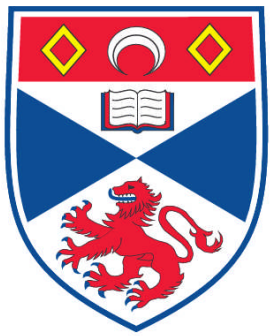
REFINEMENTS OF BIOLOGICALLY INSPIRED MODELS OF REINFORCEMENT LEARNING

Luca Aquili

A Thesis Submitted for the Degree of PhD
at the
University of St. Andrews



2010

Full metadata for this item is available in the St Andrews Digital Research Repository
at:
https://research-repository.st-andrews.ac.uk/

Please use this identifier to cite or link to this item:
http://hdl.handle.net/10023/886

This item is protected by original copyright

Refinement of biologically inspired models of reinforcement learning

A thesis for the degree of PhD
Submitted September 2009

by

Luca Aquili

University of St Andrews

School of Psychology
Faculty of Science

I Luca Aquili, hereby certify that this thesis, which is approximately 31000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September 2006 and as a candidate for the degree of PhD in September 2006; the higher study for which this is a record was carried out in the University of St Andrews between 2006 and 2009.
date signature of candidate

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.
Date signature of supervisor

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. We have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Access to Printed copy and electronic publication of thesis through the University of St Andrews.

date signature of candidate signature of supervisor

Acknowledgements

First of all, I would like to thank my supervisor Dr. Eric Bowman. Over a three year period, he has not only been an excellent teacher but also a marvellous person to work with. I would also like to thank all the people that I have worked with in the animal unit during the course of my thesis. These are, Dr. Iraklis Petrof (for his help with electrode making), Mary Latimer (for her help with histological procedures), Dr. David Wilson (for his help and advice on electrophysiology), and Dr. David Tait (for his help with stereotaxic procedures.

In addition, I would like to thank my family; my father Franco and my mother Birgitta for their continuous support all these years, Magda, and my good friends Conor, Dirck, Farid, Gerald and James.

Abstract

Reinforcement learning occurs when organisms adapt the propensities of given behaviours on the basis of associations with reward and punishment. Currently, reinforcement learning models have been validated in minimalist environments in which only 1-2 environmental stimuli are present as possible predictors of reward. The exception to this is two studies in which the responses of the dopamine system to configurations of multiple stimuli were investigated, however, in both cases the stimuli were presented simultaneously rather than in a sequence.

Therefore, we set out to understand how current models of reinforcement learning would respond under more complex conditions in which sequences of events are predictors of reward. In the two experimental chapters of this thesis, we attempted to understand whether midbrain dopaminergic neurons would respond to occasion setters (Chapter 3), and to the overexpectation effect (Chapter 4). In addition, we ran simulations of the behavioural paradigms using temporal difference models of reinforcement learning (Chapter 2) and compared the predictions of the model with the behavioural and neurophysiological data.

In Chapter 3, by performing single-neuron recording from VTA and SNpc dopaminergic cells, we demonstrated that our population of neurons were most responsive to the latest predictor of reward, the conditioned stimulus (CS) and not the earliest, the occasion setter (the OS). This is in stark contrast with the predictions of the model (Chapter 2), where the greatest response is seen at the OS onset. We also showed at a neural level that there was only a weak enhancement of the response to the discriminative stimulus (S_D) when this was preceded by the OS. On the other hand, at a behavioural level, bar pressing was greatest when the S_D was preceded by

the OS, demonstrating that rats could use the information provided by the OS, but that dopamine was not controlling the conditioned response.

In Chapter 4, our population of dopaminergic neurons showed that they would preferentially respond to only one of the two conditioned stimuli (CS_A, CS_B) in the overexpectation paradigm. The predictions of the model (Chapter 2) suggested that when the two stimuli would be presented in compound, there would be an inhibitory response if the reward magnitude was kept constant and an excitatory response if the reward magnitude was doubled. The lack of neural firing to one of the two conditioned stimuli, however, does not make for easy interpretation of the data.

Perhaps, one of the conditioned stimuli acted as if it were overshadowing the other, resulting in no response to the second CS. Interestingly, at a behavioural level, we did not see increased licking frequency to the compound stimuli presentation, a result that is somewhat at odds with the previous literature.

Overall, the results of our experimental chapters suggest that the role that midbrain dopaminergic neurons play in reinforcement learning is more complex than that envisaged by previous investigations.

1 Chapter 1. General Introduction

My PhD research considers the role that midbrain dopamine neurons play in reinforcement learning. To this end, there are four major themes that are central to the development of my thesis and that I would like to review. The first theme of importance is that of associative learning. That is, the conditions under which animals learn the relationship between stimuli and responses. The second theme of importance is that of reinforcement learning models. In particular, the application of neural network models to understanding the functional role played by the dopamine signal in information gathering and broadcasting. The third theme that shall be covered is a review of the general anatomy of the dopamine system, with specific focus on the ventral tegmental area (VTA). Finally, I will review the functional role of dopamine activity in the midbrain. More specifically, I will look at the role that VTA dopamine neurons play in reinforcement learning.

1.1 Overview of adaptive behaviour and learning

The aim of this section (1.1) is to provide a background on associative learning based on the behavioural theories and findings of the past century. These theories and research findings highlight the complexities under which animals learn to associate stimuli and responses. Key to the development of this thesis is firstly an understanding of the circumstances that can affect the learning of an association (temporal contiguity between a stimulus and reward, the sensory modality of the stimulus, the number of trials in a session, etc…). Secondly, the behavioural paradigms reviewed in this section provide us with an understanding of the psychological phenomena that are being tapped into when animals are undergoing these tasks.

1.1.1 Associative learning: Classical and operant conditioning

The ability to learn the relationship between stimuli and responses is a defining characteristic of the process of associative learning (Wasserman & Miller 1997). This is in contrast with other processes of nonassociative learning such as habituation, where the continuous presentation of a single type of stimulus results in a diminished (habituation) response to that stimulus (Hollis 1997). Starting with the first theory of associative learning by Thorndike (Thorndike 1898), and moving to the pioneering work of Pavlov (Pavlov et al 1928), two major forms of associative learning have been identified: Pavlovian (or *classical*) conditioning and instrumental (or *operant*) conditioning. Both forms of conditioning share commonalities but also differ in a number of important ways.

The aim of this section, therefore, is to highlight what these differences are and to introduce behavioural paradigms that are used in Pavlovian and operant conditioning to unravel various characteristics of associative learning. This section will start with a description of classical conditioning.

In Pavlovian conditioning a neutral sensory stimulus, such as a light or tone (CS, or *conditioned stimulus*), signals the occurrence of a biologically salient event, for example presentation of food (*US*, *unconditioned stimulus*). This is often signified by the notation CS⇒US. After a number of repeated pairings between the two stimuli, learning is identified as the development of a *conditioned response* (CR; e.g. salivation) to the conditioned stimulus (Pearce & Bouton 2001).

There are a number of different classes of CR's, these include: autonomic, conditioned approach, conditioned place preference, and conditioned stimulus preference to name but a few. CS evoked changes in autonomic responses include changes in systolic blood pressure, breathing rate, pupil dilation, heart rate, or skin conductance and salivation (Grossberg et al 2008). Conditioned approach behaviour

occurs when repeated pairing of a CS with a US, results in CRs during CS presentation, which encompass approach to both CS and the location of US delivery (Blaiss & Janak 2008). In a conditioned place preference paradigm, an animal is initially allowed to freely explore two compartments. In the conditioning stage, one of the two compartments is associated with a positive stimulus (most commonly, the administration of a drug of abuse). In the testing stage, the animal is allowed to explore either of the two compartments, and the amount of time spent in the conditioned compartment is taken as a measure of the reinforcing properties of the drug (Russo et al 2008).

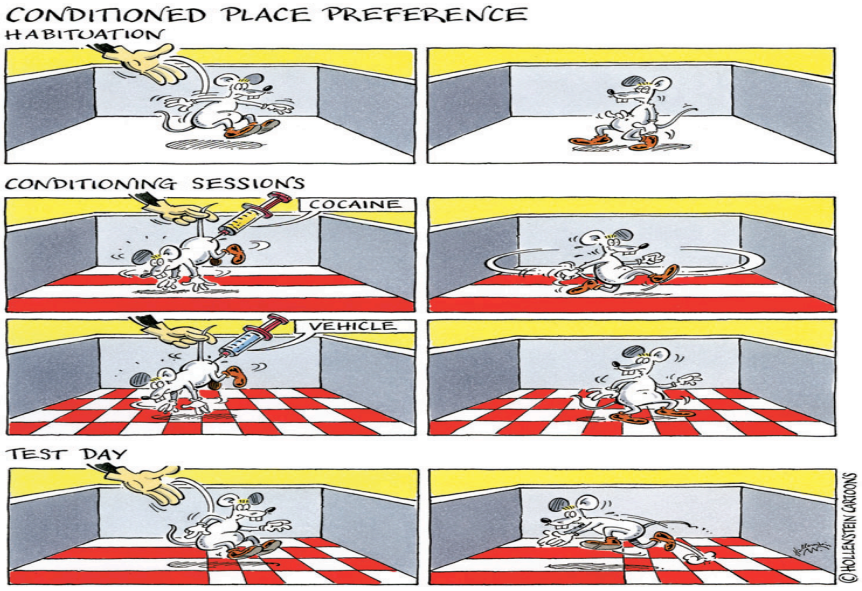


Figure 1: Adapted from (Sanchis-Segura & Spanagel 2006)

Finally, conditioned stimulus preference describes the preference for a given CS in the presence of a specific reinforcer (Quertemont & De Witte 2001). Overall, however, one key feature of Pavlovian conditioning is that the presentation of the US is independent of any conditioned response (CR) evoked by the conditioned stimulus (Wasserman & Miller 1997). From a functional perspective, therefore, Pavlovian conditioning is a form of adaptive behaviour in which an organism can gain advantages by anticipating biologically salient events given the presence of contextual

and environmental cues (Hollis 1997). In instrumental conditioning, however, the presentation of the US is dependent upon the instrumental response (R, typically the pressing of a lever), and because a stimulus signals that a response is required in order to obtain the US, this is known as a discriminative (S_D) and not a conditioned stimulus.

Therefore, one key feature that distinguishes operant from classical conditioning procedures is that behaviour is in effect controlled by its consequences. Moreover, recent studies have demonstrated that there are two distinct processes that govern instrumental conditioning. In the initial stage of acquisition, performance is goal-directed and modulated by the action-outcome (A-O) contingency (Yin et al 2004). Therefore, devaluation of the outcome has great impact on performance, as does contingency degradation (presenting the reinforcer without the operant response) (Corbit et al 2002). Nevertheless, in the later stages of training, performance becomes less goal-directed and insensitive to outcome devaluation and contingency degradation. That is, responding becomes habitual and stimulus-response driven (S-R) (S= the environmental context in which the R has a history of being reinforced) (Dickinson & Balleine 1990).

1.1.2 Research phenomena in associative learning

Here I aim to highlight the key research phenomena that have been used to study associative learning in animals. These behavioural paradigms illustrate the importance of understanding the conditions under which learning takes place. Moreover, I introduce the research phenomena of occasion setting and overexpectation that will be used in the experimental chapters (3&4) of this thesis.

1.1.2.1 Temporal and spatial contiguity

The delay or interval between CS⇒US presentation in Pavlovian conditioning and CR⇒US in operant conditioning is a significant factor that can affect response acquisition (Wasserman & Miller 1997). A number of studies looking at the effect of the interstimulus interval (ISI) in Pavlovian conditioning, for example, have shown that in order to achieve optimal learning, a short interval should be used (that is, the interval between CS offset and US onset), rather than the simultaneous presentation of CS⇒US (Rescorla 1988). However, other factors such as the nature and intensity of both the US and the CS, the number of trials per conditioning session, as well as the length of the intertrial interval can affect performance (Lennartz & Weinberger 1992).

In operant conditioning, increasing the delay between CRs⇒US reduces the potency of a reinforcer, an effect also known as temporal discounting of delayed reward (Wilkenfield et al 1992). In addition to temporal contiguity, spatial contiguity among contextual cues and the required behavioural response can also affect performance. In fact, in instrumental conditioning procedures, discrimination is improved if the response (R) spatially coincides with the discriminative stimulus (S_D, previously described in 1.1.1) (Rumbaugh et al 1989). Similarly, in Pavlovian conditioning paradigms, performance is improved if the CS and the US are located near to one another (Rescorla 1987).

1.1.2.2 Extinction

When a previously reinforced CS⇒US association undergoes subsequent pairing without reinforcer presentation, the CRs due to this treatment will diminish and eventually extinguish. This phenomenon has historically been explained as a case of unlearning, where old memories are replaced and destroyed by new ones (Pavlov, 1927). However, more recent accounts explain extinction as a case of further learning,

where the first- learned information is available along with new learning that has occurred as a result of extinction. Experimental evidence supports this view. If a previously extinguished CS is preceded by a strong novel stimulus (a stimulus similar to the CS), CRs to the CS can be restored.

Moreover, increasing retention intervals can provoke *spontaneous recovery o*f a previously extinguished CS (Calton et al 1996). In instrumental conditioning procedures, *spontaneous recovery* is aided if a short interval between training and extinction exists (Rescorla 2004). In addition, the extinction effect can be context specific. That is, if acquisition and extinction occur in the same context, CRs to the CS⇒ US+ will be lost. However, if testing occurs in a different context, CRs can re-emerge (Pearce & Bouton 2001).

1.1.2.3 The preexposure effect

A substantial exposure to a CS before CS⇒ US pairings can retard behavioural acquisition of the association (Lubow & Moore 1959). There are two main theories that attempt to explain this effect: attentional-perceptual and context associability theories. Attentional theories propose that subjects' attentional capabilities to the CS would be reduced after being pre-exposed to the stimulus (Lubow et al 1976). Hence, the theory implies that attention is a fundamental element of acquisition (Wasserman & Miller 1997) However, other theories focus on context-CS associations that when formed during CS preexposure interfere with the acquisition of the CS⇒US association (Wagner 1981). A number of studies indeed suggest that context extinction or shift between CS preexposure and CS⇒US pairings attenuates the CS preexposure effect (Hall & Pearce 1979).

1.1.2.4 Biological insignificance: sensory preconditioning

It has been argued that in order for learning to occur, the presence of a biologically relevant stimulus (the reinforcer) may not be paramount (Wasserman & Miller 1997). Therefore, the ability of a neutral stimulus in the absence of a US to elicit CRs would provide crucial evidence of the generality of associative principles. Early investigations on sensory preconditioning (Kimmel 1977), in fact, showed that if a whistle was paired with a light (CS_B&CS_A), followed by the light paired with a footshock (CS_A-US), and then the whistle was presented alone (CS_B), a CR was detected (leg flexion) despite the whistle having never been paired with the footshock (CS_B-US). This early demonstration showed that learning of biologically non-significant or neutral stimuli can occur, and is mediated by a process of associative conditioning, more specifically, by S-S (stimulus-stimulus) associations.

1.1.2.5 Second-order conditioning

Second-order conditioning or higher conditioning, is yet another example of the ability of a neutral stimulus, a stimulus that has never been directly paired with a relevant biological US, to elicit CRs (Pearce & Bouton 2001). In a slightly different manner to sensory preconditioning, CS_A is first paired with a US ($CS_A \Rightarrow US$), followed by $CS_B \Rightarrow CS_A$. In such instance, a CR develops to CS_B, which once again demonstrates that neutral stimuli can gain associative properties in the absence of a direct link with a biologically significant US (Rescorla 1973). The effect has been demonstrated in a number of species including honeybees, goldfish, and quail (Amiro & Bitterman 1980; Bitterman et al 1983; Crawford & Domjan 1995).

1.1.2.6 Serial conditioning

In a behavioural paradigm whereby CS_B precedes CS_A, which is in turn followed by US ($CS_B \Rightarrow CS_A \Rightarrow US$), the associative link may be formed between

$CS_B \Rightarrow US$ or between $CS_A \Rightarrow US$. However, one may be able to show that during this serial conditioning procedure, an association has been created between $CS_B \Rightarrow CS_A$ which does not require a direct link with the US (Wasserman & Miller 1997). One way of demonstrating this has been by showing that CRs to CS_B are stronger when $CS_B \Rightarrow CS_A \Rightarrow US$ are presented than CRs to CS_B in instances where CS_B and CS_A have been trained separately ($CS_B \Rightarrow US$; $CS_A \Rightarrow US$) (Schreurs et al 1993). Another way of demonstrating that learning can occur independent of conditioned stimuli-unconditioned stimuli (CS⇒US) representations is by the use of an autoshaping procedure, whereby pigeons are presented with the following:

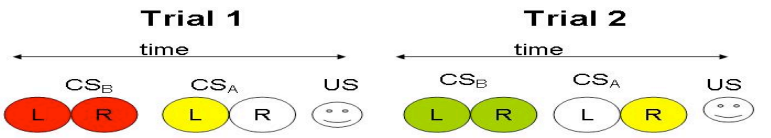


Figure 2: On any trial, CS_B involved the illumination of both keys (left-right) with the same colour (either red, or green, Trial 1 or 2). CS_A involved the illumination of one key only, and which key would be illuminated (left or right), was dependent on whether CS_B was red or green.

In such instance, CS_B is followed by CS_A, and then by the US ($CS_B \Rightarrow CS_A \Rightarrow US$) (see Figure 2 for clarification). In this paradigm, pigeons have been found to pick the left key (L) during CS_B with more frequency than the right key (R), if the illuminated key of CS_A was the left (L), and *vice versa*. This has been interpreted as evidence of specific associations between CS_B and CS_A (Wasserman et al 1978).

1.1.2.7 Cue competition: Overshadowing and blocking

Overshadowing and blocking effects occur when two conditioned stimuli come in competition with one another, due to salience or better ability to predict the

US by one of the two stimuli (Pearce & Bouton 2001). In overshadowing, CS_A and CS_B are presented together, followed by the US. CRs are then tested by presenting CS_A and CS_B individually. The general findings are that stronger CRs will normally develop with the stimulus (CS_A or CS_B) which is more intense or salient (Mackintosh & Reese 1979). If, on the other hand, CS_A is first trained with a US ($CS_A \Rightarrow US$), followed by simultaneous presentation of $CS_A \Rightarrow CS_B$ followed by the US ($CS_A \& CS_B \Rightarrow US$), stronger CRs will be elicited to CS_A than CS_B when presented individually (Balaz et al 1982). This is the phenomenon of blocking, whereby the prior pairing of one stimulus with a US, prevents subsequent conditioning to the second stimulus, an effect first described by the psychologist Kamin (Kamin, 1968).

The occurrence of overshadowing and blocking effects is, however, modulated by a number of complex factors such as the number of compound trials and by the sensory modality of the CS (Bellingham & Gillette 1981; Palmerino et al 1980).

1.1.2.8 Conditioned inhibition

A CS that predicts the non-occurrence of a US produces reduced CRs and is formally known as a conditioned inhibitor (Wagner & Rescorla 1972). More specifically, a stimulus becomes an inhibitor when paired with another stimulus that predicts US presentation, but the US does not occur (Wasserman & Miller 1997). Moreover, the inhibitory strength of an inhibitor is modulated by the magnitude of the omitted US (Pavlov 1927). There are at least two procedures that can reproduce this effect reliably. One such procedure, involves first presenting a CS with a US ($CS_A \Rightarrow US$) alone, interspersed with a compound presentation of a second CS (CS_B) with no US ($CS_A \& CS_B \Rightarrow$ No US) (Pavlov 1927). The other procedure entails presenting the compound stimuli ($CS_A \& CS_B$) in an unpaired manner (Rescorla 1968).

