 for Phase 2 learning, fulfilling all three criteria. First 60, 000 of a total of 251,000 sweeps necessary to reach stopping criterion.

Simulations 4, 5 and 6: 0.999-0.995

Criteria 1 and 3 for a satisfactory fast weight decay rate were also met using a fast decay rate of 0.995 for Task 2. However, Criterion 2 was not fulfilled as full knowledge did not transfer sufficiently to the slow weights before the stopping criteria were met (Figure C.2). The effect was even less using a slower decay rate of 0.999 (Section B, Figure B.2).

It was decided, therefore, that the optimum decay rate for the fast weights for Phase 2 learning was 0.99 as it satisfied all three criteria for that phase and showed the best interaction between the fast and slow weights. Coupled with the optimum decay rate of 0.999 for Phase 1 learning, the next task was to find which of simulations 1, 2 or 3 had the optimum combination of decay rates for all three phases.

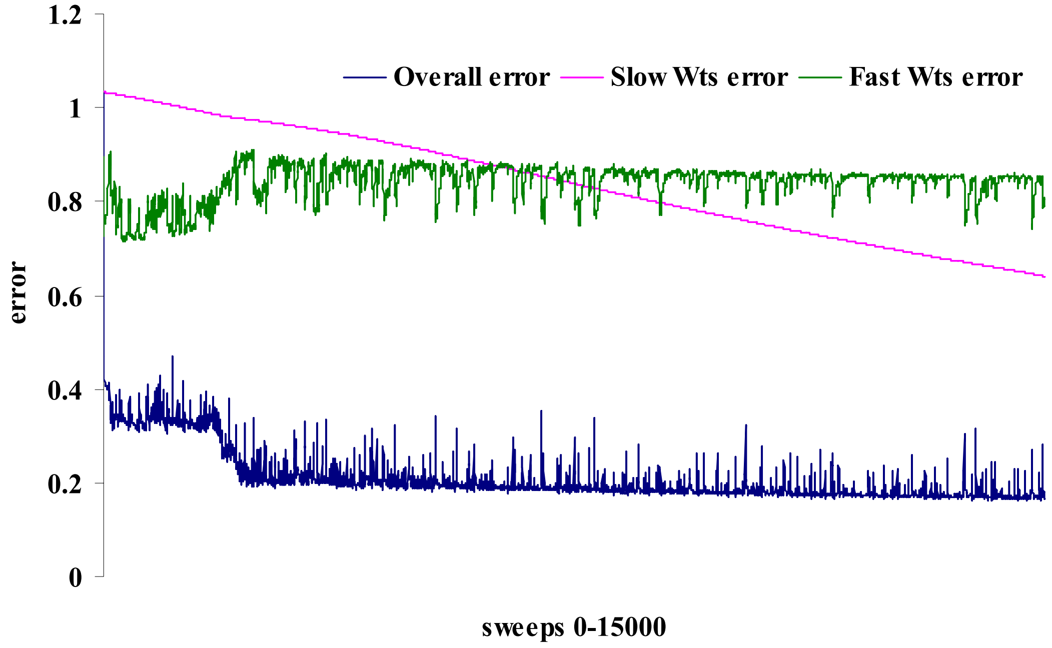


Figure C.2 Suboptimal combination of 0.999-0.995 for Phase 2 learning, fulfilling only two of the three criteria.

To Find Optimum Fast Weight Decay Rate for Phase 3

For Phase 3 the weights of Phase 2 were saved before training resumed on 10 of the 20 associations previously learned during Phase 1, without continuing training on the 5 Phase 2 associations. Using 0.999-0.99 as a base, being the optimal combination for Phases 1 and 2, simulations were continued for Phase 3 to ascertain which of Simulations 1, 2 or 3 gave the best performance. The only criterion need to be satisfied during this phase was for an initial reduction to be seen in the error for the fast weights, where the fast weights were being seen as providing a temporary context in which the old associations were present again without permanently interfering with the new associations. The other criteria for Phases 1 and 2 do not apply here as the aim of this phase is to regain temporary access to old information in the fast weights and not to relearn the old information in the slow weights. However, it should be mentioned that if Phase 3 training is allowed to continue until the stopping criterion has been reached, the associations from Phase 1 represented to the network during that phase will be gradually assimilated into the slow weights and disrupt the Phase 2 associations.