1.1.2.9 Occasion setting

The idea of a conditioned stimulus acting as an occasion setter goes back to Skinner in 1938, when he described that a discriminative stimulus does not itself elicit a response but sets the occasion for the response to occur (Schmajuk et al 1998). More specifically, a stimulus (CS) is said to have simple associative functions when it produces conditional responses (CRs) due to signalling the occurrence of an unconditioned stimulus (US) (Schmajuk & Buhusi 1997). In contrast, if a stimulus indicates the relationship between another CS and the US, it is said to act as an occasion setter (OS) or facilitator (Holland 1995). The OS, therefore, signals that another cue is to be reinforced instead of creating a direct link with the US (Schmajuk & Buhusi 1997). The distinction between stimuli acting in occasion setting fashion as opposed to simple conditioned stimuli has been largely provided using feature positive/negative discriminations. Ross and Holland (Ross 1983), for example, presented a serial sequence constituted by OS⇒CS ⇒+/CS ⇒-

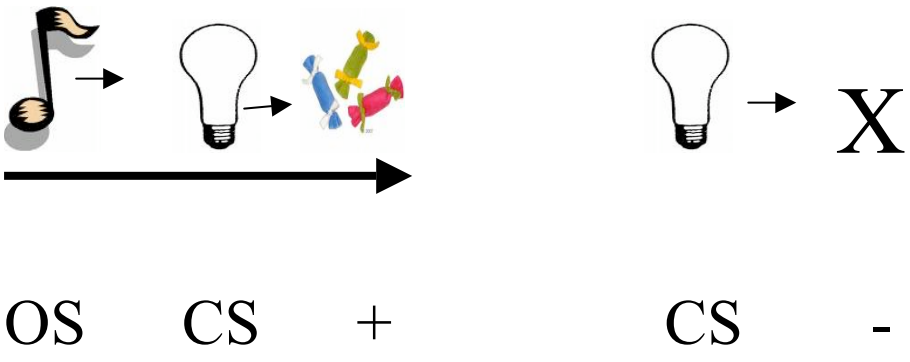


Figure 3: Occasion setting training: an animal is only able to collect reward+ if CS was preceded by OS

If it is the case that the OS is directly associated with the reward (+), then presenting the CS alone with no reward (-), after training of an OS⇒CS⇒ +/CS⇒- (that is, OS followed by CS followed by reward/ and CS followed by no reward, as in Figure 3) procedure, should reduce responding. However, the results showed that this was not clearly the case. Another demonstration of the occasion setting properties of a

stimulus is known as the "transfer" test. Studies have, for example, shown that by training OS with a different CS one can reduce CRs (Schmajuk et al 1998). This has been taken as evidence that the OS acts to specifically facilitate responding to the CS, and that changing the identity of the CS, disrupts the OS-CS relationship with reward (Holland & Lamarre 1984).

1.1.2.10 Overexpectation

The phenomenon of overexpectation is a counterintuitive, yet interesting effect that is predicted by a number of associative learning models, including the Rescorla-Wagner, which will be introduced in the next chapter. The effect refers to a decrease in responding to individual stimuli after these have been individually conditioned with the US, then paired with the US, and finally retested individually (Khallad & Moore 1996). A schematic illustration of an overexpectation experiment is here shown:

Acquisition training	Overexpectation training	Test
CS_A+	$CS_{AC}+$	$CR\|CS_B > CR\|CS_C$
CS_B+		
CS_C+		

Figure 4: In the acquisition stage, the animal is separately presented with conditioned stimuli A,B,C, followed by reward (+), In the overexpectation training, stimulus A and C are paired together followed by reward. In the test phase, CRs to stimulus B and C are contrasted and compared.

More specifically, Lattal and colleagues (Lattal & Nakajima 1998), were able to show that after individually conditioning stimulus CS_A, CS_B, and CS_C with a reinforcer (+), and pairing CS_A and CS_C together ($CS_{AC}+$), the conditioned response (CR) to stimulus $_C$ in the test phase relative to the control CS_B was diminished. The effect has been explained on the basis of error correction models (Rescorla & Wagner 1972). Briefly, the model predicts that learning is shaped by the difference between anticipated and obtained reinforcement (Dawson & Spetch 2005). Stimulus CS_A and CS_C, are seen as having associative strength *V*.

This is compared and contrasted with the maximum associative strength that can be supported by the US (λ). Once CS_A and CS_C are presented together, their associative strength (*V)* is summed (2*V*). However, as the US (λ) remains unchanged (does not double), there will be a negative discrepancy between total reward (λ) and anticipated reward (2*V)*. Hence CS_A and CS_C associative strength decrease. The results of the Rescorla-Wagner model will be further explored in one of the experimental chapters of this thesis.

1.1.3 Theories of associative learning

A number of theories have been advanced that have attempted explaining the empirical literature of associative learning. Here we aim to review and assess the most prominent ones and relevant to this thesis.

1.1.3.1 Rescorla and Wagner model (1972)

The Rescorla and Wagner model (1972) is perhaps the first model to view conditioning as occurring on a trial-by-trial basis. It was also the first model to treat each cue of a multiple compound presentation interacting with one another. Hence, the associative strength of a CS⇒ US pairing, changes as a function of the presence of other CSs (Miller et al 1995). This change in associative strength is determined by the difference between the maximum associative strength supported by the US and the sum of all associative strengths of all CS's present in a given trial (Pearce & Bouton 2001). In addition, the salience of the CS and the features of the US (learning parameters) modulate the extent of associative strength change (Wasserman & Miller 1997). As a result, the model predicts that learning is determined by the level of "surprise" that an organism or entity comes to expect from the US (Miller et al 1995).

More specifically, by the difference between the US presented in the trial, and the predicted value obtained from all the CS's present in the trial. Therefore, as

the number of trials increase, and provided that the same CS(s) and US are presented, the difference between what is presented and expected decreases, and learning diminishes (Miller et al 1995). The Rescorla-Wagner model (1972) has been particularly successful in predicting the effects of stimulus generalization, discrimination, conditioned inhibition, overshadowing, blocking and overexpectation (Wasserman & Miller 1997). However, the model has also had its troubles predicting spontaneous recovery from extinction paradigms, CS-preexposure effects, second order conditioning, and learned irrelevance (uncorrelated preexposure presentation of CS-US pairings, which leads to retardation of subsequent formation of excitatory or inhibitory associations between the two stimuli) (Bonardi & Yann Ong 2003; Miller et al 1995).

1.1.3.2 Pearce-Hall (1980)

The Pearce and Hall (1980) model makes specific predictions about the causal role played by attention in regulating associability. For example, it has been found that if a CS_A is paired with a US_1 (mild shock) for a substantial number of trials, and then this same CS_A is paired with US_2 (strong shock), there will be a weaker CR than the CR to a CS_B paired with the US_2 (Hall & Pearce 1979). This phenomenon has been interpreted as CS_A impairing the formation of new associations given that CS_A was an accurate predictor of the shock (Pearce & Bouton 2001). Hence, the associability of a stimulus will be highest when an unexpected US is due to occur.

The Pearce-Hall (1980) model has been able to successfully predict the CS-preexposure effect and the negative transfer effect (see 1.1.2.3), although a measure of associability other than the orienting response to a stimulus needs to be found (Wasserman & Miller 1997).

1.1.3.3 Pearce (1987, 1994)

The Rescorla-Wagner model (1972) conceptualises each cue in a multiple compound presentation as acting individually. Therefore, the associability strength of an CS_{AB}+ pairing (CS_A and $CS_B \Rightarrow$US), for example, will be determined by the associative strength of CS_A summed with that of CS_B, as if each conditioned stimulus was acting in "elemental" fashion (Deisig et al 2001). However, another interpretation of a CS_{AB}+ pairing is that an organism perceives CS_{AB} as an overall entity, and therefore the elements of the compound come to be associated as a single block. This interpretation is otherwise known as the "configural" approach (Pearce & Bouton 2001).One of the major proponents of a configural approach to learning is Pearce (1987, 1994). In Pearce's (Pearce 1987; 1994) model, and in contrast to Rescorla-Wagner predictions, the associability strength of a CS_{AB}+ compound is determined by the similarity of CS_A and CS_B to CS_{AB} (which will be .5, provided that they are equally salient). As a result of this similarity, responding to CS_{AB} will be equal to CS_A or CS_B alone (Pearce 1994).

The model has shown to be better able to predict the effects of stimuli interaction in multiple compound paradigms than the Rescorla-Wagner model has. This is of particular importance with regards to the predictions it makes about the overexpectation effect which will be further explored in Chapter 2 and 4. The Pearce model is also able to make more accurate predictions than Wagner's model with regards to negative patterning effects. In fact, in negative patterning discriminations, when CS_{AB}- presentations are followed by individual CS_A+, CS_B+ pairings, the Rescorla-Wagner model would predict greatest response to the compound presentation (CS_{AB}-). Pearce's theorem, however, is able to correctly predict the animal enhanced response to the single elements (CS_A+, CS_B+). This has been shown in a variety of species, including rats and honeybees (Deisig et al 2001; Pearce 1987).

1.2 Reinforcement learning models

A primary objective of this thesis is to refine the current neural network models of adaptive behaviour that are inspired by the biology of reinforcement learning. Thus, the purpose of this section is twofold. The first purpose is to provide a general background on what these methods are. The second purpose is to understand how specific neural network models (temporal difference methods) can be used to understand the functional role played by the dopamine signal in reinforcement learning.

1.2.1 An introduction to reinforcement learning

The goal of a reinforcement learning agent is to learn to maximise rewards. Reinforcement learning differs from supervised learning in that the learner is not told what to do, but is required to discover the most beneficial actions (evaluative feedback) through trial and error search, and delayed reward (Sutton & Barto 1998). One of the greatest advantages that reinforcement learning offers over supervised learning is its greater flexibility in unknown environments, whereby the learner's own experience is crucial in determining the best action, rather than having to adapt a limited set of instructions to novel situations as it occurs with supervised learning (Barto 1995).

On the other hand, reinforcement learning agents encounter a number of challenges that other forms of learning do not experience. One of these is the temporal credit assignment. That is, how to establish which actions in the past had the effect of causing the desired outcome. The other major challenge is the trade-off between exploration-exploitation (Sutton & Barto 1998). That is, in order to maximise rewards, an agent has to strike a balance between exploring new actions

that may lead to better rewards, and exploiting the knowledge that it already has (Sutton & Barto 1998). There have been two major ways of solving the exploration-exploitation dilemma; namely, by using e-greedy and softmax action algorithms. E-greedy policies exploit knowledge most of the time, but with small probability they explore an action at random, independent of action-values estimates. A drawback of this method is that by giving equal weight to all actions when exploring, they are just as likely to select the worst of all action as to select the next-to best action. Softmax policies improve this shortcoming by varying the action probabilities as a function of graded value estimates (Sutton & Barto 1998).

In order to understand the mechanics that drive a reinforcement learning system, it is helpful to identify its constituent parts. Thus, the fundamental parts of a reinforcement learning system are made up by a policy, a reward function, a value function, and a model of the environment (Bertsekas & Tsitsiklis 1996). A policy determines the behaviour of the agent at a given time, given a set of environment stimuli (Kaelbling 1993). A reward function defines the intrinsic desirability of external rewards (Sutton & Barto 1998). A value function estimates the total amount of reward an agent can expect to receive in the long-term. This is in contrast with the reward function that only determines the desirability of a reward in the immediate present (Barto 1995). Finally, a model of the environment is also important, as it allows an agent to determine what to do at a given point, given a variety of possible outcomes in the future (Sutton & Barto 1998). A policy differs from a model of the environment on the basis that it determines behaviour on a "stimulus-response" action, without planning capabilities.

In summary, an agent interacts with the environment through a series of time steps. At each time step, the agent is in a given state that determines the action that it will take. Certain actions will lead to particular rewards, which will in turn form the

basis of further actions. In addition, the agent needs to strike a balance between selecting immediate rewards and rewards delayed in time. This decision is achieved through a discount rate function, which determines the value of future rewards (Sutton & Barto 1998).

1.2.2 Three solutions to the reinforcement learning problem

Here we will review three classes of methods that have been applied to solve the reinforcement learning problem. Namely, dynamic programming and Monte Carlo methods but we will focus most of the attention to temporal difference methods, as these are most relevant to the study of the dopamine system in reinforcement and to the development of this thesis.

1.2.2.1 Dynamic programming

One of the defining characteristics of dynamic programming (DP) is that it computes optimal policies on the assumption of a perfect model of the environment (Markov decision process, MDP) (A Markov property (or MDP) refers to a state signal of the environment that retains all relevant information but without a complete history of previous events) (Sutton & Barto 1998). In a more realistic case, where the agent has partial knowledge of the world/environment, the model is known as a partial MDP (Bertsekas & Tsitsiklis 1996). In order to compute optimal policies, a DP system uses two special types of algorithms: a policy evaluation and a policy improvement. A policy evaluation simply calculates the value functions for a given policy, whereas a policy improvement calculates an improved policy given the value function for that policy (Sutton & Barto 1998). The other special characteristic of a DP system is its use of "bootstrapping". That is, the values of each state are computed on the basis of estimates of future states (Giegerich 2000). DP methods are overall fast and efficient, however, they are not suitable in finding optimal reinforcement

learning solutions, and are thought to be of limited applicability due to the difficulty that DP systems have in handling large state sets (Sutton & Barto 1998).

1.2.2.2 Monte Carlo methods

Whereas DP methods computed optimal policies on the assumption of a perfect model of the environment, Monte Carlo methods (MCM) discover optimal policies from on-line and simulated experience (Sutton & Barto 1998). Therefore, in order for learning to occur, MCM do not require perfect probability distributions of all possible transitions as it is the case with DP methods, but only to generate sample transitions (Kalos & Whitlock 1986). However, whilst sample experience is a unique aspect that differentiates MCM from DP methods, MCM methods still possess a policy evaluation, a policy improvement, and finally a policy iteration as DP methods do (Sutton & Barto 1998). Policies and value estimates are updated only once "episodes of experience" are terminated, and hence MCM solve a reinforcement learning problem by averaging sample returns (Barto 1995). Due to the restricted applicability of MCM, mostly in solving tasks that are delineated by a set of "episodes", other approaches such as temporal difference (TD) paradigms have been adopted as they appear to be better capable of quickly learning and altering their behaviour.

1.2.2.3 Temporal difference algorithms

The underlying architecture of temporal difference (TD) models much more resembles that of Monte Carlo methods than that which we find in dynamic programming methods. Most prominently, TD models do not require a model of the environment as DP methods do (Sutton & Barto 1998). However, TD methods differ to MCM in a number of important ways. For example, TD methods update their estimates in an on-line, one-time step fashion, whereas with MCM, the update of their

estimates can only occur at the end of an episode (Suri & Schultz 2001a).This difference has a crucial impact on the speed by which learning takes place. If the episodes are long ones, then MCM are substantially slower than TD methods (Sutton & Barto 1998).

In addition, MCM can be slower than TD methods due to having to discount episodes on which exploratory actions are taken, whereas TD methods learn from each transition and do not need to ignore experimental actions (Suri & Schultz 1998b). Nevertheless, TD methods like MCM and to some extent DP methods are based on the idea of generalised policy iteration to reach optimal values. TD methods, however, have been favoured over MCM and DP in solving the reinforcement learning problem, due to a combination of better able to fit the empirical data, greater simplicity, minimal computation, and the possibility of being applied on-line (Sutton & Barto 1998).

1.2.2.4 The architecture of a temporal difference model of reinforcement learning

One of the greatest merits of TD methods is that they have been successfully applied in a variety of contexts. For example, they have been used to learn to play backgammon and to balance a pole on a cart wheel (Schultz et al 1997). More importantly, and from a biological point of view, TD models have been used to replicate foraging behaviour of honeybees, the learning of voluntary eye movements, and the simulation of human decision making (Friston et al 1994; Montague et al 1995; Montague et al 1996). And of particular relevance to this thesis, they have been used to understand the role that the dopamine signal plays in terms of information construction and broadcasting (Montague et al 1996). Whilst being relatively simple to compute, it remains important to understand some of the key algorithms that make up a TD model of reinforcement learning. Firstly, the model assumes that sensory

cues are used to predict all future rewards within a learning trial (Schultz et al 1997). The equation is as follows:

$$V(t) = E[y^{\circ} r(t) + y^{1} r(t+1) + y^{2} r(t+2) + \ldots$$

Figure 3: Adapted from Schultz et al (1997)

V (t) represents the prediction of all future rewards. r(t) corresponds to reward (r) at a given time (t), and E the expected sum of all future rewards up to the end of the trial. $0 \leq y \leq 1$ signifies that rewards that are delayed in time have less "importance" (discounted) than those that arrive sooner.

The prediction as to which sensory cue or reward is presented is assumed to be based only on the current and not the past sensory cues (Schultz et al 1997). In order to estimate the sum all future rewards (V (t)), the agent compares and adjusts its predictions at each time step, instead of having to wait for all its future rewards to assess its prediction (Barto 1995; Daw & Doya 2006; Schultz et al 1997; Suri & Schultz 1998b). Therefore, an error is generated between prediction and outcome, which is laid out in the following equation:

$$\delta(t) = r(t) + y\hat{V}(t+1) - \hat{V}(t)$$

Figure 4: Adapted from Schultz et al (1997)

This TD error δ(t) acts as a prediction error signal at time t+1, and aids in improving the estimates of V (t) (Schultz et al 1997). In order for the temporal prediction between sensory cues and rewards to be accurate, the agent requires having not an adjustable weight (w) per sensory cue, but a set of weights, one for each timestep cue's onset. This form of temporal representation is otherwise known as serial-compound stimulus, and it is an integral feature of most TD models of reinforcement learning (Schultz et al 1997; Sutton & Barto 1998). The equation which describes stimulus representation through time is as follows:

$$\hat{V}(t) \equiv \hat{V}(x(t)) = \sum_i w_i x_i(t)$$

Figure 5: Adapted from Schultz et al (1997)

Correlation between stimulus representation and prediction error, improves the performance of adjustable weights. In addition, the change in weights from one trial to the next by the learning rate for cue x (t), helps to bridge the gap between the true value of V (t) (update rule) (Schultz et al 1997; Suri 2002). From a biological perspective, adjustable weight can also be thought as synaptic connections. The update rule is shown in the equation below, where α_x stands for the learning rate parameter (The learning rate parameter determines the number of trials that that are needed for the development of cue responses, and the disappearance of responses to rewards).

$$\Delta w_i = \alpha_x \sum x_x(t)\, \delta\,(t)$$

Figure 6: Adapted from Schultz et al (1997)

1.2.2.5 TD models and their relation to the dopaminergic system

TD models of reinforcement learning produce prediction error signals that very much resemble the neuronal activity showed by midbrain dopaminergic neurons (Montague et al 1996; Schultz 1999; Suri & Schultz 1998b).That is, the presentation of an unpredicted reward, and the earliest reward predictive stimulus, produces positive prediction errors (better than expected), whereas omitting a reward results in negative prediction errors (worse than expected) (Schultz et al 1997). This has a striking resemblance to the activity of dopamine neurons, as reviewed in (See 1.4.4): the following figure demonstrates more clearly the findings:

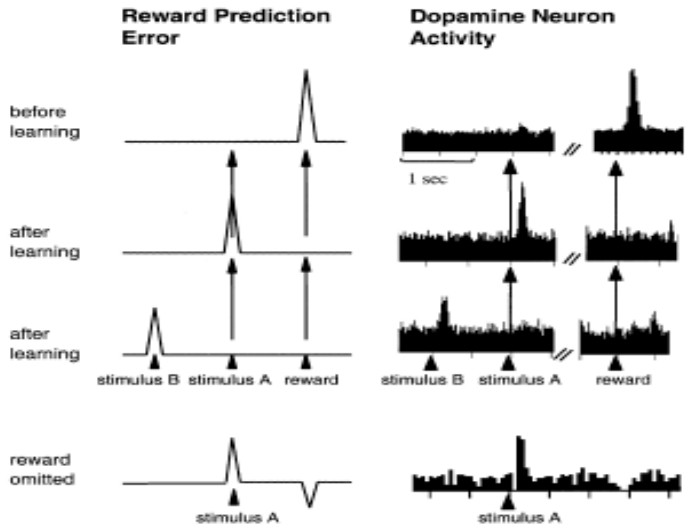


Figure 7: Responses of TD model and dopamine neurons to stimuli presentations. Adapted from Suri et al (2002)

TD models of reinforcement learning such as those developed by Suri and Schultz have been able to reproduce the activity of dopamine neurons in a number of contexts (Suri & Schultz 1999). These include: 1) presentation of an unpredicted reward 2) unexpected omission of reward 3) blocking 4) delayed reward 5) before, during, and after learning that a stimulus (CS) precedes reward 6) when two stimuli precede a reward with fixed timed intervals 7) when intervals between two stimuli are unpredictable 8) with novel, physically salient stimuli 9) with rewards that occur earlier than expected 10) with unexpected omission of a stimulus that is reward predicting (Suri & Schultz 2001a; Waelti et al 2001).

In the following sections, we will explore two TD models of reinforcement learning that are particularly relevant to the work presented here.

1.2.2.6 The Montague model of TD learning

Montague *et al* (Montague et al 1996) developed a model based on TD learning algorithms that compared previous neurophysiological data of dopamine neurons in monkeys performing a spatial choice and a delayed response task with the model output (Ljungberg et al 1992; Schultz et al 1993). In addition, the theory made

testable predictions of the activity of dopaminergic neurons in humans performing a card choice experiment (Egelman et al 1995). The model was successful in being able to account for the neurophysiology data and is illustrated schematically here below:

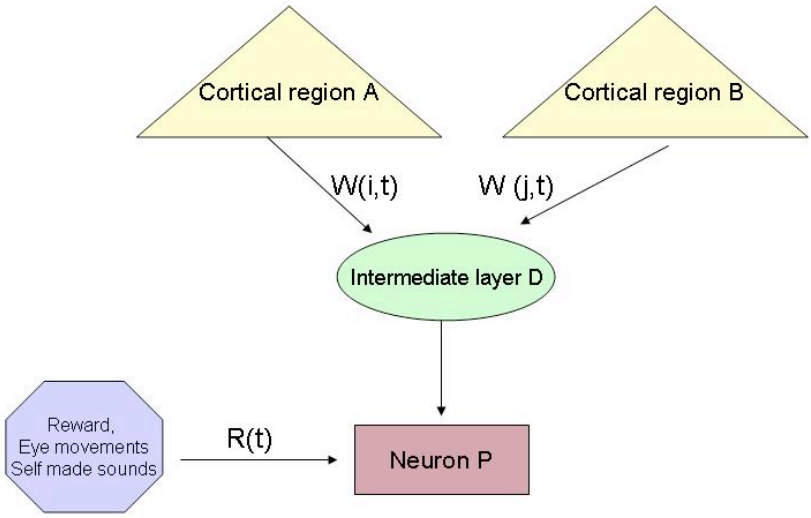


Figure 8: Adapted from Montague et al (1996).

P, represents a small number of dopamine neurons which receive inputs from the two cortical representations of sensory events (i,j), and from rewarding/salient events in the environment and within the organism r(t) (Montague et al 1996). (i), stands for cortical domain, and (t) for time. The cortical domain (i) (here we will focus on i only for illustrative purposes) is linked to adjustable weights (w(i,t)), which in turn influence the strength on P at time (t) after its onset (Montague et al 1996).

The connections from the cortex to the dopamine neurons P are conceived as indirect, first synapsing at an intermediate region (neuron D)(the output of D signalling changes in weighted sensory inputs only). Theoretically, therefore, weight changes are thought of as occurring anywhere from the cortex, to the intermediate

region and finally reaching P. The time representation of a sensory stimulus is based on serial compound stimulus algorithms, which assign a value *x* to each timestep following stimulus presentation (Sutton & Barto 1998). Finally, decisions (for more details on the decision rule, please see 1.2.1) are ultimately based on the prediction error rule, which calculates the difference between the expected future reward and the reward received, and stamps a transition as better than expected $\delta(t) > 0$ or worse than expected $\delta(t) < 0$ (Montague et al 1996).

1.2.2.7 The Pan *et al.* model of TD learning

Pan *et al.*'s TD learning model was based on Montague's work, but with three important differences: it modified the eligibility-trace decay parameter, the learning rate, as well as limiting the amplitude of the negative prediction error (Pan et al 2005). Altering these three parameters of the model had a profound effect on being able to better reproduce the activity of dopaminergic neurons whilst undergoing the same sequence of cues and reward.

The tasks that both rats and model were presented with consisted of a number of single and two-cue paradigms, interspersed with random reward presentation, cues only, and cue omission. The reason for limiting the amplitude of the negative prediction error was based on the observation that dopaminergic neurons respond in asymmetrical manner to excitation and inhibition, as compared to baseline activity (from 5Hz, dopaminergic neurons can be excited to 100Hz, but can only be inhibited to 0Hz) (Pan et al 2005). In previous work, the 5Hz baseline activity has been equated to a state of 0 prediction error (Schultz et al 1997). Therefore, in Pan's model, the range of positive and negative prediction errors (above and below 5Hz), has been scaled (the maximum prediction error was limited to -0.05, and firing at 100Hz

equated to values close to 1) in such a way as to better reflect the activity of dopaminergic cells (Pan et al 2005).

The eligibility trace parameter (ETP), on the other hand, determines the extent to which pathways encoding events that occurred in the past are eligible for undergoing learning change (Sutton & Barto 1998). In other words, ETP stands for the rate at which sensory events are forgotten by the system: the more of a trace that is forgotten before the US occurs, the less the system learns about it. More specifically, in TD learning models whereby the ETP is set to 0, which signifies immediate loss of the memory trace for the sensory event after its offset, only the most recent state is changed by the TD error (Pan et al 2005). If the ETP approaches 1, however, more of the preceding states are changed (Sutton & Barto 1998). The effect of an ETP set to 0 is to produce a stepwise migration of responses from rewards to cues, so that responses to both cues and rewards do not overlap (Pan et al 2005).

In Pan's work, however, dopaminergic cells at least early in training responded to both cues and reward presentation, suggesting that TD models using an ETP set to 0 are unable to replicate the activity seen *in vivo* (Montague et al 1996). Using an ETP set to 0.9, which means a 10% reduction in the strength of the trace with each time step, Pan *et al.* demonstrated a better match between the model and the neurophysiological data. In addition, the model included changing of the learning rate parameter (LRP).

The learning rate parameter determines the number of trials that are needed for the development of cue responses, and the disappearance of responses to rewards (Pan et al 2005). By having a low learning rate parameter, the model was able to successfully predict the number of trials needed to abolish reward responses. Finally, and more importantly perhaps from a physiological perspective, the results suggest that the dopaminergic cells retain responses to rewards and conditioned cues for a

prolonged period of time, and only after extensive training, do the firing activity to the rewards disappear. More recently, Pan *et al.,* introduced two modifications to their original model (Pan et al 2008). They recorded dopaminergic activity during a behavioural extinction paradigm, and used the empirical findings to build an extended TD model based on their previous work (Pan et al 2005). The model included the addition of two sets of weights, one that dealt with zero to positive values (excitatory/positive), and the other that dealt with zero to negative ones (inhibitory/negative).

Interestingly, excitatory and inhibitory weights were not only activated by the prediction error, as in other TD models, but also by spontaneous decay at each time step. Therefore, weight changes driven by positive/negative prediction error were referred to as learning and unlearning respectively, and weight changes by spontaneous decay were referred to as forgetting. Crucially, for the ability of the model to simulate dopaminergic activity, four parameters determined the differential rate of change for positive and negative weights.

Firstly, positive weights are strengthened (by positive prediction error) when cues are followed by rewards (p+ or learning) and weakened when cues do not follow rewards (p- or unlearning). Both rates of changes are modulated by a parameter (α). Secondly, negative weights are strengthened by a negative prediction error at a rate set by a parameter (β).Thirdly and fourthly, when no prediction error occurs, both positive and negative weights undergo forgetting at rates modulated by a parameter (ψ^{+}) and a parameter (ψ^{-}). Whilst previous TD models had been able to simulate the elimination of responses through unlearning, this new modified model successfully displayed additional characteristics of behavioural extinction, namely, spontaneous recovery, speeded relearning, and better recovery with longer inter-test intervals.

1.3 General anatomy of the dopamine system

The aim of this section is to review the anatomical, pharmacological, and electrophysiological properties of one of the pivotal structures that is believed to be involved in reinforcement learning: the ventral tegmental area (VTA). The VTA is also the brain region of interest in the two experimental chapters of this thesis (Chapter 3 & 4) and where I will be performing extracellular single neuron recording.

1.3.1 Ventral tegmental area anatomy

The ventral tegmental area (VTA) is part of what we come to know as the mesencephalic dopamine system, which is further subdivided into the nigrostriatal and the mesocortolimbic system (Mathon et al 2003).

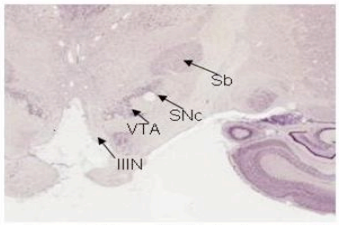


Figure 1: Schematic illustration of VTA Mucaca mulatta.
IIN= oculomutor nerve; SNc= Substantia Nigra Pars Compacta; Sb= Subthalamic nucleus

Adapted from BrainMaps.org: Nissl staining of VTA neurons in the rhesus monkey (Mucaca mulatta)

One of the pathways originates in the substantia nigra (A9), projects to the dorsal striatum, and is known as the nigrostriatal pathway (Ikemoto 2007). The other, originates in the VTA (A10) and projects (mainly) to the nucleus accumbens and the prefrontal cortex (Ferreira et al 2008). The VTA sits ventral to the red nucleus and

medial to the substantia nigra in the midbrain (Fields et al 2007). The first identification of the VTA, as an independent and separate unit, was made in the opossum brain which was named "nucleus tegmenti ventralis"(Tsai 1925). However, it was only later on that the region was identified as VTA or "ventral tegmental area of Tsai" when it was noted that the lateral hypothalamic area projects to the VTA but not to the substantia nigra, providing evidence of a hodological (based on the interconnections of brain areas) distinction between the two areas (Nauta 1958).

A more recent investigation using retrograde axonal tracing has been able to show that the striatum receives most dopaminergic input from neurons in the SNpc (Subtantia nigra *pars compacta*) labelled for the G-protein *Girk2*, and that the frontal cortex receives preferential dopaminergic innervation from VTA labelled for the calcium binding protein calbindin (Thompson et al 2005). Further work, identified neurochemical cell groups throughout the rat brain using the terms A1-12 and B1-9 nuclei, with dopaminergic neurons identified as A8, A9, and A10; the A10 nuclei being synonymous with VTA (Dahlstroem & Fuxe 1964). Largely on the basis of this neurochemical labelling, the VTA was envisaged to encompass three midline nuclei; the rostral, central and the interfascicular nucleus (Oades & Halliday 1987). More specific cytoarchitectonic divisions of the VTA, however controversial these may be, include the paranigral nucleus (PN), the parabrachial pigmented area (PBP), the parafasciculus reflexes areas (PFR), and the ventral tegmental tail (VTT) (Ikemoto 2007).

One of the distinguishing characteristics of these four subdivisions is related to the dopaminergic density: the PN and PBP high in density, whereas PFR and VTT low (McRitchie et al 1996). Overall, in the rat brain there are 30000 VTA neurons, of which fewer than 60% are dopaminergic, with the remaining ones being glutaminergic and GABAergic (Fields et al 2007). Defining the exact borders of the

PBP has proved more difficult. However, cell bodies here tend to be large, and without unified orientations (Ikemoto 2007). Finally, posterior to the PN, the VTT is characterised by small cell bodies. Whilst clearly separated from the PN and PBP, the cytoarchitectonic features of the VTT remain similar enough for it to be included in the VTA nomenclature (Ikemoto 2007).

1.3.2 Inputs to the VTA and their neurotransmitter profile

A number of CNS sites target the VTA. Glutaminergic innervations to the VTA originate from the prefrontal cortex (PFC), the lateral hypothalamus (LH), the bed nucleus of stria terminalis (ST), and the superior colliculus (SC) (Fields et al 2007). A mixture of glutaminergic, cholinergic, and GABAergic inputs to the VTA are provided by the pedunculopontine tegmental nucleus (PPTg) and the laterodorsal tegmental nucleus (LDT) (Oakman et al 1995).

The PPTg, in turn, receives sensory input from the superior and inferior colliculi, the lemniscal nuclei, the trigeminal complex and the parabrachial nucleus (Winn 2008). GABAergic inhibitory afferents to the VTA include the ventral pallidum (VP) and the nucleus accumbens (NAcc) (Geisler & Zahm 2005). In addition, converging noradrenergic and serotonergic inputs originate from the locus coeruleus and the dorsal raphe nucleus respectively (Geisler & Zahm 2005). Finally, additional innervations to the VTA stem from the central amygdala and the preoptic area of the hypothalamus (Wallace et al 1992).

1.3.3 VTA projections

The largest projections from the VTA are to the NAc core and shell, the PFC, the central and basolateral amygdala, and the lateral hypothalamus (LH) (Berger et al 1974). However, the VTA also makes a number of smaller projections to other areas such as the hippocampus, the lateral septal area, and the entorhinal cortex (Swanson

1982). VTA projections appear to be richest in dopamine in the NAcc (where the percentage of projecting VTA neurons identified as dopaminergic is 65-85%), the lateral septal area (72%), and poorest in the PFC (30-40%) and hippocampus (6-18%) (Margolis et al 2006). However, there also appear to be GABAergic and glutaminergic VTA projections to the PFC and the NAcc (Carr & Sesack 2000; Lavin et al 2005). Interestingly, in cultured VTA neurons, administration of the D_2 antagonist sulpiride increases glutaminergic EPSPs, suggesting that in some neurons, glutamine acts as a cotransmitter of dopamine (Sulzer et al 1998). However, whether cotransmission of dopamine and glutamine occurs *in vivo*, remains to be ascertained (Fields et al 2007).

1.3.4 VTA inputs and their circuitry

In order to understand the VTA circuitry, we must take into account the neurotransmitters of its inputs, as well as where VTA neurons project. Whilst the VTA circuitry is only partially understood, a number of significant connections have been identified. In fact, we know for example that PFC afferents innervate VTA dopamine neurons that in turn project back to the PFC (Fields et al 2007). However, other PFC afferents target GABAergic VTA neurons that do not project back to the PFC but to the nucleus accumbens (NAcc) (Carr & Sesack 2000). In addition, only a small proportion of lateral hypothalamus (LH) neurons synapse with VTA dopamine neurons projecting to the NAcc, nevertheless, a more significant number of LH neurons innervate VTA dopamine neurons projecting to the PFC (Omelchenko & Sesack 2006). Excitatory and inhibitory laterodorsal tegmental nucleus (LTD) neurons project to both dopaminergic and GABAergic VTA neurons projecting to the PFC, but LTD+ (excitatory) neurons synapse with VTA dopamine neurons projecting

to the NAcc, whereas LTD- (inhibitory) neurons synapse with VTA GABAergic neurons projecting to the NAcc (Fields et al 2007).

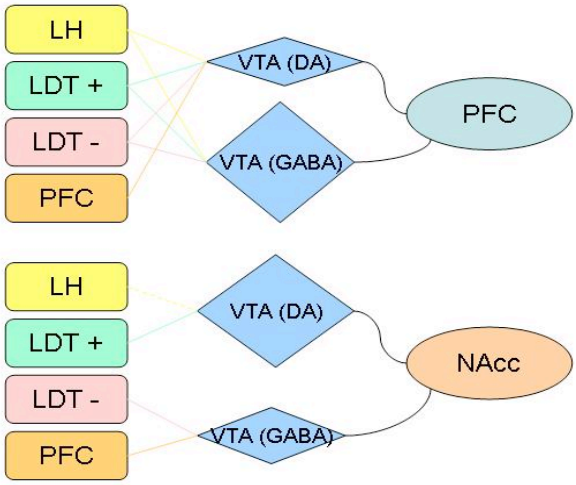


Figure 2: Schematic diagram of VTA circuitry. VTA neurotransmitter content, and its projections determine the type of innervation that VTA neurons will receive. Each input to the VTA is represented in a different colour for easier visual discrimination. The size of VTA (DA) and VTA(GABA) is used to show the proportion of neurotransmitter content. LH (lateral hypothalamus, LDT (laterodorsal tegmental nucleus) PFC (prefrontal cortex. + and – signs represent patterns of excitatory and inhibitory inputs. Adapted from Fields et al (2007).

1.3.5 VTA D_1 receptors

The presence of D_1 receptors in the VTA is estimated to be moderate to low (Adell & Artigas 2004). D_1 receptor antagonist perfusion in the VTA produces release of glutamate and GABA (Kalivas & Duffy 1995). Moreover, blockade of D_1 administration impairs cocaine self-administration, suggesting a mediating role of glutamate and GABA in controlling the activity of dopamine neurons (Cameron & Williams 1993). In addition, concurrent stimulation of D_1 and D_2 receptors in the NAc is required for a reduction of dopamine release in the VTA (Rahman & McBride 2001).

1.3.6 VTA D_2 receptors

In contrast to D_1 receptors, D_2 receptors are highly expressed in the VTA of rodents (Wamsley et al 1989). These, appear to be mainly located on extrasynaptic

plasma membrane (defined as the external limiting lipid bilayer of cells) of dendrites, and they mostly function as autoreceptors (Adell & Artigas 2004). In the VTA, systemic administration of agonists and antagonists diminishes the firing rate of dopaminergic neurons, suggesting that D_2 autoreceptors have a central role in dopamine inhibition (Adell et al 2002). However, D_2 receptors can also function as heteroreceptors (defined as a receptor modulating the release of neurotransmitters other than its own ligand) in non dopaminergic VTA neurons (Adell & Artigas 2004). Mostly, this occurs through D_2 receptors located in glutaminergic terminals which regulate the firing rate of dopaminergic neurons in the VTA (Koga & Momiyama 2000).

1.3.7 VTA interaction with non dopamine neurotransmitters

1.3.7.1 Serotonin

Serotonergic inputs to the VTA mostly arise from the dorsal and medial raphe nucleus (DR, MnR) (Herve et al 1987). Electrical or pharmacological stimulation of the DR results in inhibition/excitation of dopaminergic neurons in the VTA. This is dependent upon the specific 5-HT receptor manipulated (in anaesthetised recording) (Adell & Artigas 2004). At least, four serotonin receptors have been identified in the VTA. The 5-HT_{1A} receptors, for example, appear to have a role in tonic inhibition of dopamine neurons. Depletion of 5-HT stores, in fact, abolishes the ability of 5-HT_{1A} receptor agonists to increase dopamine release in the VTA (Prisco et al 1994). The 5-HT_{2A} receptors, on the other hand, contribute to the phasic, excitatory response of dopamine VTA neurons. Stimulation of 5-HT_{2A} receptors in vitro augments the firing rate of VTA dopaminergic neurons, and administration of a 5-HT_{2A} antagonist abolishes the effect (Pessia et al 1994).

The 5-HT_{2C} receptors, on the contrary, appear to have a similar role to the 5-HT_{1A} receptors in producing a tonic inhibition of VTA dopaminergic neurons (Adell & Artigas 2004). 5-HT_{2C} receptor agonists decrease the firing rate of VTA dopaminergic cells, whereas 5-HT_{2C} antagonists increase the frequency of bursts (Di Giovanni et al 1999). Finally, the 5-HT_3 receptors offer similar functions to the 5-HT_{2A} receptors in regulating the tonic action of dopaminergic cells (Adell & Artigas 2004). 5-HT_3 receptor antagonists, in fact, reduce the basal firing rate of active VTA dopaminergic cells (Miyata et al 1991). Altogether, these findings suggest that 5-HT release in the VTA can have inhibitory/excitatory effects on VTA dopaminergic neurons depending on the particular 5-HT receptor involved (Adell & Artigas 2004).

1.3.7.2 Noradrenaline

The locus coeruleus is an important site for the modulation of dopaminergic activity in the VTA through noradrenergic transmission (Herve et al 1982). Abolishment of D_2 somatodendritic autoreceptors function in the VTA, concomitant with the stimulation of α_1-adrenoreceptors, increases VTA dopaminergic cell firing (in vitro) (Linner et al 2001). On the other hand, dopaminergic cell firing remains stable if extracellular noradrenaline concentration in the VTA is enhanced (in the alert rat) (Chen & Reith 1994).

In addition to α_1-adrenoreceptors, the VTA also expresses the α_2 family (Lee et al 1998). No direct evidence has yet been established between α_2-adrenoreceptors and VTA dopamine release, however, a number of studies have shown that dopamine cell firing is increased by the administration of the α_2-adrenoreceptor agonist clonidine (Georges & Aston-Jones 2003; Millan et al 2000).

1.3.7.3 Acetylcholine

Cholinergic fibers originating from the peduncolopontine tegmental nucleus and the laterodorsal tegmental nucleus represent an anatomical and functional hallmark by which acetylcholine and dopamine in the VTA interact (Adell & Artigas 2004). The stimulation of the two main classes of acetylcholine receptors, the nicotinic and the muscarinic, contributes to dopamine release within the VTA and participates in reward and locomotor activation (Yeomans et al 2001). Electrophysiological and anatomical evidence confirms that muscarinic and nicotinic receptors can be found in the VTA (Clarke & Pert 1985; Schilstrom et al 2000).

The stimulation of muscarinic and nicotinic receptors in the VTA provokes burst firing of dopaminergic cells in the VTA, nucleus accumbens, and prefrontal cortex (in vitro) (Gronier et al 2000). These increases in dopamine release are greater in the prefrontal cortex than in the nucleus accumbens, in agreement with the anatomical evidence that shows that cholinergic afferents to the VTA are more specific in targeting the mesocortical dopaminergic cells (Garzon et al 1999).