Simulation 3, with decay rates of 0.999-0.99-0.999, provided the best combination. It can be seen from Figure C.3 that the criterion for Phase 3 had been met as there was a dip in the error for the fast weights during the first 150 or so sweeps, where the fast weights dominated due to the reintroduction of 10 of the 20 original associations in Phase 1. With Simulations 1 and 2 (decay rates of 0.99 and 0.995 for the fast weights in Phase 3, respectively) the fast weights decayed too rapidly for the complexity of this task and did not dominate sufficiently during the initial stages of training to provide a temporary context with the change in associations (not shown).

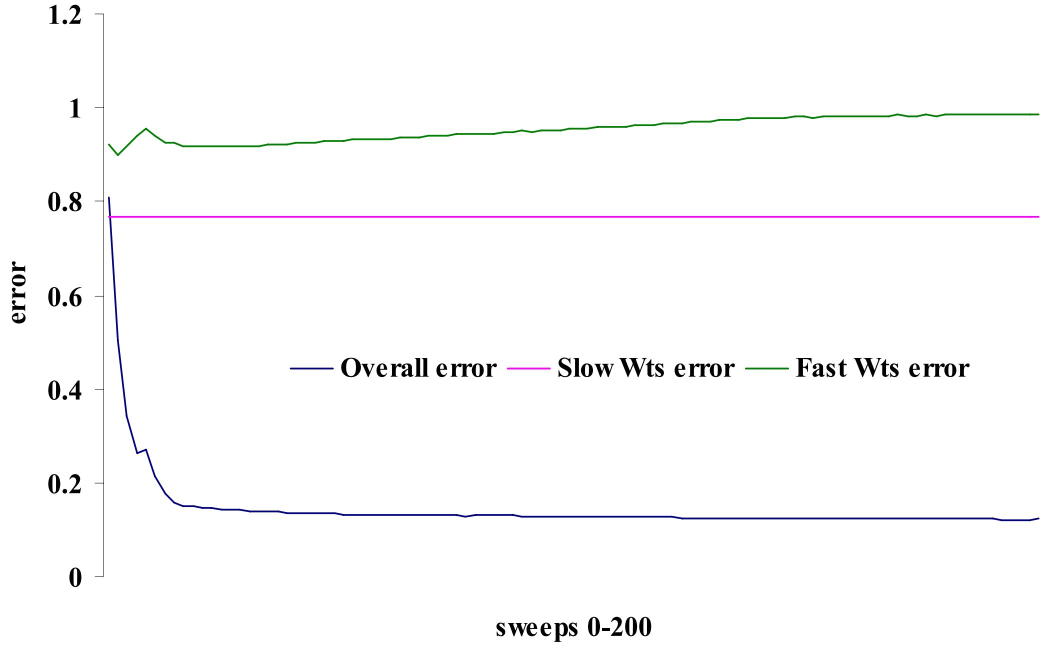


Figure C.3. Optimal combination of 0.999-0.99-0.999 for Phase 3 fulfilling the criterion of an initial reduction in the error for the fast weights which should decay before the stopping criteria of the task has been met.

CONCLUSIONS

The simulations in this experiment demonstrated that the interaction between the fast and slow weights is enabled by the decay rate for the fast weights where the weights rapidly decay towards zero by some fraction, *h*, after each weight change. It was clear that tasks with different learning complexities will require different decay rates for the fast weights in order for the knowledge of the task to be transferred to the slow weights by the time the stopping criterion has been reached. It would appear that the decay rate of the fast weights is a critical factor in the interaction between the fast and slow weights in this model: A fast decay rate (0.99) is appropriate for tasks with a lower complexity that train in a minimal number of sweeps where the contribution of the fast weights is minimal, such as in Phase 2 where the network has to learn only 5 new associations; and a slower decay rate (0.999) is more appropriate for tasks with a higher complexity that take longer to train and are facilitated by the fast weights, such as in Phases 1 and 3. Accordingly the optimal combination for decay rates over Phases 1, 2 and 3 is 0.999-0.99-0.999.