1.3.7.4 GABA

GABAergic cells play an important part in the modulation of dopaminergic activity of VTA neurons (Adell & Artigas 2004). The VTA contains approximately 15-20% GABAergic neurons, which project to the NAcc and PFC, and receive inputs from the NAcc, VP and PPTg (Carr & Sesack 2000). Both $GABA_A$ and $GABA_b$ receptors have been identified within the VTA (Churchill et al 1992). The administration of the $GABA_a$ antagonist bicuculline in the VTA enhances dopamine release, and provides evidence for $GABA_a$ receptors control of dopaminergic activity (in the alert rat) (Adell & Artigas 2004).The $GABA_b$ receptors seem to play a similar inhibitory role. In fact, VTA application of the $GABA_b$ receptor antagonist (CGP

55845A) increases local dopamine release, whereas the agonist baclofen reduces dopamine cell firing (in the alert rat) (Giorgetti et al 2002; Lacey 1993).

1.3.7.5 Glutamate

The VTA receives glutaminergic afferents from the medial prefrontal cortex (mPFC) (Lu et al 1997). The control of dopaminergic activity by glutaminergic neurotransmission is achieved not only through glutaminergic receptors found in the VTA, but also via an indirect prefrontal-VTA glutaminergic pathway (Adell & Artigas 2004). Within the VTA, glutaminergic receptors have been found of both ionotropic and metabotropic nature (Albin et al 1992). The ionotropic receptors are further subdivided into AMPA/kainate and NMDA (Adell & Artigas 2004). The type of modulation by the AMPA/kainate receptors on dopaminergic neurons is twofold: in mesocortical dopaminergic cells, there appear to be tonic excitatory control, whereas in the mesolimbic pathway, the regulation is of phasic nature (in the alert rat) (Takahata & Moghaddam 1998).

NMDA receptors, on the other hand, produce more radical changes to the profile of dopaminergic VTA neurons: the main effect being a transformation from "pacemaker-like activity to a burst firing pattern" (in the alert rat) (Adell & Artigas 2004; Chergui et al 1993). Stimulation of metabotropic receptors, in contrast, increases the firing rate but not the burst firing pattern of VTA dopamine neurons (Zheng & Johnson 2002).

1.3.8 Electrophysiological characteristics of VTA dopamine neurons

The identification that dopaminergic VTA neurons display a specific "neural electrophysiological signature" has, by and large, stemmed from *in vivo* and *in vitro* recordings of cells in the substantia nigra *pars compacta* (SNpc), whereby

approximately 90% of the neurons are dopaminergic (Margolis et al 2006). However, the initial pioneering work was based on indirect in vivo pharmacological manipulations of dopamine neurons in the SN_{pc} and SN_{pr} (substantia nigra pars reticulata) (Guyenet & Aghajanian 1978). Using a dopamine neuron-selective neurotoxin (6-OHDA), the group identified two separate clusters of electrophysiological responses (Fields et al 2007). One cluster of neurons (type 1) showed a slow firing rate with intermittent burst-like activity, wide action potentials, and slow axonal conduction velocity (Fields et al 2007). The other cluster of neurons (type 2) displayed higher firing rates, briefer action potentials, and faster conduction velocity (Guyenet & Aghajanian 1978).

As the medial forebrain bundle (MFB) was lesioned, a smaller proportion of the neurons of the first cluster responded to antidromic stimulation, whereas the second cluster of cell responses remained intact, suggesting that the first cluster of neurons were dopaminergic (Fields et al 2007; Margolis et al 2006). Such suggestion is based on the evidence that the MFB contains ascending and descending fibres, most of which are dopaminergic in nature, that reach the nucleus accumbens and other forebrain regions from their origin in the VTA (Pillolla et al 2007). Therefore, the MFB acts a bridge between NAcc and VTA dopamine release. Additional work based on intracellular recordings in vivo and in vitro confirmed the identity of the two clusters of cells in the SN_{pc} and SN_{pr}; type 1 being dopaminergic, and type 2 GABAergic (Grace & Bunney 1980; 1983; Richards et al 1997). Moreover, an in vitro investigation in the SN_{pc} revealed the presence of another physiological marker, that is, a hyperpolarization-activated current, non specific cation current (Ih) in type1 but not type 2 cells (Lacey et al 1989).

However, adopting the same criteria to identify dopaminergic neurons in the VTA as in the SN_{pc} has proved more controversial. For a start, the proportion of

dopamine versus non-dopamine (GABAergic) cells in the VTA is more equally distributed. Also perhaps, because the cytoarchitectonic boundaries of the VTA are less clearly defined than those of the SNpc (Margolis et al 2006). Studies have shown, for example, that mixtures of dopaminergic and non-dopaminergic neurons within the VTA express the hyperpolarization-activated current (Johnson & North 1992; Jones & Kauer 1999). In addition, differences in action potential duration between dopamine (TH-positive, that is, the enzyme tyrosine hydroxylase acts as a dopaminergic marker) and non dopamine cells (TH-negative) within the VTA in vitro, have proved elusive, despite some reports of a "tendency" for TH-positive cells to have longer action potentials duration (Margolis et al 2006; Ungless et al 2004).

There are also difficulties in identifying dopaminergic VTA cells on the basis of inhibition by D_2 agonist administration, as non-dopaminergic cells appear to be inhibited also, and only a subset of dopamine neurons responds to the inhibitory effects of the D_2 agonist (Margolis et al 2006). Finally, the postsynaptic hyperpolarization of VTA cells by a KOR (kappa opiate receptor) receptor agonist only affects a subset of TH-positive neurons, and the effect is additionally mitigated in vivo by KOP inhibition of "glutamate release onto non-dopaminergic VTA neurons" (Margolis et al 2005; Margolis et al 2006). Therefore, given the absence of reliable physiological and pharmacological criteria to identify dopaminergic neurons, more recent *in vivo* VTA recordings studies tend to include all the detected neurons in the analysis (Margolis et al 2006).

1.3.9 Extracellular characteristics of dopamine action

Whilst it is important to understand the firing pattern of putative VTA dopamine neurons when an animal is executing a given behavioural action, the actual correlation between dopamine and behaviour is complicated by the way in which

dopamine is cleared after its release (Margolis et al 2006). Electrical stimulation by single pulse of the VTA or MFB, in fact, causes long lasting dopamine transients in the NAcc (several seconds) (Phillips et al 2003). A similar effect has been shown to occur after stimulus presentation in rats (Roitman et al 2004). In contrast, the concentration of glutamate or GABA after their release tends not to exceed the 1ms range (Clements 1996). Hence, whilst glutamate and GABA are rapidly removed from the synaptic cleft, dopamine activity appears to be temporally and spatially diffuse (Margolis et al 2006).

This is shown in single cell VTA recording studies, whereby bursts of activity last in the order of 200ms, but voltammetric measurements in the NAcc show dopamine transients lasting several seconds (Pan et al 2005; Phillips et al 2003). Moreover, there are additional factors that complicate our understanding of the steps that follow dopamine release from a burst of action potentials. Namely, the quantity of dopamine release is determined not only by firing rate per se, but also by the neuron's recent firing history (Margolis et al 2006; Montague et al 2004b). In addition, the duration of the dopamine signal is modulated by the density of dopamine transporters; the PFC, for example, with fewer dopamine transporters than the NAcc, displays longer dopamine signals than the NAcc (Cass & Gerhardt 1995).

Finally, the presence of dopamine D_2 autoreceptors in dopamine terminals can in turn modulate dopamine release (Kennedy et al 1992). In summary, it appears clear that there are some difficulties in interpreting the electrophysiological characteristics of VTA neurons as being dopaminergic, but that there are also mitigating factors after affect dopamine release past VTA neuronal firing, which need to be taken into account if one's ultimate aim is to evaluate the relationship between dopaminergic VTA signal and behaviour.

1.4 Functional role of dopamine activity in the midbrain (VTA)

The scope of this section is to first provide a broad overview of the functional role of midbrain dopaminergic neurons in drug addiction, disease, and in adaptive behaviour. The second objective of this section is to introduce four theoretical interpretations of the role that the dopaminergic system plays in reinforcement learning (the prediction error, the anhedonia, the incentive salience and the neuroethological hypotheses). These interpretations are of particular relevance to this thesis, as they will help us put into perspective the empirical findings of Chapter 3 & 4.

1.4.1 The mesolimbic dopamine system as a primary target of drugs of abuse

Midbrain dopaminergic neurons are known to be responsive to psychostimulants, opiates, ethanol, cannabinoids and nicotine.

The three major psychostimulants known to act via dopaminergic mechanisms are cocaine, methamphetamine, and MDMA (Pierce & Kumaresan 2006). Pharmacological, microdialysis, imaging, electrophysiological and lesion studies support this claim. For example, the inhibition of the dopamine transporter (DAT), in humans, is correlated with the subjective positive reports ("high") of intravenous cocaine administration (Volkow et al 1997). Moreover, lesions of the nucleus accumbens (NAcc) produce decrements in cocaine and *d*-amphetamine self-administration responding in rats (Gerrits & Van Ree 1996). Microdialysis investigations in rats and monkey reveal that dopamine extracellular levels in the striatum augment during *d*-amphetamine or cocaine self-administration (Czoty et al 2000; Wise et al 1995).

Within the opiates family, heroin remains the most widely abused substance, despite recent reports of widespread usage of OxyContin (non generic narcotic pain

reliever) (Pierce & Kumaresan 2006). In the animal literature, rats will self-administer opioids into the VTA and the NAcc (Bozarth & Wise 1981; Olds 1982). Some studies suggest that the reinforcing efficacy of opioids self-administration, particularly in the NAcc, is due to the direct inhibition of GABAergic neurons, bypassing dopamine release (Hakan & Henriksen 1989). However, the bulk of the evidence suggests that "μ opioids receptors inhibit GABA release in the VTA", which results in dishinibition of dopaminergic transmission and augmented dopamine release in the NAcc (Pierce & Kumaresan 2006).

With regards to ethanol, its effects on the mesolimbic system are well known. One likely mechanism underlying ethanol administration is an increase in the firing rate of dopamine neurons in the VTA, followed by dopamine release in the NAcc (Bunney et al 2001). The increased firing rate of dopaminergic neurons appears to be mediated by potassium channels and $GABA_A$ receptors located in the VTA (Pierce & Kumaresan 2006). Two lines of evidence support the view of a dopaminergic VTA-NAcc interaction as a leading mechanism of ethanol consumption. Firstly, ethanol is self-administrated into the VTA, and such an effect is impaired by autoreceptors stimulation in the VTA (Rodd et al 2004). Secondly, the administration of ethanol in rats and monkeys augments dopamine release in the NAcc (Bradberry 2002; Weiss et al 1993).

The dopaminergic system has also been linked with the abuse of cannabinoids in humans (mostly through marijuana consumption). In the CNS, cannabinoids receptors (CB_1 in particular) are expressed in small quantities in the VTA and in larger quantities in the NAcc shell (Julian et al 2003; Pickel et al 2004). At a functional level, mice that are devoid of the CB_1 receptor show impaired goal-directed and motivational behaviour (lever pressing for food reward) (Baskfield et al 2004). Similarly to the effects of opioids, cannabinoids augment the firing rate of

dopaminergic neurons by inhibiting GABA release in the VTA (Pierce & Kumaresan 2006).

Finally, there appears to be a strong interaction between nicotine and dopamine particularly within the NAcc and the VTA (Pierce & Kumaresan 2006). In the VTA and in the NAcc, for example, there are nicotinic receptors found on dopaminergic and GABAergic cell bodies, as well as in glutamatergic terminals (Klink et al 2001; Wonnacott et al 2005). Infusions of nicotine in either the VTA or the NAcc shell produces extracellular dopamine release in the NAcc (Ferrari et al 2002; Pontieri et al 1996).In addition, nicotine administration increases the firing rate of dopaminergic neurons, and similarly, nicotinic antagonists in the VTA stop the extracellular dopamine release in the NAcc (Fu et al 2000). Furthermore, lesioning the NAcc impairs the maintenance and acquisition of nicotine self-administration (Corrigall et al 1992).

1.4.2 The dopaminergic system and disease

Midbrain dopamine neurons have also been linked to the aetiology of Parkinson's disease, schizophrenia and depression. The discovery (over 40 years ago) of a prominent loss of dopaminergic cells in the nigrostriatal system has led to the vast majority of Parkinson's patients to be treated with products that aim to restore dopaminergic function, such as Levodopa (Riederer et al 2007). Levodopa is a naturally occurring dopamine precursor, and as such, its administration in conjunction with a DDCI (dopa-decarboxylase inhibitor) to reduce peripheral dopaminergic side effects, became mainstream (Birkmayer & Mentasti 1967). To this date, it remains the "gold-standard" of Parkinson's disease therapy (Mercuri & Bernardi 2005).The main benefit that Levadopa seems to offer over dopamine agonists is the ability to recreate

the regulated release of dopamine by supporting both tonic and phasic bursts of dopamine (Riederer et al 2007).

With regards to the aetiology of schizophrenia, one key finding is that acutely psychotic patients release in great excess striatal dopamine in response to amphetamine administration (Di Forti et al 2007). Based on the "motivational salience" idea that reward associated stimuli are used by organisms for goal-directed behaviour, and that such mechanism is controlled by dopamine striatal release, when dopamine release is excessive (as in the case of schizophrenics), the attention to even the least significant of environmental stimulus assumes disproportionate dimensions leading to the delusions experienced by the psychotic individual (Abi-Dargham et al 2000; Kapur et al 2005). Interestingly, an animal model (in the rat) of schizophrenia has demonstrated that high affinity of the D_2 dopamine receptor as a result of brain impairment is responsible for making an animal more sensitive to changes in its environment (Seeman et al 2005).

Finally, it is also worth mentioning that despite the known involvement of serotonin in depression, a number of studies suggest that the dopaminergic system may play a complementary role (Nestler & Carlezon 2006). Some of these studies, for example, have shown that stress activates dopaminergic neurons in the VTA which in turn send their inputs to the NAcc (Rada et al 2003). More interestingly perhaps, chronic stress appears to activate similar long term adaptations in the VTA-NAcc as those observed after chronic administration of a drug of abuse (Everitt & Wolf 2002; Nestler & Carlezon 2006).

1.4.3 Positive reinforcement in the mesolimbic system

We have so far seen the role that the mesolimbic system plays in drug reinforcement and in disease. However, a number of structures within this pathway

are responsible for producing adaptive responses (maternal behaviour, and attraction), as well as in responding to sensory stimuli (music and humour). This section will briefly review some of the main findings.

In an fMRI study, for example, the ventral striatum was activated when attractive faces looked toward a viewer, but activity decreased when looking away from a viewer, thus suggesting that the ventral striatum may not only encode attractiveness per se, but the interaction between attractiveness and eye gaze (Kampe et al 2001). Similarly, the VTA and the right caudate nucleus have been identifies as playing a role in mate choice selection. Participants who were described as being "in love" were shown pictures of their beloved ones, and fMRI analysis showed that these two brain regions were most active during such presentation (Fisher et al 2005).

In addition, in the rat, there is evidence that the mesolimbic system is recruited in the production of maternal behaviour. In fact, lesions of either the VTA or NAcc disrupt maternal behaviour, and administration of a D_1/D_2 antagonist into the NAcc impairs it (Numan 2007). Interestingly, whilst there is evidence for a role of the mesolimbic system in reproductive behaviour, there are also reports that this pathway is also involved in more cognitive-orientated, typical human attributes such as the hedonic appreciation of music and humour. In fact, a recent study showed that the amygdala, the VTA, the NAcc, and the anterior thalamus, were activated during the presentation of funny cartoons compared to neutral ones (Mobbs et al 2003). Furthermore, listening to subjectively "pleasurable" music also resulted with strong activations of the VTA and the NAcc (Menon & Levitin 2005).

Studies of the NAcc in rodents and minipigs have shown that dopamine is involved in aspects of novelty seeking. A recent study using PET technology showed that increasing dopamine release in the NAcc by amphetamine administration was correlated with an increase in exploratory behaviour of unfamiliar novel objects in

minipigs (Lind et al 2005). Interestingly, the dopamine system supports a case for its role in economic choice behaviour. A recent fMRI investigation using a monetary task, in fact, revealed that the ventral striatum (NAcc in particular), is more activated during a large gain or loss option than during a small gain or loss (Ino et al 2009).

1.4.4 Midbrain dopaminergic responses to rewards, and reward predicting stimuli: the prediction error hypothesis

In order to have a clear understanding of how a group of neurons may contribute to the execution of a given behaviour, it becomes pivotal to possess a precise temporal correlation of spike activity with behaviour (Fields et al 2007). In vivo electrophysiology, and in particular, single-unit recording in monkeys and rats has had a definite influence on our understanding of the function of midbrain dopamine neurons (Pan et al 2005; Schultz 2002; Schultz et al 1997). Dopaminergic neurons in the VTA and in the substantia nigra, for example, respond to primary food and liquid rewards, but also to auditory and visual stimuli that predict reward, and to salient and auditory stimuli (Schultz 2007a). The type of response seen is one of neurons exhibiting burst activity, or phasic activation, characterised by short latency (70-100 ms), short-duration (100-200 ms) and brief intervals (10-50 ms) (Redgrave et al 2007; Schultz 2007a) (but See more in 1.3.5.1).

This is in contrast with the slower or tonic dopamine signal which appears to be associated with movement and cognition deficits in Parkinson's disease (Schultz 2001). The bulk of the evidence suggests that the unexpected presentation of a reward causes activation, whereas a reward that is fully predicted elicits no activation, and finally, omitting a reward at an expected time, causes suppression of activity (Schultz 1999; Schultz 2007a; Tobler et al 2003).

In addition, the extent of the error is reflected in a graded neuronal response, rather than in binary fashion (Schultz 2007b). This discrepancy, between reward

expectation and reward presentation, is otherwise known as “prediction error”, and in the case in which this signal induces a response to reward-predicting CSs it can be called “opportunity gain”. The prediction error or neuronal signature is also sensitive to the timing of the reward, such that delayed rewards produce inhibitions at the expected reward time, and adaptive excitations develop at the new time (Hollerman & Schultz 1998). The neuronal response to conditioned stimuli (CS), however, is unable to distinguish between the spatial position of the reward and the sensory stimulus attributes (Schultz 2007b; Tobler et al 2003; Waelti et al 2001).

In fact, despite being able to discriminate between reward predicting CSs and neutral stimuli, the dopaminergic signal is rather nonspecific to a given reward modality (Redgrave et al 2007). The response is however modulated by the motivation of the animal, and the type of choice amongst rewards (Morris et al 2006; Satoh et al 2003). In addition, during blocking procedures, the neuronal responses are inhibited in acquisition (Waelti et al 2001).

Dopaminergic midbrain neurons also respond to the presentation of novel and physically intense stimuli (Horvitz et al 1997; Ljungberg et al 1992). However, if such stimuli are presented repeatedly without reward presentation, the neuronal responses habituate (Ljungberg et al 1992). Nevertheless, a more recent study showed that phasic responses to stimuli with no reinforcement consequences in the alert animal can be sustained by unpredictable and infrequently presented light flashes or tones (Takikawa et al 2004). It is therefore not entirely clear whether the activations seen during the presentation of novel and/or physically intense stimuli are modulated by their attention catching properties, or by reward-related mechanisms (Redgrave et al 2007).

However, given that other forms of attention inducing stimuli such as conditioned inhibitors, aversive events, and reward omission paradigms produce

much weaker activations, it would appear that dopaminergic neurons assign specific negative-positive values to stimuli presentation rather than encoding simple attentional salience (Mirenowicz & Schultz 1996; Schultz 2007b; Tobler et al 2003).

1.4.4.1 Firing modes of dopaminergic neurons *in vivo*

We have already reviewed some of the electrophysiological properties of VTA neurons, and the intrinsic difficulties there are in identifying them as being dopaminergic (see 1.3.8 and 1.3.9). This section, assumes that the cells discussed are dopaminergic, and compares and contrasts their firing mode in freely moving versus anaesthetised rats, as well as reporting the firing mode of dopamine cells in task related activities (in the freely moving rat only). Therefore, this section investigates the firing mode of dopaminergic cells under three conditions: 1) The burst-like activity of dopamine cells in the absence of a behavioural task 2) The more regular "clock-firing" pattern in the absence of a behavioural task. 3) The phasic pattern in task related activities.

1) A substantial number of studies in anesthetised animals report that the vast majority of dopamine cells in the absence of specific stimuli do not respond in bursts (Freeman et al 1989; Grace & Onn 1989). A recent investigation in freely moving rats has also shown that only 21% of dopamine cell spikes fell into bursts (as characterised by three spikes bursts per 500 spikes recorded) (Hyland et al 2002). This proportion is therefore similar to that reported in animals anesthetised with chloral hydrate (in the absence of specific stimuli), but lower when other anaesthetics are used (Schultz & Romo 1987; Tepper et al 1995). There were, however, other differences. Intraburst frequencies in the freely moving rat, for example, were reported to be higher in the freely moving than the anesthetised rat (Hyland et al 2002). In addition, in the

anesthetised preparations, there have been reports of increased interspike intervals during bursts (Grace & Bunney 1984).

2) A number of studies have shown that the regular, "clock-like" firing of dopaminergic cells is typically observed in *in vitro* preparations (Grace & Onn 1989; Hyland et al 2002). On the other hand, in both the anaesthetised and in the freely moving rat, such clock-like regular firing is rarely observed. However, it has been pointed out that such discrepancy between the *in vivo* and *in vitro* results may be due to the type of statistical tool used. In fact, when auto-correlation analysis is used, this difference may be smaller than believed (Hyland et al 2002). Studies have shown that when this method of analysis is adopted, there are reports of dopaminergic cells showing "clock-like" firing in paralysed and alert rats also (Hyland et al 2002; Wilson et al 1977).

3) In task-related activities, the phasic related bursting is characterised by reduced intraburst intervals, and hence, higher intraburst frequencies. Moreover, those cells that respond in phasic manner during reward delivery, display similar electrophysiological properties to the cells that are recorded in the absence of stimulus presentation, if observed outside reward delivery (Hyland et al 2002).

1.4.5 The role of the dopaminergic system: beyond the prediction error hypothesis

There are additional theoretical interpretations to the role that dopaminergic neurons play in reward other than the prediction error hypothesis, which very much views dopamine to be linked with reward learning. The (an)hedonia hypothesis developed by Wise, the incentive salience hypothesis developed by Berridge, and more recently, the neuroethological perspective proposed by Alcaro (Alcaro et al 2007; Berridge & Robinson 1998; Wise et al 1978).

1.4.5.1 The anhedonia hypothesis

One of the first roles attached to dopamine in reward was identified by Wise, which proposed that the dopamine system modulates the pleasure for food, sex, drugs of abuse, and other unconditioned incentives (Wise & Bozarth 1985). This conclusion was based on the discovery that after the administration of a neuroleptic dopamine blocker (Pimozide) in the hungry rat, there was attenuated lever pressing for food reward, leading Wise to conclude that the neuroleptic had induced a state of anhedonia in the rat (Wise et al 1978). Following Wise's findings, a number of other studies have reported that dopaminergic modulation is responsible for creating an hedonic imbalance, and that during drug withdrawal, organisms seek to re- establish such imbalance (Dackis & Gold 1985; Koob et al 1997). Moreover, other studies have instead identified the dysregulation of dopamine neurotransmission as "blunting of reactivity in the NAcc" (Di Chiara & Tanda 1997). Finally, neuroimaging investigations have correlated the level of positing liking (drug pleasure ratings) with dopamine receptor occupancy in the NAcc (Small et al 2003).

There is, however, plenty of evidence which goes against the anhedonia hypothesis. Firstly, microdialysis and electrophysiological studies show that the dopamine system is often activated before the presentation of the reinforcer, and that such activation is stronger to a conditioned stimulus predictive of reward than to reward consumption per se (Hernandez & Hoebel 1988; Martel & Fantino 1996a; b; Schultz et al 1997; Tobler et al 2003). Secondly, a number of taste reactivity studies show that blockade of dopamine transmission does not produce disliking of palatable foods, and that enhancement of dopamine transmission does not induce increased natural reward impact in the rat (Kaczmarek & Kiefer 2000; Pecina et al 1997). Thirdly, lesions of the NAcc in the rat result in increased breaking points in progressive-ratio responding schedules. These data suggest that lesions of the NAcc

impair the ability to judge the increasing cost of reward (Bowman & Brown 1998). Finally, in humans, subjective ratings have been found to be most strongly correlated with "wanting a drug" than liking a drug (Leyton et al 2002).

1.4.5.2 The incentive salience hypothesis

Berridge proposed that the dopaminergic system may not be involved with the consummatory phase (or liking) *per se*, but with the seeking aspect (or wanting) of reinforcement (Berridge & Robinson 1998). Pharmacological investigations, in fact, show that blockade of dopamine activity in the NAcc, leave the consummatory phase intact, but impair maze-running speed (Ikemoto & Panksepp 1996). In addition, as we have gathered from taste reactivity studies, dopamine appears not to be required for "liking" of palatable foods, but for increased "wanting" for sweet taste (Pecina et al 2003). Similarly, increased dopamine neurotransmission through amphetamine administration increases approach behaviour to stimuli previously associated with reward (Wyvell & Berridge 2000; 2001). Incentive salience, therefore, is best described as "conditioned motivation triggered by and assigned to a reward-related stimulus" (Berridge 2007).

A number of reports on feeding behaviour, however, challenge this view, for example, as particular pleasant/unpleasant tastes activate dopamine neurotransmission (Roitman et al 2005). It has also been argued that, in order for a stimulus to achieve increased salience, a motivational component must in turn be supported by the ability to learn the stimulus-related contingencies (Alcaro et al 2007).

1.4.5.3 The neuroethological perspective

Berridge's proposal specifically focused on motivational properties that are activated by external stimuli and that in turn drive behaviour. Therefore, the environment is seen as the guiding force of behaviour, rather than the affective states

or internal drives of an organism (Alcaro et al 2007). The neuroethological perspective, on the other hand, attributes a secondary importance to the effects of external stimuli on driving behaviour, whilst putting internal drives (or seeking disposition) as the true motivational engine of an organism. Such seeking disposition is argued to be a set of instinctual behavioural tendencies which encompass locomotion, orienting movements, sniffing, and vocalisations in rats (Burgdorf & Panksepp 2006). Indeed, there is experimental evidence that, for example, exploratory behaviour often precedes bar pressing for electrical stimulation (Ikemoto & Panksepp 1996).

Consequently, the strength of the conditioning is determined by the extent to which an unconditioned stimulus can arouse the seeking disposition (Alcaro et al 2007). Once the seeking disposition is aroused, a process of associative learning similar to that described by prediction error hypotheses is believed to drive behaviour; facilitated by memory consolidation in hypothalamic areas and through learning mechanisms in the dopaminergic midbrain (Alcaro et al 2007; Cahill 1997).

1.5 Aim of the present thesis

Following the demonstrations of an involvement of dopaminergic neurons in reward prediction in Pavlovian and simple operant conditioning tasks, it has become necessary to elucidate how such neurons would respond under more complex conditions in which sequences of events are predictors of reward.

The aim of my thesis, therefore, is to refine the current neural network models of adaptive behaviour that are inspired by the biology of reinforcement learning. Although most recent research on reinforcement learning is targeted at increasing our understanding of addiction, reinforcement learning is also important for the development of computational algorithms for machine learning. In this regard reinforcement learning is a useful algorithm because it is unsupervised, relying on trial-and-error learning under conditions in which the optimal solution is unknown. At a more fundamental level, an understanding of reinforcement learning is also important for our basic scientific understanding of habit formation, decision-making and microeconomics (e.g., see (Egelman et al 1998)).

Currently, reinforcement learning models have been validated in minimalist environments in which only 1-2 environmental stimuli are present as possible predictors of reward. In the first stage of the project we will test current models of reinforcement learning under a configuration of multiple stimuli in which sequences of events are predictors of reward. An example of this type of situation is "occasion setting" in which reward is contingent on a given stimulus (the conditioned stimulus in Pavlovian conditioning, a discriminative stimulus in instrumental conditioning; the following example will refer to the simplest case of Pavlovian conditioning) only when the stimulus has been preceded by another stimulus (the occasion setter) (see Fig 3, 1.1.2.9). When the occasion setter is present the reinforcer reliably follows the conditioned stimulus and when it is absent no reinforcement is delivered after the

conditioned stimulus. Thus, the presence of an occasion setter prepares an animal for responding and indicates that the relationship between a conditioned stimulus and reward is in effect.

To date there have been only two studies on the responses of the dopamine system to configurations of multiple stimuli, and in both cases the stimuli were presented simultaneously rather than in a sequence. This occurred in the case of blocking and a conditioned inhibition paradigm (Balleine & Dickinson 2006; Tobler et al 2003). By observing the responses of dopamine neurons to conditioned stimuli in the presence and absence of an occasion setter, we will be able to answer three questions: First, are the responses of dopamine neurons to conditioned stimuli "gated" by the presence of the occasion setter? Second, does the degree of dopaminergic responses to conditioned stimuli systematically vary between the behavioural conditions (absence/presence of occasion setter) predictive of whether rats exhibit conditioned behavioural responses? The answer to this question will allow us to infer the extent to which the dopamine system is involved in response preparation *per se versus* driving the neural plasticity necessary for reinforcement learning. Third, will the dopamine neurons respond to the occasion setter in the same way as they respond to conditioned stimuli? The temporal difference models of the role dopamine plays in reinforcement learning respond in a similar way to the occasion setter as they do to the conditioned stimulus, but in most circumstances animals do not exhibit conditioned behavioural responses to occasion setters.

The first step will be to use the neural network simulation of reinforcement learning developed by Montague, Dayan & Sejnowski (Montague et al 1996) modified to take into account recent neurophysiological work that provides more biologically realistic estimates of trace eligibility (Pan et al 2005). The reinforcement-learning model will be presented with sequences of events that predict reward and

control trials in which single events are presented without consequent reward. This will allow us to determine whether the current reinforcement models are able to learn the sequences that are predictors of reward. The second part of Stage 1 will be to test the predictions of the model compared to neurophysiological responses of dopamine neurons in alert rats performing behavioural tasks that are functionally similar (instrumental versus Pavlovian conditioning) to the ones presented to the neural network model. Once the empirical data are collected they will be compared to the model to determine if there are discrepancies.

In the second stage of the project, we will test the model under another complex condition of events, known as "overexpectation". The first phase of the overexpectation training includes presenting two conditioned stimuli (CS_A, CS_B) independently and pairing them with an unconditioned stimulus (US). In the second phase, CS_A, and CS_B are simultaneously presented and paired with the US. Previous studies have shown that, following the CS_{AB} compound, responding to CS_A or CS_B is reduced relative to a control condition where no compound training has occurred (Dawson & Spetch 2005; Khallad & Moore 1996; McNally et al 2004). The effect, whilst counter-intuitive, is predicted by the Rescorla-Wagner model of associative learning, and provides strong support for the idea that learning is dictated by the difference between anticipated and obtained reinforcement (Rescorla & Wagner 1972).

To date, there have been no studies matching the responses of dopamine neurons in the "overexpectation" paradigm, to the predictions of temporal difference learning models. Therefore, by using the same neural network simulation adopted for the occasion setting experiment, and by performing extracellular single-cell recording in vivo, we will be able to answer two fundamental questions: First, do dopaminergic neurons respond in a reduced manner to CS_A, CS_B, stimuli, following their compound

presentation? Second, are dopaminergic responses to CS_A, CS_B, and CS_{AB} presentations dependent upon the balance of reinforced trials that each stimulus receive (e.g. A+ 40%, B+ 40%, AB+ 10%, A+ 5%, B+ 5% ; or A+ 20%, B+ 20%, AB+ 40%, A+ 10%, B+ 10%)? In order to answer these questions, the reinforcement learning model will be presented with a sequence of events that is functionally identical to the behavioural task presented to alert rats, whilst neurophysiological recording of dopaminergic cells is taking place.

Once the empirical data are collected they will be compared to the model to determine if there are discrepancies. Overall, we will use these two behavioural tasks and neurophysiological recording combined with neural network simulations to unravel the role that midbrain dopamine neurons play in reinforcement learning under conditions in which multiple stimuli are present as potential predictors of a reinforcer.

2 Chapter 2. Modelling the occasion setting and the overexpectation effect

2.1 Abstract

Dopaminergic neurons' responses in the primate and rat midbrain are well described by temporal difference models (TD) of reinforcement learning. In particular, TD models have been able to reproduce the activity of dopamine neurons in a number of contexts that include: the presentation of an unpredicted reward, in delayed reward paradigms, with novel, physically salient stimuli, etc (Suri & Schultz 1999). However, whilst TD models have been validated in behavioural paradigms in which only 1-2 environmental stimuli are present as possible predictors of reward, there is no knowledge to date about the responses of the dopamine system to configurations of multiple stimuli presented in a sequence. Thus, there is an underlying interest in understanding how TD models may respond to a sequence of events. Here we report the results of TD simulations under the behavioural paradigms of occasion setting and overexpectation. The data for the occasion setting simulations show that dopamine would produce the greatest firing to the earliest predictor of reward, the occasion setter. The output of the model also suggests, however, that the occasion setter would not facilitate responding to the conditioned stimulus.

TD simulations of the overexpectation effect suggest that responses to the combined conditioned stimuli (CS1 & CS2) would be greater than to the individual CS's (CS1 or CS2). In addition, the output of the model shows that inhibition would occur when a single reward was delivered after the combined CS. The results of the model are interpreted in light of the neuronal and behavioural data previously acquired.

2.2 Introduction

In chapter 1, (see 1.2.2.6 and 1.2.2.7) we reviewed two temporal difference methods that have been central to the construction of this thesis (Montague et al 1996; Pan et al 2005). Here, we will specifically discuss the significance of four of the parameters of the models, as modifying these can have a direct impact on simulations of dopaminergic activity. The aim of the simulations reported herein is in turn to systematically vary those four parameters (over a factor of 10), whilst the model is presented with a sequence of events that is functionally similar to the behavioural experiments of occasion setting and overexpectation (Chapter 3 and 4). Varying the parameters allows us to create variability in the output data of the model, based on empirically verified parameters values (Montague et al 1996, Pan et al 2005). Whilst in computational neuroscience there is no standard number of simulations that can be ran (i.e. varying the parameters over a factor of 10, 100, or 1000), the results of 6250 simulations will tell us whether there is any indication that the range of the parameters values chosen needs further exploring (if for example in some of the simulations, the relative strength of the dopaminergic output to one of the stimuli is greater than that to the other). Finally, the output data of the model will allow us to make direct comparison with the behavioural and neurophysiological data collected.

Montague's model (1996) was an attempt to build a theoretical framework that could demonstrate how fluctuations of activity in the dopaminergic system would represent a reward prediction error signal that is then broadcasted to cortical and subcortical targets, and that would in turn drive adaptive behaviour. The model was, therefore, an attempt to reconcile the physiological recordings in alert monkeys and rats that show that midbrain dopaminergic neurons respond to unexpected reward delivery, conditioned stimuli, and novel stimuli (see more 1.4.4) (Fiorillo et al 2008; Hyland et al 2002; Schultz 1998).

Briefly, the model was envisioned as made up of three parts: a cortical region, an intermediate layer, and a subset of dopamine neurons (for illustrative purposes, see 1.2.2.6). Dopamine neurons receive convergent input from either cortical region or from rewarding/salient events in the environment and within the organism. The input coming from the cortical region is first passed through an intermediate layer. Each input from the cortical region is then associated with weights that in turn determine the strength of its influence on the dopamine neurons.

The intermediate layer is conceived as an attentional filter where weight changes take place as a function of experience. The output of the dopaminergic neurons is conceived as the sum of its net excitatory and inhibitory input plus basal activity. Crucially, the changing output of dopaminergic neurons is based on the comparison between expected and actual reward. A difference between actual and predicted reward, generates a prediction error. A positive prediction error would represent that the current state is better than expected (increase in dopamine) and a negative prediction error would represent that the current state is worse than expected (decrease in dopamine) (Montague et al 1996). The predictions (or expectations) that occur are displayed in the way in which weight changes develop.

The four parameters that we know affect the model output are the learning rate, the temporal discount rate, the eligibility trace decay, and the dopamine inhibition parameter. The last two parameters have been incorporated explicitly in Pan's but not Montague's model (Pan et al 2005). The learning rate determines the magnitude of weight changes by the prediction error signal. In other words, the rate dictates the speed in which prediction error signals to conditioned stimuli and rewards develop or are lost (Pan et al 2005). So far, a good match of dopaminergic activity has been achieved using low learning rates (0.05-0.3) (Montague et al 1996; Pan et al 2005; Suri & Schultz 1999). From a purely computational point of view, learning rate

refers to how fast a network is trained. Studies in machine learning have shown that a low learning rate increases the stability of learning (Singh & Sutton 1996; Tesauro 1992).

The temporal discount rate, on the other hand, refers to the decreasingly motivational value of delayed rewards (Suri & Schultz 1999). The best match for the neurophysiology data has been achieved using a low discount rate (0.98) (Montague et al 1996; Pan et al 2005). Using various stimulus-reward intervals, in fact, it has been shown that the dopaminergic signal decreases by 20% per second (or 2% per 100ms of a timestep) (Suri & Schultz 1999).

The eligibility trace parameter (see 1.2.2.7) allows the modification of previous weights by the prediction error signal. Setting a low trace value means that weight changes take place only for events that have mostly recently occurred, and vice versa. In other words, the interpretation of a system based on a low eligibility trace value is that it can only hold the memory of an event from one moment to the next (Pan et al 2005). On the other hand, a higher value signifies that the system is capable of bridging events that are far removed in time. A number of hypotheses have been advanced regarding the biological substrate of long eligibility traces, amongst these, sustained firing in the striatum and prolonged changes in calcium concentration (Calabresi et al 1992; Calabresi et al 1997). However, evidence for this is scarce (Suri & Schultz 1998). Nevertheless, whatever their biological underpinning may be, studies by Pan and Suri have shown that setting long eligibility traces (0.9 and 0.96) allows to better mimic the dopaminergic response to conditioned stimuli (Pan et al 2005; Suri & Schultz 1998; 1999). Namely, that responses to predicted rewards are retained even after responses to conditioned stimuli have developed. This is in contrast with a gradual stepwise migration of responses, as predicted by previous temporal difference models (Montague et al 1996; Schultz et al 1997).

The fourth and final parameter is the dopamine inhibition value. This parameter was introduced by Pan *et al* (2005), and it refers to the extent that dopaminergic cells are asymmetrical in their positive and negative prediction error signal amplitudes. More specifically, the baseline spiking activity of dopaminergic cells (~5Hz) can be equated to a prediction error state of 0. The maximum dopaminergic inhibition that is possible is equal to (~0Hz). However, when dopamine cells are excited they can reach values of approximately (~100Hz). Therefore, as the positive and negative amplitudes are asymmetrical, Pan *et al* (2005) set a limit (-0.05) to the negative prediction error so that the range of dopaminergic activity was scaled with the range of positive and negative prediction error amplitude. The result was an improvement in matching the small inhibition of dopamine neurons by limiting the amplitude of the negative prediction error.

The aim of these simulations, therefore, is to use Pan's *et al* (2005) parameters' values that best fit into their neurophysiological findings, and to vary these values systematically over a range of 10x. The model will then be presented with a sequence of events that is functionally similar to the behavioural experiments of chapter 3 and 4. The results of the model will allow us to compare and contrast the neurophysiological and behavioural findings that we previously acquired.

2.3 Methods

2.3.1 Creating and running simulations

We ran simulations using Pan et al (2005) values for the learning rate, temporal discount, trace decay, and dopamine inhibition parameter, and we varied these original values over a factor of 10.

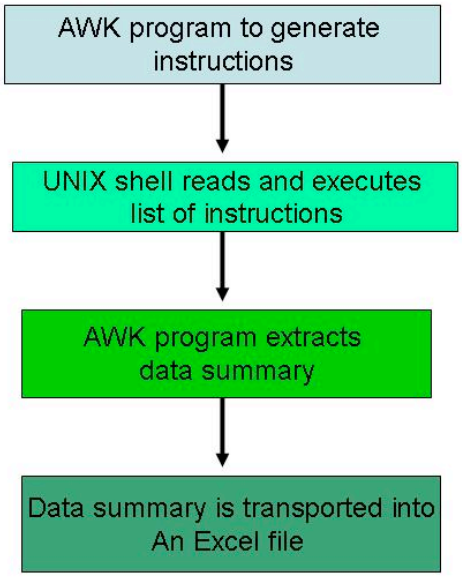
Learning rate	Temporal discount	Trace decay	DA inhibition
0.0174	0.0063	-0.0333	-0.0158
0.0309	0.0112	-0.0593	-0.0281
0.0550	0.0200	-0.1054	-0.0500
0.0978	0.0356	-0.1874	-0.0899
0.1739	0.0632	-0.3333	-0.1581

Table1: The row highlighted in yellow shows the original Pan et al (2005) values for the learning rate, the temporal discount, the trace decay, and the dopamine inhibition parameter. The first row and the last row of each column show how each parameter was varied by a factor of 10 (i.e. 0.0174-0.1739 for learning rate etc…). This produced a total of 625 combinations of parameters for which 10 simulations were run for each task (occasion setting and overexpectation).