Section D
The Contribution of the Fast Weights

In order to investigate the contribution of the fast weights in the dual weights model I looked at the effect of training over the three phases using just the slow weights. Two control conditions were implemented, where progression was made though the three phases using slow weights only using learning rates of either 0.001 or 0.41, being the optimum learning rate values for slow and fast weights determined in the pilot study and used in the experiments detailed in the previous two sections. This involved setting the fast weights to zero, however, it was still necessary for a decay rate to be applied to the fast weights in order to allow for any random initialisation weights to decay to quickly to zero. The decay rates were set at the optimum combination of 0.999-0.99-0.999 and a momentum of 0.9 was applied to the slow weights.

The control conditions were compared to a third condition using fast and slow weights, being the results from Simulation 3 in the previous experiment, the optimal combination of decay rates for the three phases of 0.999-0.99-0.999. The three conditions in this experiment were all performed from the same initialisation point and so by comparing the results of these two control conditions using slow weights only to the results obtained from Simulation 3 of the previous experiment, using both fast and slow weights, it will be possible to ascertain the contribution of the fast weights over the three phases of the experiment.

RESULTS

As the three conditions in this experiment were all performed from the same initialisation point it was possible to ascertain the contribution of the fast weights over the three phases of the experiment by comparing two control conditions using slow weights only to a condition using fast and slow weights. However, as previously mentioned, it is important to note that the total error of the system should not be confused with the total weights of the system, where fast and slow weights sum together to give the total weights. The fast and slow weights are both free variables that can operate together to give many different solutions to a problem, unlike using slow weights only with a single solution. Therefore in Figures D.1 to D.3 the errors plotted for the fast and slow weights will not sum together to give the total weights for Simulation 3 and, likewise, they are not comparable directly to the errors for either of the controls. However, direct comparison can be made between the overall error for Simulation 3 and the errors for each of the control conditions, while the fast and slow weight errors give an indication of which of the two is contributing the most to the total weights at each stage of learning.

The numbers of sweeps taken to train each of the three conditions over the three phases of training are given in Table D.1 and the results are discussed in the next three sub-sections.

Table D.1 showing number of sweeps taken to reach the relevant stopping criteria for the three phases of training for three conditions: (i) Simulation 3, the optimum combination of decay rates of 0.999-0.99-0.999 ascertained from Section C, using both fast and slow weights with learning rates of 0.41 and 0.001, respectively; (2) Control 0.001 using slow weights only and a learning rate of 0.001; and (3) Control 0.41 using slow weights only with a learning rate of 0.41. All three conditions were preformed from the same initialisation point.

Condition	**Phase 1** (20 associations)		**Phase 2** (5 associations)		**Phase 3** (10 associations)	
	Decay rate	Number Sweeps	Decay rate	Number Sweeps	Decay rate	Number Sweeps
Simulation 3 LR Slow: 0.001 LR Fast: 0.41	0.999	83,001	0.99	251,001	0.999	1,001
Control 0.001 LR Slow: 0.001 LR Fast: Zero	0.999	227,001	0.99	43,001	0.999	10,001
Control 0.41 LR Slow 0.41 LRFast: Zero	0.999	2,001	0.99	41,001	0.999	766

Phase 1

As would be expected it takes longer to reach the stopping criterion using slow weights only and a learning rate of 0.001 (227,001) than a learning rate of 0.41 (2,001 sweeps). Whereas Simulation 3, using both fast and slow weights with the same learning rates takes an intermediate number of sweeps (83,001).

Figure D.1 shows the effect of adding the errors for the slow weights for the two control conditions during Phase 1 to Figure B.1 in Section B, for Phase 1 of Simulation 3 which contains the errors for both fast and slow weights for Phase 1. By comparing the overall error of a dual weighted network (dark blue line) to the two control conditions using slow weights only, it can be seen that Control 0.001 (red line) is slower to learn during the early stages, while Control 0.41 (light blue line) learns almost immediately. A closer inspection of the fast weights for Simulation 3 shows that they make a substantial contribution to the early stages of learning, and a network using fast and slow weights with learning rates of 0.41 and 0.001 learns more rapidly during the initial stages of training and takes fewer sweeps to reach the stopping criterion than a network consisting of slow weights only and a learning rate of 0.001.