The TD algorithms of Pan et al (2005) were then implemented in a RealBasic (REAL Software, Inc., Austin, USA) written by Dr Eric Bowman. An AWK program was written to create a macro file with instructions for running 10 simulations of each of the 625 combinations of parameters. In the training stage of each simulation the order of trial types was pseudo randomly varied, creating variability in the output from simulation to simulation. Finally, a UNIX shell program in which an AWK program was embedded extracted the relevant results and put them in to a tab-delimited file that served as a database of the results of the simulations.

For changes in dopamine activity evoked by sensory stimuli (occasion setters & conditioned stimuli), the dopamine output of the model was summed over the first three ticks (a tick is a symbolic unit of time: 100ms) of the simulated stimulus.

Responses to the reinforcers were calculated as the output for the single tick in which the simulated reinforcer was delivered. A schematic illustration of how the TD model was implemented is shown here below (all scripts and data files can be found in the appendix).



2.3.2 Occasion setting simulation task

Figure 1: We modelled the occasion setting procedure in Pavlovian manner, as opposed to the operant task in the behavioural paradigm. The OS was presented between tick 4 and tick 17. CS presentation occurred between tick 12 and 17. The US presentation occurred at tick 20. In the condition where reward was delivered, the OS and the CS overlapped. There were 1500 training trials. 500 of which (OS alone), 500 (CS alone) and 500 (OS-CS-US). The probe trials were identical in nature as the training trials, except that they include 1 trial only per condition.

2.3.3 Overexpectation simulation task

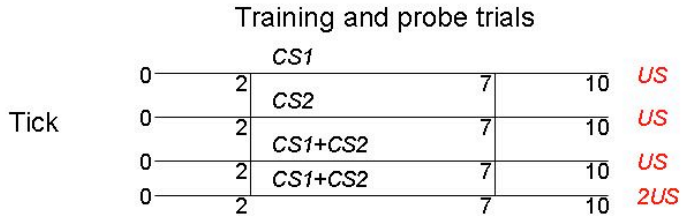


Figure 2: CS1 and CS2 were presented between tick 2 and tick 7 in both training and probe trials. US presentation always occurred at tick 10. In the training procedure, there were 1500 trials. 500 of which the CS1 is followed by reward (US). 500 of which CS2 is followed by reward. 500 of which CS1+CS2 is followed by reward. In the testing or probe phase, there was 1 trial only for each condition. Each condition was the same as in the training procedure expect for doubling of reward (2US) in one of the probe trials.

2.4 Results

2.4.1 Occasion setting simulations

We compared the output of the model with regards to the relative response to the OS versus the CS alone. Thus, for each simulation, the dopaminergic output of the OS is divided by the response to the OS plus that to the CS (OS/ (OS+CS)). For example, if the model had produced an output of 1 for OS and 1.5 for CS, this would have equalled to: (1/ (1+1.5) = 0.4). That is, a preferential responding to the CS alone. In all 6250 simulations, we found preferential responding to the OS versus the CS alone.

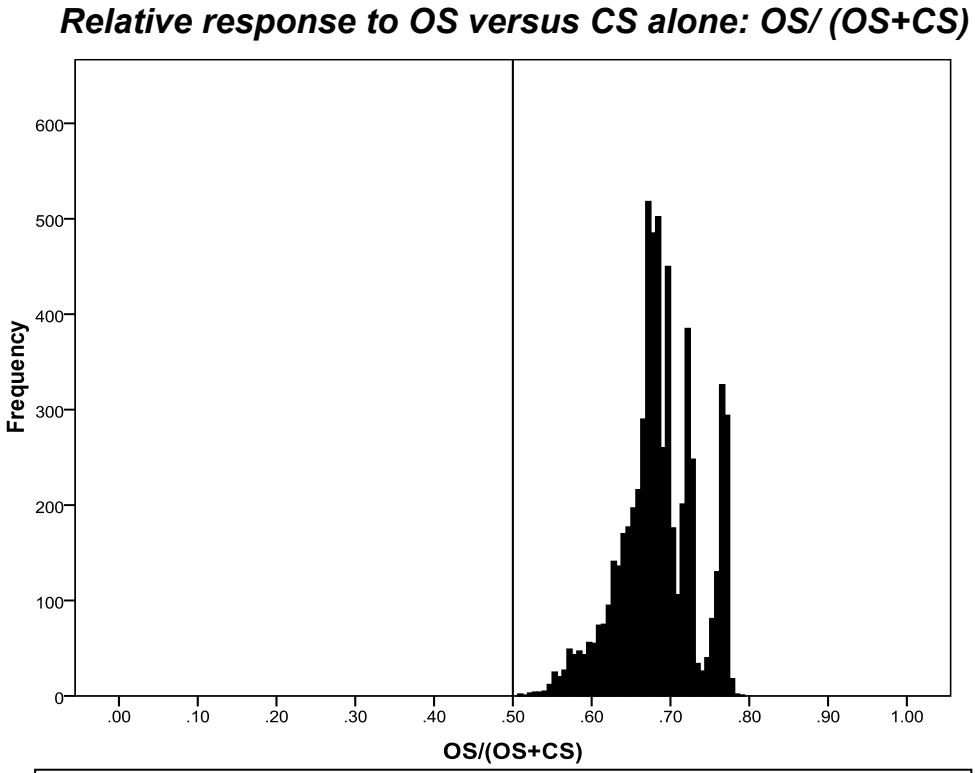
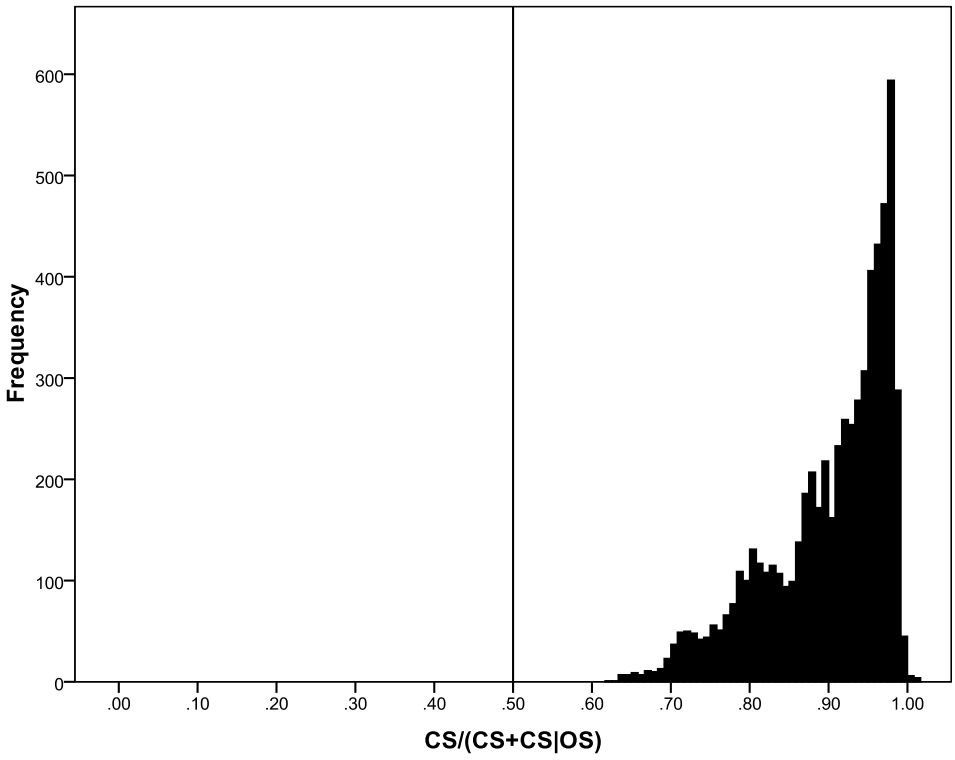


Figure 3: The results of 6250 simulations of dopaminergic responses to OS versus CS after training. Values on the horizontal axis are based on equation: $Response_{OS}/(Response_{OS}+Response_{CS})$:
Responses on the .50 mark indicate equal responding to OS and CS. Responses lying between .00 and 0.49 indicate preferential responding to CS (with .00 the maximum value), and responses lying between .51 and 1.00 indicate stronger responding to OS (with 1.00 the maximum value). In all simulations responses to the OS were greater than to the CS.

In addition to looking at the relative response to the OS versus the CS alone, we investigated whether the presence/absence of the OS modulated the strength of the response to the CS. Thus, for each simulation, the dopaminergic output of the CS is divided by the response to the CS plus that to the CS given the presence of the OS (CS/ (CS+CS | OS)). For example, if the model had produced an output of 1 for CS and 1.5 for CS given the presence of the OS, this would have equalled to: (1/ (1+1.5) =0.4)). That is, a preferential responding to the CS preceded by the OS. In all 6250 simulations we found preferential responding to the CS alone.

Relative response to CS alone versus CS after OS: CS/(CS+CS|OS)



.

Figure 4: The results of 6250 simulations of dopaminergic responses to CS alone versus CS after OS. Based on equation CS/(CS+CS|OS) responses on the .50 mark indicate equal responding to CS alone versus CS after OS. Responses lying between .00 and 0.49 indicate preferential responding to CS after OS (with .00 the maximum value) and responses lying between .51 and 1.00 indicate preferential responding to CS alone (with 1.00 the maximum value). In all simulations, responses to the CS alone were greater than to the CS preceded by the OS.

Finally, we wanted to find out whether there was any relationship between the response to the OS as a function of the response to the CS alone. Using a Spearman's rank correlation, we demonstrated that across all 6250 simulations, the stronger the response to the CS, the stronger the response to the OS

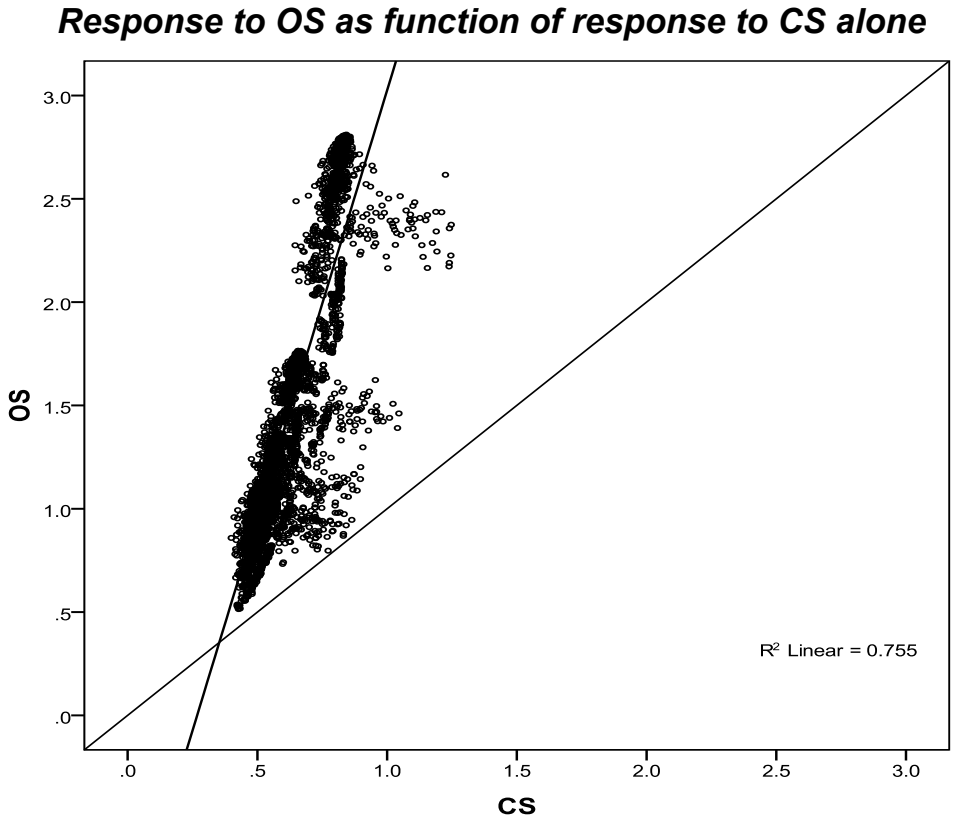


Figure 5: Spearman's rank correlation between strength of CS and OS responses. ($r_{(6249)}$=0.871, p<0.001). The simulations predict that the stronger the response to the CS, the stronger the response to the OS. Although a regression line is used to illustrate the trend, a rank correlation is calculated because the univariate distributions for the responses to the OS and CS were multimodal.

2.4.2 Summary of occasion setting simulations

The main findings of the occasion setting simulations are: 1) All simulations (n=6250) have shown that there was preferential responding to the OS versus CS alone (Figure 3). 2) All simulations (n=6250) have shown that there was preferential responding to the CS alone versus CS after the OS (Figure 4). 3) There was a positive correlation between CS and OS strength of responding (Figure 5).

2.4.3 Overexpectation simulations

We plotted the response to the individual CS's (average of CS1 & CS2) versus the response to the compound stimuli (CS1 & CS2 presented together). For all 6250 simulations, there was a clear additive effect of the two stimuli being presented together. These results are a clear consequence of the TD model architecture based on the Rescorla-Wagner learning rule, which derives the associative strength of the compound stimuli (CS1 & CS2 presented together), from the sum of the associative strength of CS1 and CS2 presented individually.

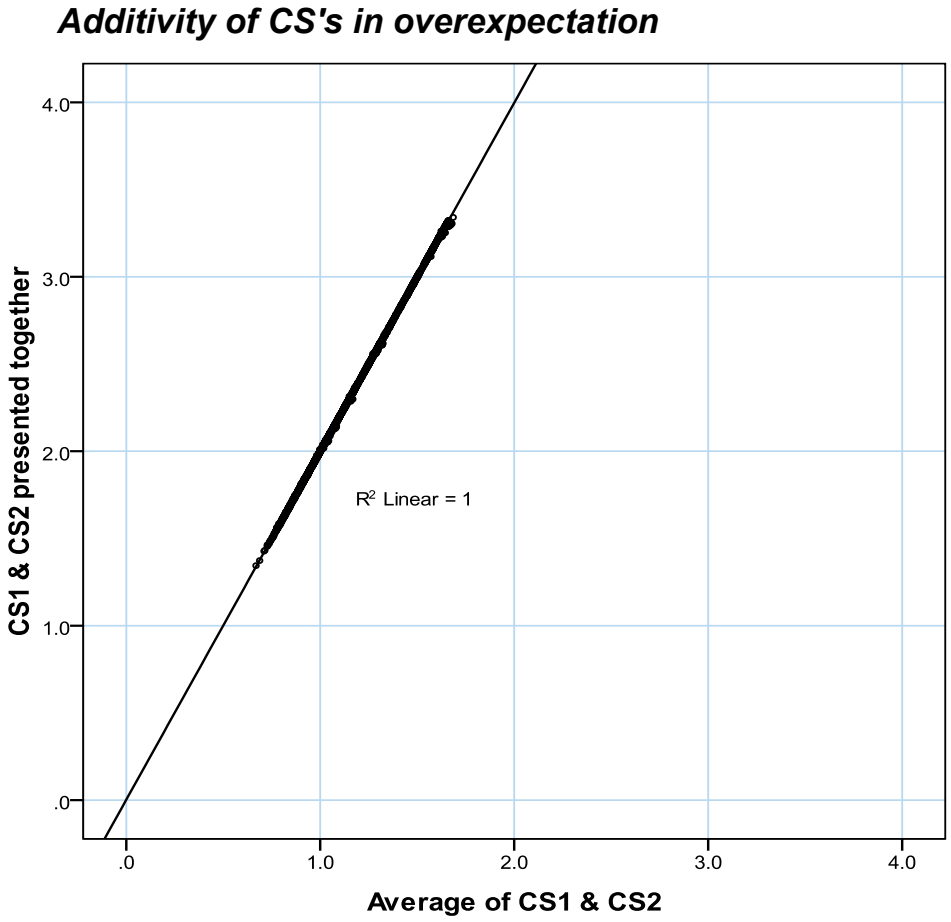


Figure 6: Dopaminergic response displaying the average of CS1 and CS2 versus CS1 & CS2 presented together. R^2 =1 ($F_{(1,6249)}$ =1.116, p <0.001). The additivity of the responses arises from the architecture of the model, which results in elemental rather than configural learning.

The first requirement for the overexpectation effect to occur is that there is summation when CS1 & CS2 are presented together (Figure 4). However, the other requirement is a mismatch between the reward predicted and that obtained during the compound trial. Thus, we plotted the response to a double versus single reward after the combined CS presentation (CS1 & CS2 presented together). The results demonstrate that the model is sensitive to changes in reward magnitude, and the regression analysis clearly shows that excitation to the double reinforcer was 2x the inhibition to the single reinforcer (a slope of approximately -2).

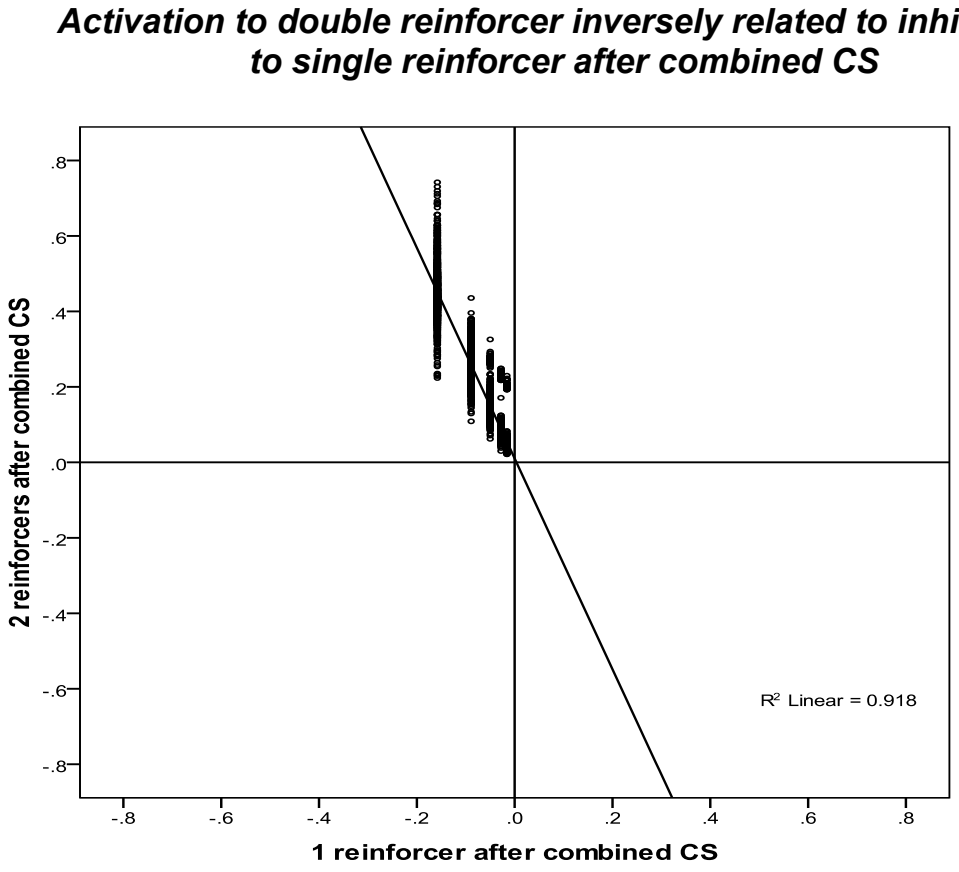


Figure 7: Dopaminergic response to single reward (low reward quantity) after combined CS presentation (1 reinforcer after combined CS) versus double reward (high reward quantity) after combined CS (2 reinforcers after combined CS). The simulated dopamine activity evoked by the single reinforcer was inhibited (left half of graph), indicating less reward than expected, while the simulated response to the double reinforcer was excited (top half), indicating greater than expected reward. The simulations show that the greater the inhibition to the single reward, the greater the excitation to the double reward (R^2=.918 ($F_{(1\ 6249)}$=69496, p=<0.001).

2.4.4 Summary of overexpectation simulations

The main findings of the overexpectation simulations are:

1) The model reacted equally to both CS1 and CS2.

2) The model reacted twice as strongly to the combined CS as to the individual CS's.

3) The model was excited when double reward was delivered after the combined CS.

4) The model was inhibited when a single reward was delivered after the combined CS.

5) The more inhibition to the single reward, the greater excitation to the double reward.

2.5 Discussion

The first finding worth discussing is that the results of the simulations have clearly shown that varying the models parameters (by a factor of 10) has had very little impact on the qualitative output of the model. This was the case for both the occasion setting and the overexpectation simulations. We will start by summarising the findings for the occasion setting simulations, and then those of the overexpectation simulations. A note, however, should first be added regarding the way in which the model and the behavioural experiment differed. To begin with, we ran simulations of the occasion setting in a Pavlovian fashion, as opposed to the operant conditioning nature of the behavioural experiment. Hence we here refer to CS, rather than $S_{D.}$ There are two main reasons for this. Firstly, we did this so as to contain the computational costs of implementing an instrumental procedure, but keeping intact the temporal representation of the configuration of stimuli (the sequence OS→S_D in our behavioural neurophysiology experiments is functionally similar to the OS→CS sequence modelled here). Secondly, because it has been shown empirically that the evoked dopamine response in operant conditioning tasks is very similar to that seen in Pavlovian tasks (Schultz 1993). The other difference between the model and the behavioural paradigm is that in the occasion setting model we did not include a third stimulus indicating immediate reward availability (a reward signal, referred to as the *CS* in chapter 3). This is because in preliminary simulations, we found that adding a third stimulus amounted to no qualitative difference in the output data of the model.

The results of the simulation can be summarised as follow. Firstly, the model predicted that in all cases (n=6250), there would be a greater response to the OS alone than to CS alone. That is, the model predicted a gradual shift in response to the earliest predictor of reward, the OS. Secondly, the model predicted that in all

simulations, there would be greater responding to the CS alone than to the CS preceded by the OS. To sum up, varying the model parameters had little impact on shifting responses from the OS to CS, or enhancing responses to the CS in the presence of the OS.

With regards to the overexpectation simulations, we also found that varying the parameters had little impact on the qualitative output of the model. The results of the simulations can be summarised as follows. Firstly, we found that the model reacted approximately equally to both CS_1 and CS_2. Secondly, we found that the model reacted twice as strongly to the combined CS presentation (CS_1+CS_2) as to the individual CS's (see 2.4.3). The model also predicted excitation when double reward was delivered after the combined CS (reflecting a "better than expected" signal of the model"), and inhibition when a single reward was delivered after the combined CS (reflecting a "worse than expected" signal of the model) (see 2.4.3).

Overall, by using Pan's baseline parameter settings, and varying them by a factor of 10, there was very little impact on the qualitative output of the model in either behavioural circumstance. This increased our confidence that the hypotheses derived from considering the properties of the TD model would be worth testing. The next two chapters summarise our attempt to do this.

3 Chapter 3. Neural responses in the dopaminergic midbrain to occasion setters

3.1 Abstract

Midbrain dopamine (DA) neurons respond to a variety of reinforcers (water, food, sex and other affiliative social behaviours, addictive drugs, and intracranial self-stimulation). DA neurons are also activated by conditioned stimuli that predict primary reinforcers, which has led to the inference that the DA system encodes reward expectation and that plasticity in the DA system minimizes reward prediction error. Computational models of reinforcement learning predict that the DA system should respond to the earliest reliable predictor of reward, provided the representation of the stimulus has not faded (Montague et al 1996, also see the first series of simulations in Chapter 2).

Thus the issue of how the system responds to static, unchanging stimuli that indicate reward predictability is of interest. One class of such stimuli are occasion setters (OS's), which are contextual sensory stimuli that signal the contingency between another stimulus and reinforcement is in effect. We recorded from DA neurons while thirsty rats (*n*= 6) performed a bar pressing task for liquid reinforcement. Bar-pressing was reinforced only when an OS (tone or houselight) and a discriminative stimulus (S_D; houselight or tone) were presented simultaneously. The OS was presented alone for 10 sec and then followed by an overlapping S_D that was presented for 30 sec, during which each bar press was followed by a reward signal (reward magazine light) and availability of a small volume (0.05 ml) of saccharin solution.

Bar-pressing was not reinforced in control trials in which the OS was presented alone, the S_D was presented alone or no stimuli were presented. Data from 6 rats in recording sessions indicate that dopaminergic cells responded most strongly to

the most proximal stimulus (the reward signal) and not the earliest predictor of reward (the OS). Analysis of data from 45 neurons in the dopaminergic midbrain indicated that there was weak neural enhancement to the S_D when it was preceded by the OS.

At a behavioural level, however, the rats showed selective discrimination of the S_D preceded by the OS. Therefore, overall our data show that the rats use the information from the OS to determine behavioural output but that their dopamine systems did not respond strongly to the OS *per se*.

3.2 Introduction

Reinforcement learning occurs when organisms adapt the propensities of given behaviours on the basis of associations with reward and punishment. It is a useful algorithm because it is unsupervised, relying on trial-and-error learning under conditions in which the optimal solution is unknown. At a more fundamental level, an understanding of reinforcement learning is also important for our basic scientific understanding of habit formation, decision-making and microeconomics (e.g., see (Egelman et al 1998)). Recent neural network models of reinforcement learning are based on the neurophysiology of the mammalian dopamine system (Barto 1994; Contreras-Vidal & Schultz 1999; Dayan & Balleine 2002; Egelman et al 1998; McClure et al 2003; Montague & Berns 2002; Montague et al 1996; Montague et al 2004a; Schultz 2002; Schultz et al 1997; Suri & Schultz 1998a; 1999; 2001b).

The main finding of this research is that the dopamine system appears to function to minimise errors in the prediction of reward (or to maximize the truthfulness of opportunity gain signals) derived from the behavioural context through a process called *temporal difference learning*. This allows an organism to predict the time and amount of future rewards or punishments. The system also can guide the organism's behaviour when it tries to re-create the circumstances under which reward has been obtained previously.

An example of the role dopamine plays in minimising reward prediction errors can be seen in the phenomenon of classical (Pavlovian) conditioning, in which a sensory stimulus becomes associated with reward throughout repeated pairings. During the first stages of classical conditioning during which the association has not been formed, dopamine neurons respond with a burst of action potentials after the reward. The type of response seen is one of neurons exhibiting burst activity, or phasic activation, characterised by short latency (70-100 ms), short-duration (100-200

ms) and brief inter-spike intervals (10-50 ms) (Redgrave et al 2007; Schultz 2007a). However, after repeated pairings of the stimulus with the reward, the dopamine neurons respond only after the presentation of the conditioned stimulus (e.g., see (Wilson & Bowman 2006)).

They do not respond after the reward itself, which has been accurately predicted because of the occurrence of the preceding stimulus. The dopaminergic response, however, is controlled by a number of additional factors, including: delayed reward presentation, the motivation of the animal, and the type of choice amongst rewards (Hollerman & Schultz 1998; Morris et al 2006; Redgrave et al 2007; Satoh et al 2003).

The success of current reinforcement learning models has been restricted to minimalist environments in which only 1-2 environmental stimuli are present as possible predictors of reward. However, very little is known about the responses of the dopamine system to configurations of multiple stimuli, with the exception to date of two studies whereby stimuli were presented simultaneously. This occurred in the case of blocking and a conditioned inhibition paradigm (Balleine & Dickinson 2006; Tobler et al 2003). However, the response of the dopamine system to sequences of events as predictors of reward is unknown. One such paradigm is occasion setting, whereby the reward is contingent on a given stimulus (the conditioned stimulus in Pavlovian conditioning, a discriminative stimulus in instrumental conditioning) only when the stimulus has been preceded by another stimulus (the occasion setter).

Thus, occasion setters indicate that the relationship between a conditioned stimulus and reward is in effect. At a behavioural level, occasion setting properties have been well characterised. In fact, the phenomenon of occasion setting has been distinguished from that of simple conditioning on the basis that occasion setters act on the specific association between OSs and CSs, and are not creating a direct link with

the US (Ross 1983). They are hence said to act as facilitators, or modulators of behaviour (Boakes et al 1997; Bonardi & Ward-Robinson 2001).

Therefore, in this experiment, rats will undergo the occasion setting behavioural task whilst dopaminergic neurons are being recorded. Once the biological data is collected, it will be matched with the qualitative output of the TD model (Chapter 2).

3.3 Methods

3.3.1 Subjects

16 Listar Hooded adult male rats (Harlan, UK) were housed in pairs on a light 12h: dark 12 h cycle, weighted 340 to 548 g when training began. Rats were allowed to consume water from 16.00 h to 17.00 h each weekday and from Friday 16.00 h to Sunday afternoon during experimental training. During this period, the rats' body weight was monitored so that it would not fall below 85% of their free drinking weight. Following electrode implantation, rats were housed in isolation. All procedures conformed to the United Kingdom 1986 Animals (Scientific Procedures) Act.

3.3.2 Apparatus

3.3.2.1 Behaviour

Training and testing of rats occurred in sound-attenuated chambers (34 cm · 29 cm · 25 cm; Medical Associates Inc., St Albans, VT, USA), fitted with a video camera (Santec SmartVision, modelVCA 5156; Sanyo Video Vertrieb GmbH Co., Ahrensburg, Germany) for monitoring the rats' behaviour. The chamber contained a standard retractable lever, drinking spigot, houselight and piezoelectric buzzer (model EW-233A, Medical Associates Inc.). Sodium saccharin solution (0.25% w/v) was pumped out of the drinking spigot at 0.05 mL/s from a 50-mL glass syringe (Rocket, London) by computer-controlled syringe pumps (model PHM-100; Medical Associates Inc.)

3.3.2.2 Neurophysiology

The electrode arrays contained a movable bundle of four 50-µm stainless steel microwires coated in Teflon (impedance 0.4-1.3 MΩ). The microwires could be advanced by ~317.5 µm/turn in each recording session by turning an 80-thread/inch set screw (Small Parts Inc., Miami Lakes, FL, USA). The arrays weighed between 1.3 and 1.4g and measured 6mm along the mediolateral axis and 11mm along the anteroposterior axis. During recording sessions, the rat was connected to a preamplifier headstage using field effect transistors (input impedance 100 MΩ) which was in turn attached via a flexible cable to an electrical commuter (Stoelting Co., Wood Dale, IL, USA).

In order to remove noise, and lickometer artefacts, neuronal activity was recorded differentially from each of two pairs of wires (Sasaki et al, 1993). A custom built lickometer was also used to minimise lickometer artefacts (Malcolm McCandless, University of St Andrews) which produced a signal of sufficiently high frequency (> 5 kHz) that they could be filtered out. Amplification by 100 000x was obtained from each pair of wires using a Neurolog System (Digitimer Research Instrumentation) and frequencies <1 kHz and >5 kHz were attenuated by filters. Two Quest Scientific 'Hum Bug' digital filters (Digitimer) were used to eliminate 50 Hz noise. The differential activity from the two pairs of wires was finally digitised by the CED 1401+ data acquisition system using the associated Spike2 software (Cambridge Electronic Design, Cambridge, UK).

Waveforms of putative action potentials were sampled at 20 kHz. Behavioural events were communicated from the MED-PC to the CED 1401+'s digital inputs for time-stamping. The temporal resolution of the MED-PC system was 2 msec.

3.3.3 Procedures

Rats were trained over a ~2 month period to reach the final stage of occasion setting training (See Fig.1). The initial training stages were adopted from a previous study (Wilson & Bowman 2006). The later stages of training were instead a modification of a previously published work (Holland 1995).

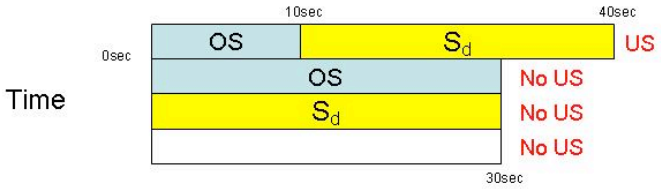


Figure 1: Task. Thirsty rats (n=16) were reinforced with saccharin solution for bar pressing only if the OS (tone or light) overlapped a subsequent discriminative stimulus (S_D, light or tone). Each trial lasted 40 secs. When presented, the OS lasted throughout each trial and the S_D was presented in the last 30 secs. The rats could earn multiple reinforcers during periods in which OS and the S_D were presented together. A reward signal (*CS*) was presented prior to each reinforcer (US).

3.3.3.1 Stage 1: Reward magazine training

Rats were trained in one 30- min session to lick the reward spigot to obtain saccharin solution. Rats were only able to gain access to saccharine reinforcement at a variable interval time schedule in which the first lick after 2, 4, 8 or 16 s (randomly chosen on each trial) was reinforced. No reinforcement was delivered outside these time schedules. A lick made after one of the four variable time schedules was simultaneously followed by the presentation of a conditioned stimulus (the reinforcement signal, CS, that is, reward magazine light).

This was followed by the delivery of 0.05 mL (0.05 mL/s) of sodium saccharin solution (0.25% w/v) whilst the RS was continuously presented. The rats consumed the primary reinforcer directly at the spigot during its delivery.

3.3.3.2 Stage 2: Modified FR1 training

Rats were then introduced to bar pressing for reward delivery on a modified FR1 (fixed-ratio responding) schedule of lever pressing for 60 minutes. Reward delivery here occurred as in the previous stage, except for having implemented two changes. First, rats were able to gain access to reward by licking on a variable interval time schedule of 32s and 64 s (randomly chosen on each trial). Second, a lever was protruded for the entire 60 minutes, and each bar press at any given time was followed by the delivery of sodium saccharin solution. Thus, rats were able to gain reward either through licking during the variable time schedule or by bar pressing (at any time). Rats that completed 50 trials (16 out of 16) moved on to the next stage of training.

3.3.3.3 Stage 3: Standard FR1 training

Reward delivery did not occur by licking of the reward spigot as in stages 1 and 2. Rats were only rewarded when bar pressing for reward on a FR1 schedule (1 bar press=reward delivery) over 60 minutes. An arbitrary set point of 50 bar presses per session was introduced, before a rat could reach the next stage of training. Rats that did not achieve 50 responses (2 out of 16) were given further identical sessions.

3.3.3.4 Stage 4: Discriminative stimulus training

The next stage was designed to place lever pressing under the control of a discriminative stimulus (S_D). The rats were split pseudorandomly into two groups, either being presented with a housetone or houselight as the S_D associated with sodium saccharin delivery. Each session lasted 60 minutes, with 30s random blocks of S_D-UCS pseudo randomly interleaved with 30s blocks of no S_D-no UCS. Rats were hence only rewarded when bar pressing under the 30s continuous presentation of the S_D.

Bar presses under S_D presentation versus no S_D were recorded as the discrimination rate. The latency from bar pressing to licking for sodium saccharin delivery was also recorded. Rats (16 out of 16) that achieved an above 80% rate discrimination for S_D-UCS were allowed to go on to the next stage.

3.3.3.5 Stage 5: Occasion setting training

The final stage of training was conducted over the next ~11 sessions. There were four kinds of trials in each 60-min session. Only one such kind of trial allowed the rat access to the sodium saccharin solution. The four kinds of trials consisted of the following (see *Figure 1*, 3.3.3). The rats were randomly presented with: 1) S_D for 30s-no reward for bar pressing 2) OS (occasion setter, a light for the group of rats that were previously trained using a tone as the discriminative stimulus and vice versa) for 30s-no reward for bar pressing 3) no OS or S_D for 30s-no reward for bar pressing 4) OS for 10s followed by OS+S_D for 30s in which time bar pressing was effective for reward delivery.

The rats' ability to discriminate between the OS+S_D condition versus the other trials was calculated as a discrimination rate, that is, the rate of bar pressing responses per second. A 2x2 repeated measure ANOVA was used to determine whether the rats

could significantly discriminate the occasion setting trial, and could hence be ready for electrode array implantation, or required further training.

3.3.4 Surgery

Following behavioural training, rats underwent surgery to implant an electrode array which was fixed onto the skull. Rats were anaesthetized using mixture of Isoflurane (5% for induction, 2% for maintenance) and oxygen (1.0 L/min). A presurgical nonsteroidal, nonopiate analgesic Rimadyl™ (0.5 mL/kg; 5% w/v carpofen; Pfizer Ltd, Kent, UK) was injected subcutaneously. In order to lower the electrode array into place, a hole was drilled stereotaxically at the top of the ventral tegmental area (5.80 mm posterior and 0.8 mm lateral to bregma; 7.4–8.4 mm ventral to skull surface).

In addition, five to seven holes were drilled around the area to which the electrode array would be attached, tapped for retaining screws (0–80 hex head set screws, 1/4 inch; Small Parts). Using the stereotaxic arm, the electrode array was lowered and dental acrylic used to retain the array attached to the cranium.

3.3.5 Histology

The following procedure is based upon previous work (Wilson & Bowman, 2006). Rats underwent ~3 weeks of neurophysiological recording, and were killed by overdose with 0.8 mL Dolethal TM (200 mg/L pentobarbitone sodium BP; Univet Ltd, Oxford, UK). Following death, they were perfused intracardially with 0.1% phosphate-buffered saline, plus a fixative (4% paraformaldehyde in 0.1 m phosphate buffer).

A freezing microtome was used to cut sections 50 μm thick. These sections were then collected in 0.1 m phosphate buffer, and every fourth was stained for tyrosine hydroxylase and Nissl bodies using standard immunohistochemistry

protocols. In order to conform the position of the electrode microwires with reference to the VTA, all stained sections were analysed under a light microscope and mapped onto standardized sections of the brain (Paxinos & Watson, 1997).

3.3.6 Data analysis

3.3.6.1 Behaviour

The behavioural analysis was restricted to sessions in which neurons were estimated to be in dopaminergic areas (as seen in tyrosine hydroxylase stained areas). The rate of bar pressing for the OS→S_D condition was corrected for the time it took to consume the reinforcer (saccharin solution was dispensed for 3s in each trial). The average bar pressing rate for each rat for each condition was calculated and analyzed by a 2 X 2 repeated-measures ANOVA in which the presence/absence of the OS was one factor and the presence/absence of the S_D was the second. Corrections for heterogeneity of variance were explored using a Levene's test but the variance across conditions was sufficiently similar that these were not used.

3.3.6.2 Neurophysiology

3.3.6.2.1 Spike sorting

Spikes were firstly sorted online in Spike 2 ™ (CED) using the waveform shape template matching, and re-sorted offline by performing principal components analysis on 20–60 data points of every waveform in the data set. The first three principal components of each spike were assigned a co-ordinate in 3-D space, to be able to cluster similar waveforms together. Separate clusters were then classified using the Normal Mixtures algorithm in Spike2 (modified to include waveforms 2.5-3SD of the centre of that cluster).

Finally, overlaid waveforms were visually inspected to reject any putative spike that seemed to be the result of a mechanical or electrical noise. The quality of the clustering was assessed by calculating the signal-to-noise ratio within each cluster. Single neurons were classified using the following criteria: there were no signs of noise at 50 Hz or its harmonics, the inter-spike interval histogram exhibited a refractory period, and there were no electrical artefacts within the cluster from the rat bar pressing or licking the spigot.

The use of peri-event histograms of the neuron's firing rate with regard to press and lick onset allowed for discrimination of electrical artefacts when these masked the true firing rate of the neuron when the rat had licked or pressed. Data contaminated by electrical artefacts were dropped from the sample. Data recorded from identical neurons over different testing session were also dropped.

3.3.6.2.2 Measuring spike duration

In Spike2, an automated algorithm was created (by Dr Eric Bowman) which calculated the action potential duration based on the middle 96% of the area of the averaged waveform.

3.3.6.2.3 Windows for spike counts

We estimated that there would be neural responses to the OS, S_D, or CS within 300ms of their onset, based on previous work (Redgrave et al 2007; Schultz et al 1997; Wilson & Bowman 2006). As a result, the firing rate in 300 msec time windows before and after a given stimulus onset were compared.

3.3.6.2.4 Classification of response type

We performed a within-subject design ANOVAs (repeated-measures factor, Epoch) on the spike frequency (Hz) of each neuron during the 300ms pre stimulus onset and 300ms post stimulus onset. This was carried out for the OS, S_D, and CS. Neurons were classified as exhibiting a response to the OS, S_D, and CS when there was a significant pairwise comparison ($p \leq 0.05$) between baseline (300ms pre stimulus onset) and stimuli time windows (300ms post stimulus onset). We also performed a 2x2 mixed ANOVAs design on the spike frequency of each neuron pre and post stimuli onset, looking at the effect of S_D alone, versus S_D preceded by the OS (OS by Epoch interaction). Neurons were classified as exhibiting a response to the OS preceded by the S_D when there was a significant OS by Epoch interaction ($p \leq 0.05$) corrected for heterogeneity of variance using the Greenhouse-Geisser correction. Please note that the use of multiple ANOVAs may result in false positives. To act as a control, thus, we included data from red nucleus cells (See Table 1 and 2).

3.4 Results

3.4.1 Behaviour

We measured the bar pressing frequency per minute of the six rats that underwent neurophysiological recording. We found that rats' bar pressing was significantly greater during the compound presentation of OS and S_D (See Fig. 2).