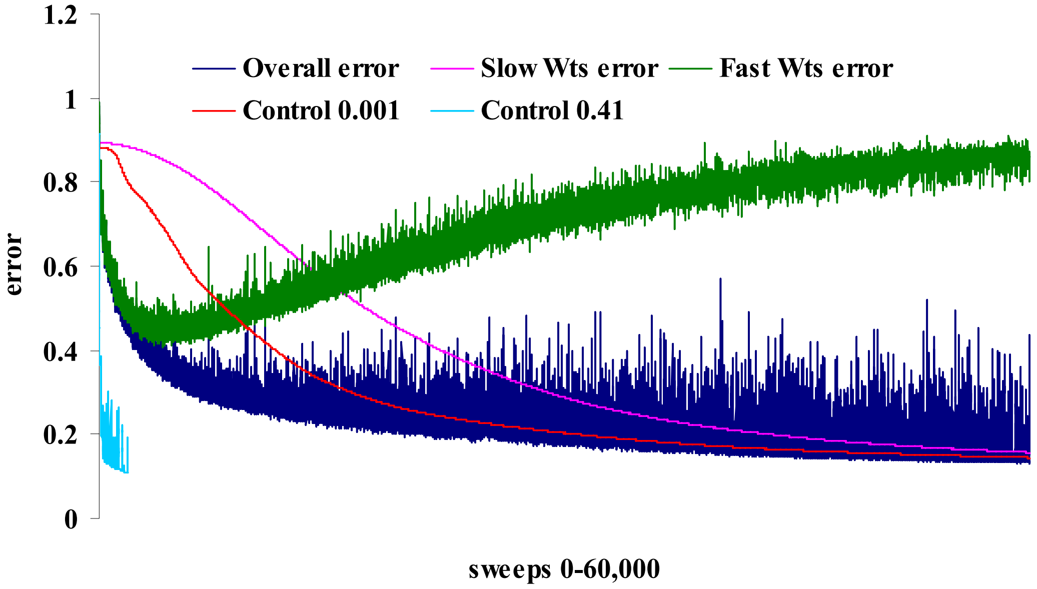


Figure D.1 Controls for Phase 1. The errors of the fast (green line), slow (pink line) and overall (dark blue line) for a dual weighted network (Simulation 3) compared to the total weights of two control conditions using slow weights only with learning rates of 0.001 (red line) and 0.41 (light blue line). The fast weights in Simulation 3 make a substantial contribution to the early stages of learning.

While the combination of fast and slow weights does not learn as fast as using a learning rate of 0.41 alone, such a high training rate often leads to error and will not always find a solution for the network, unlike a very slow learning rate, and so the combination of fast and slow weights offers an alternative to using an error-prone high learning rate only.

Control: Phase 2

Figure D.2A shows the effect of adding the errors for the two control conditions during Phase 2 to Figure C.1 in Section C above for Phase 2 of Simulation 3. Again, Control 0.001 with the lower learning rate (red line) is slower to learn initially than the dual weighted network (dark blue line), while Control 0.41 with the higher learning rate (light blue line) learns the fastest. However there is a short period of the first 30 sweeps where the overall error for the dual weighted network dominates over Control 0.41 with a high learning rate (Figure D.2B). As with Phase 1 a closer inspection of the errors for the fast weights for Simulation 3 shows that they make a substantial contribution to the early stages of learning and a network using fast and slow weights with learning rates of 0.41 and 0.001 learns more rapidly during the initial stages of training than a network consisting of slow weight only and a learning rate of 0.001.

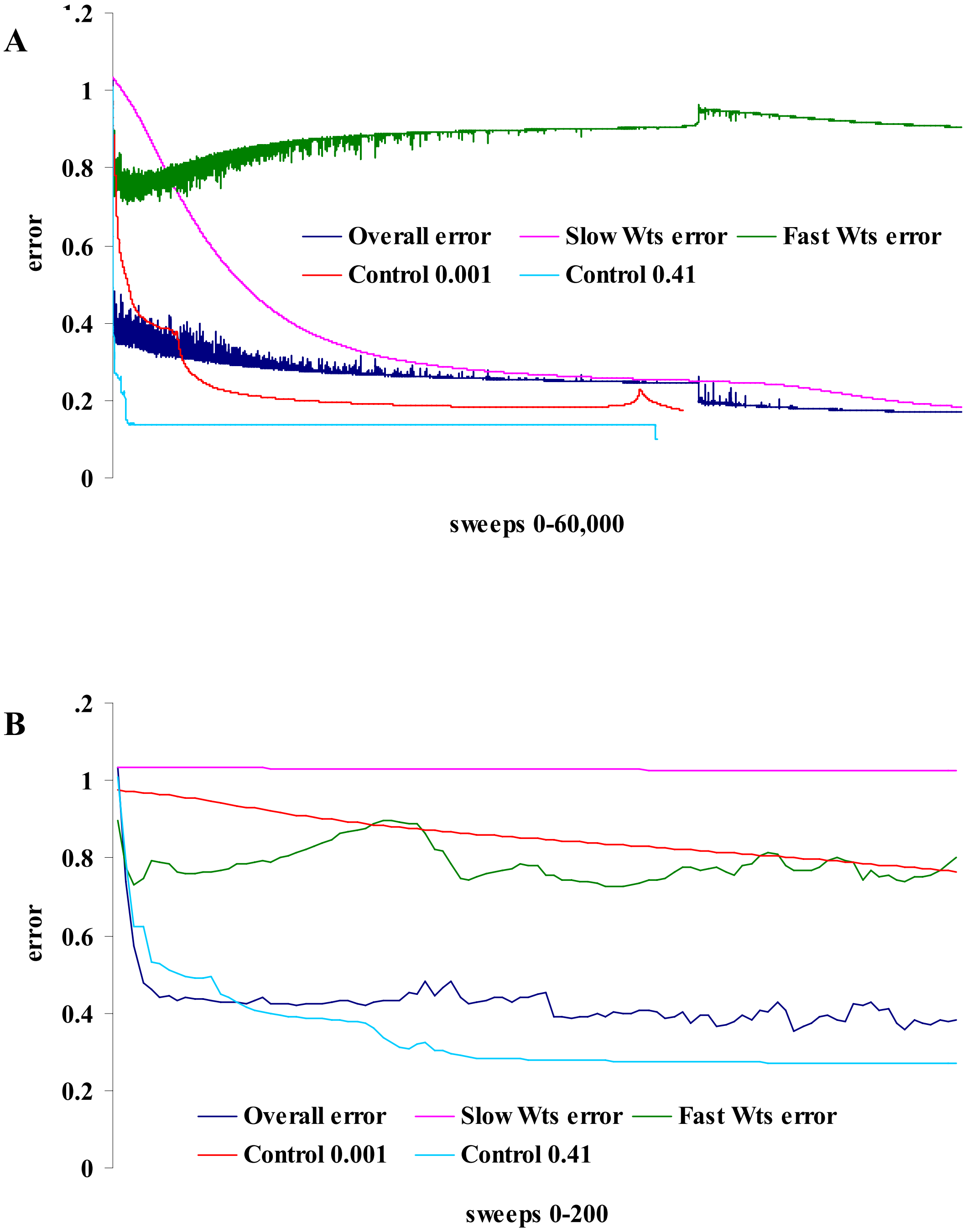


Figure D.2 Controls for Phase 2. The price to pay for early rapid learning in a dual weighted system is the time taken for the network to converge on a solution. (A) Errors for the first 60,000 sweeps. (B) The first 200 sweeps in greater detail show the temporary domination of the dual weighted system (dark blue line) during the first 30 sweeps.

However, the dual weighted network takes longer to reach the stopping criterion (251,000 sweeps) than the control conditions with learning rates of 0.0001 (43,001 sweeps) and 0.41 (41,000 sweeps). While the addition of the fast weights provides rapid learning during the initial stages of training (green line) and the overall error (dark blue line) for the dual weighted network is lower early in training than Control 0.001 (red line), it does so at the expense of the number of sweeps, or time taken to complete the task. A dual weighted network is trying to converge on a solution using two sets of weights, which is a much harder problem than a standard network with a single set of weights. Thus, the price to pay for early rapid learning is the time taken for the network to converge on a solution.

Control: Phase 3

Once again the fast weights (green line) provide rapid learning/relearning during the early stages (Figure D.3) and this is reflected in the overall error for the dual weighted network (dark blue line). What is interesting about this phase is the extent of the fall in the overall error for the dual weighted network. In the previous two phases the control condition with the high learning rate of 0.41 produced the most dramatic reductions in the error (Figures D.1 and D.2), although there was a short period during the first 30 sweeps of Phase 2 where the total weights for Simulation 3 dominated over Control 0.41 (Figure D.2B). The difference seen in Phase 3 may well be due to the relearning of the Phase 1 associations referred to by Hinton and Plaut, where retraining on some of those associations pushed back the weights to the point in time before the disturbance to the weight values occurred.

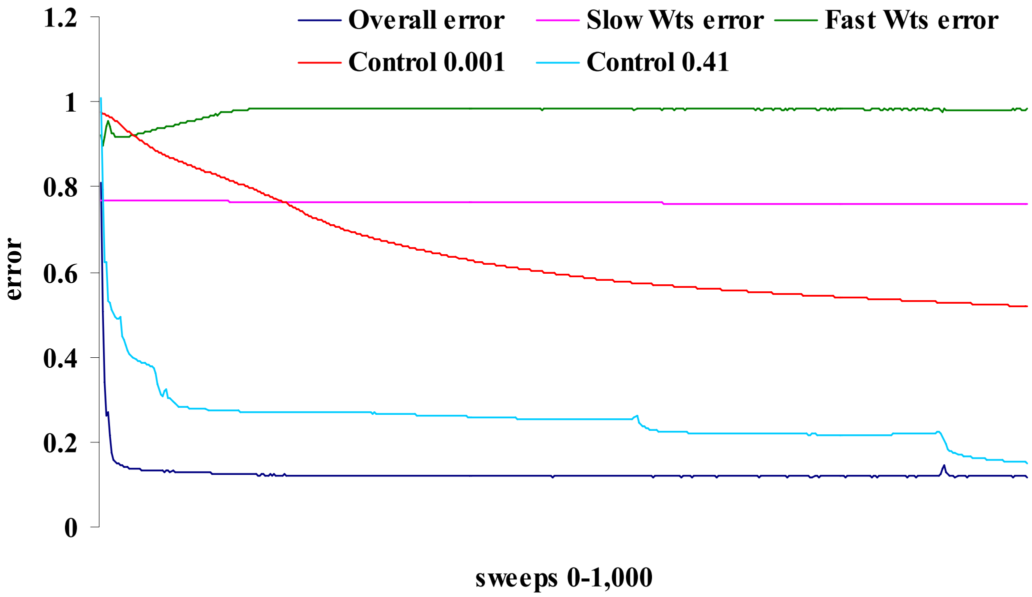


Figure D.3 Controls for Phase 3. The errors of the fast (green line), slow (pink line) and total weights (dark blue line) for a dual weighted network (Simulation 3) compared to the total weights of two control conditions using slow weights only with learning rates of 0.001 (red line) and 0.41 (light blue line).

CONCLUSIONS

The control conditions in this experiment have demonstrated the contribution of the fast weights in a dual weighted system. In all three phases the fast weights make a substantial contribution to the early stages of learning compared to a network using slow weights only and a low learning rate of 0.001. Of special interest is that additional fast weights improve learning/relearning compared to a high learning rate of 0.41 during Phase 3 (and to a limited extent during the first 30 sweeps of Phase 2). The difference seen in Phase 3 may well be due to the relearning of the Phase 1 associations referred to by Hinton and Plaut (1987), where retraining on some of those associations pushed back the weights to the point in time before the disturbance to the weight values occurred.

While the addition of the fast weights provides rapid learning during the initial stages of training, it does so at the expense of the number of sweeps, or time taken to complete the task. A dual weighted network is trying to converge on a solution using two sets of weights, which is a much harder problem than a standard network with a single set of weights. Thus, the price to pay for early rapid learning is the time taken for the network to converge on a solution.