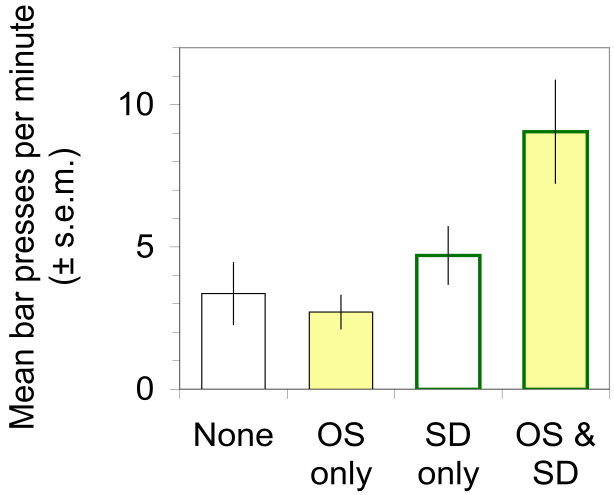


Figure 2: Bar pressing frequency for n=6 rats under the four stimulus conditions in the experiment. Responding was greatest during compound presentation of the OS (indicated in bars with yellow shading) and S_D (indicated in bars with a green border) ($F_{(1,5)} = 9.497$, $p<0.03$).

3.4.2 Neurophysiology

3.4.2.1 Electrophysiological characteristics of neurons

We recorded from 78 neurons in 6 rats responding within the occasion setting task. In order to confirm that the neurons we recorded from were from midbrain dopaminergic areas, we looked at the location of the electrode tract, and estimated the dorsal-ventral distance from the first recording session to the last. On this basis, we counted 45 out of 78 neurons that were within tyrosine hydroxylase-stained areas of midbrain dopaminergic neurons (of which a majority from ventral tegmental area and a minority from substantia nigra *pars compacta*; see Figure 6).

The remaining 33 neurons were recorded from the red nucleus, which is located dorsal to the VTA. In addition to histological examination, we then aimed to identify putative dopaminergic neurons using previous reports of average firing rate (<10 Hz) and action potential duration (>1ms) (Hyland et al 2002; Pan et al 2005; 2008). However, numerous investigations have pointed out that specific characterisation of VTA dopaminergic neurons based on action potential duration and average firing rate to be a poor criteria for distinguishing VTA dopamine neurons from non VTA dopamine cells.

For example, different groups have suggested that action potential duration of putative dopamine neurons to be in the range of >1.0 to 4.5 ms (Aghajanian GK 1973; Johnson & North 1992). Furthermore, a recent study has shown that no specific measure of action potential duration had a significant difference that correlated with TH staining (Margolis et al 2006). Therefore, we report here all neurons that were in TH stained areas and define them as possible dopamine neurons (DA'), and neurons in non TH stained areas (red nucleus) and define them as non-dopaminergic. In our population of neurons (78), we found that DA' neurons average firing frequency was non significantly higher than for non dopaminergic red nucleus neurons (11.3 Hz

versus 9.70 Hz), and their action potential duration was slightly longer than for non dopaminergic (1.13 ms versus 1.08 ms).

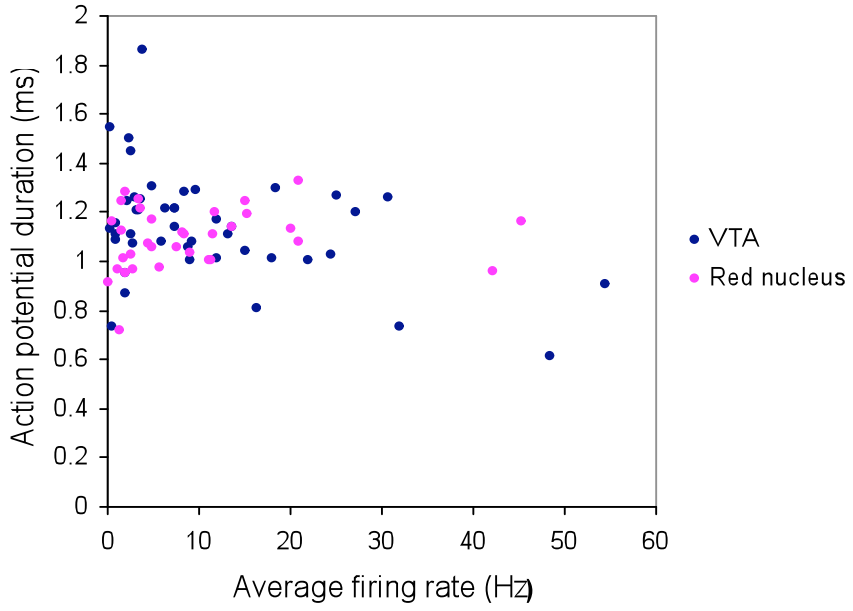


Figure 3: Scatterplot of average firing rate and action potential duration during recording sessions for VTA/SNpc neurons (n=45) and Red nucleus (n=33).Independent-samples t-tests showed that the mean action potential duration of VTA/SNpc neurons was not significantly longer than that of the remaining neurons ($t_{(76)}$ =1.219, p= 0.227). Mean average firing of VTA/SNpc neurons was non significantly different than that of the remaining neurons ($t_{(76)}$= 0.625, p=0.534)

3.4.2.2 Neural responses to stimuli presentation: action potential duration and average firing rate

In our population of neurons (78), we found that the responsive neurons' (51) firing frequency was significantly higher than that of unresponsive neurons (27) (12.83 Hz versus 6.01 Hz). Average action potential of responsive neurons was not significantly longer than that of remaining neurons (1.12 ms versus 1.10ms).

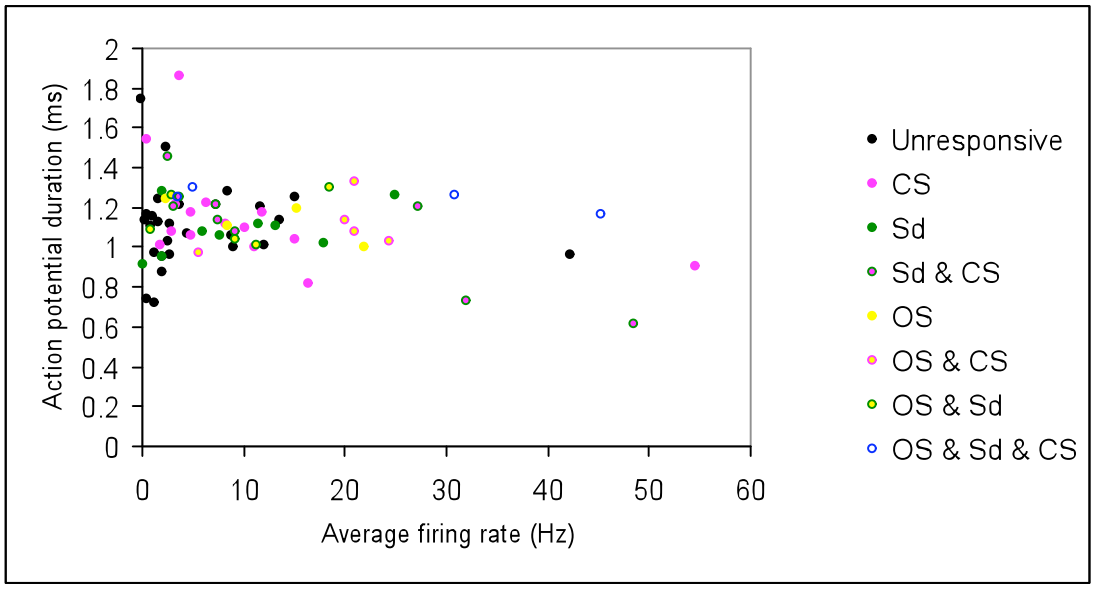


Figure 4: Scatterplot of average firing rate and action potential duration during recording sessions for responsive neurons (51) and unresponsive neurons (27), Responsive neurons here defined as neurons responding to any of the 7 stimuli combination (CS, S_D, S_D & OS, OS, OS &CS, OS & S_D, OS & S_D & CS). Independent-samples t-tests showed that the mean action potential duration of responsive neurons was non significantly longer than that of unresponsive neurons ($t_{(76)}$= -0.492, p= 0.625). Mean average firing rate of responsive neurons was significantly higher than that of remaining neurons ($t_{(76)}$= -.2846, p=0.006).

3.4.2.3 Distinguishing neural responses in the occasion setting paradigm

DA' neurons in the VTA/SNpc responded most prevalently to CS presentation (24/45; 53%) (See figure 5 for example responses from 4 neurons to stimuli presentation). This average was lower in nondopaminergic areas (red nucleus) (12/33; 36%). Neural responses to the S_D were also higher in dopaminergic versus nondopaminergic areas (VTA/SNpc= 18/40; 40% Red Nucleus= 9/33; 27%). Dopaminergic responses to the OS were the least prevalent (8/45; 18%). Non dopaminergic neurons responses to the OS were similar in proportion to responses to S_D and CS (10/33; 30%). Overall, it appears that dopaminergic neurons displayed a more selective pattern of responses to stimuli presentation (OS, S_D, and CS) than non dopaminergic neurons. Non-dopaminergic neurons, in fact, seemed to respond uniformly to all stimuli (See Table 1).

Stimulus	Neurons (%)[a]	Mean partial η^2
OS $_{(VTA/SNpc)}$ OS $(_{No\ VTA})$	8 (18%) 10 (30%)	0.056 0.063
S_D $_{(VTA/SNpc)}$ $S_{D\ (No\ VTA)}$	18 (40%) 9 (27%)	0.129 0.076
CS $_{(VTA/SNpc)}$ CS $_{(No\ VTA)}$	24 (53%) 12 (36%)	0.135 0.010

a Total=78

Table 1: Comparison of the responses neurons in the dopaminergic midbrain (top side of each row) versus neurons in adjacent overlying structures (bottom side of each row) at the OS, S_D and CS onsets by the neurons in the sample. For each neuron, the firing rate in 300 msec time windows before and after a given stimulus onset were compared. Effect sizes are reported in the form of mean partial eta squared. This refers to the proportion that a given variable "x" accounts for of the overall variance

In addition, we found a weak enhancement of the responses to the S_D by the OS. This suggests that the neural responses to the S_D are not dependent solely on the past history of reinforcement (See Table 2).

Condition	S_D-responsive neurons (%)[a]	Mean partial η^2
S_D alone	12 (57%)	0.145
OS → S_D	16 (76%)	0.203

a Total=21

Table 2*:* For neurons in VTA/SNpc, responses to the S_D were more prevalent and had greater strength in the presence of the OS. A non parametric sign test on the pairs of estimated partial η^2's from S_D alone versus S_D preceded by the OS across neurons showed that the result was non significant (p>0.05). Note that for neurons with S_D responses observed only when the OS is present or absent, n=21 (47%). Therefore, we report 3 additional neurons (Table 1, $S_{D;}$ n=18) that responded to S_D alone and S_D in the presence of the OS (OS→S_D)

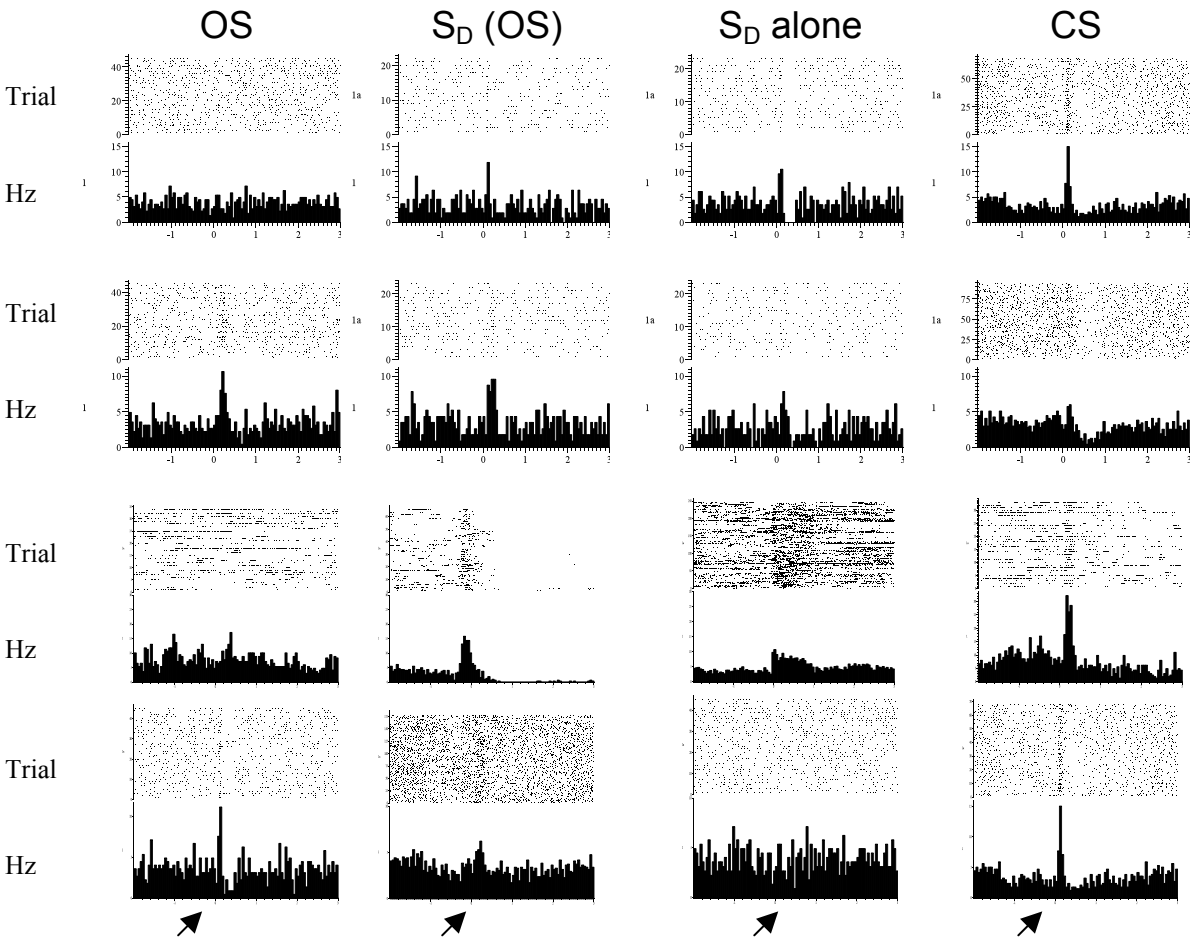


Figure 5: Example responses observed from 4 neurons in the VTA. Rasters and histograms show average firing rate (Hz) of the neurons relative to OS , S_D (OS), S_D alone, and CS onset on all trials (Trial) (black arrows=stimulus onset). The top row shows the typical response of a neuron to stimuli presentation: most pronounced firing to the CS, a slight enhancement of responding to the S_D in the presence of the OS, and very little responding to the OS alone. The other rows, on the other hand, show that occasionally, neurons fired to the OS, were less responsive to the CS (second row), and responded strongly to the S_D in the presence of the OS.

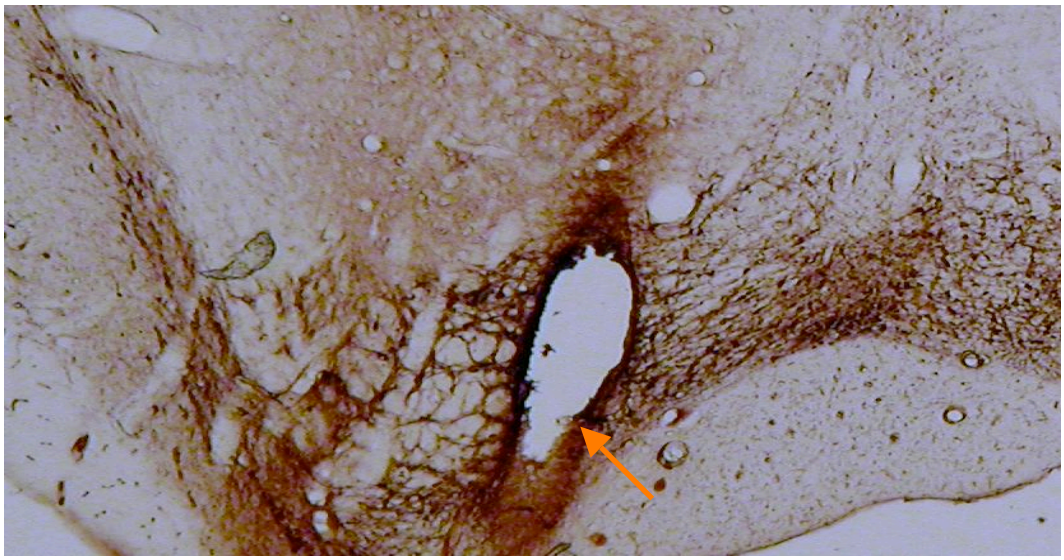


Figure 6*:* Photograph taken under light microscopy of tyrosine hydroxylase-stained brain section from one rat whereby damage (orange arrow) from microwires can be seen to be located in tyrosine hydroxylase-stained neurons (VTA).

3.5 Discussion

Our results show dissociation between the neural responses in the dopaminergic midbrain, conditioned behavioural responses, and the qualitative prediction of the temporal difference (TD) learning model used here. Namely, our TD model simulations (please see results in Chapter 2) showed that the onset of the earliest stimulus predicting reinforcement should evoke a strong response (OS⇒S_D⇒US). However, in our population of dopaminergic neurons we found very few responses to the OS. On the other hand, we find a weak enhancement of the responses to the S_D by the OS. Such enhancement is, however, not captured by models of reinforcement learning in the dopamine system, including those used here. Furthermore, and in stark contrast with the predictions of the TD model, the neurons in our sample responded most strongly to the most temporally proximal stimulus, the reward signal (CS).

Interestingly, we found that at a behavioural level the animals used the information provided by the OS, as bar pressing frequency was greatest during compound presentation of the OS and S_D: this is in accordance with previous behavioural investigations (Holland 1989; Ross & LoLordo 1986). Hence, we report here a dissociation between the enhancement of the responses to the S_D by the OS at a behavioural level and the weak enhancement of S_D responses in VTA/SNpc neurons when the S_D was preceded by the OS. In addition, we found a much more selective neural response in our dopaminergic sample to the various stimuli configurations, than we did in our subset of the presumed non-dopaminergic neurons. The non-dopaminergic neurons, in fact, responded in almost equal proportions to the OS, S_D, and CS (see Table 1).

There are at least three plausible explanations to account for the lack of neural response in the VTA/SNpc to the OS. First, the lack of response may be due to the

low contingency between the OS and the reinforcer (the reinforcer predicted reinforcer availability on only 50% of the trials). Second, the effect may be due to the rate of decay of the eligibility trace (Pan et al 2005). Please bear in mind, however, that whilst the temporal distance between OS onset and reward was >10s, the OS was on until the end of the trial, and in fact, overlapped with the S_D and the US (that is, it was not a case of trace conditioning). Therefore, this explanation refers to a set of neurons that may have been selectively sensitive to the onset of a stimulus, the OS, and not to the offset, as it would be in a trace conditioning paradigm. Third, in the training regime the S_D was initially trained to indicate responding would be reinforced and that might have lessened the dopamine systems responses to the OS in that the S_D may have become a very strong predictor of reward and hence resistant to extinction (therefore, the OS would have become partially redundant) (Pan et al 2008). The rationale for including this stage of training was based on previous behavioural studies which have included this training and have demonstrated that rats used the information provided by the OS, as significant discrimination was achieved when the S_D was preceded by the OS (Bonardi 2001; Holland 1995). We can speculate that the function of this stage of training is to build up contingencies of reinforcement and as a result, improve overall behavioural performance.

There are two previous papers that deal specifically with the issue of reward probability and dopaminergic conditioned responses, and both papers suggest the first possibility of our present findings is plausible but ultimately not viable (Fiorillo et al 2003; Schultz et al 1993). In one of these studies, two monkeys were trained in a Pavlovian procedure whereby distinct visual stimuli indicated the probability (p=0, 0.25, 0.5, 0.75, 1) of reward being delivered after a fixed 2-s delay (Fiorillo et al 2003). The results showed that the magnitude of dopaminergic cells' response to the CS increased with respect to increased reward probability (i.e. highest at P=1 and

lowest at P=0). In the present study, the OS was a P=0.5 predictor of reward, the S_D alone was a P=0, the S_D (preceded by OS) a P=1, and the CS was a P=1 also. In light of these results, the dopaminergic response that we report for the CS matches that seen in Fiorillo's work. However, the monotonic phasic increase seen at different probability values (P=0-1) in previous work, does not resemble the response pattern observed in the present study. In fact, there were more responsive dopaminergic cells to the S_D alone (P=0), than to the OS (probability of being followed by period of reinforcer availability=0.5). In addition, whilst there was a weak enhancement of the responses to the S_D preceded by the OS (P=1), there were nevertheless more active neurons during the equally (however, the CS could be interpreted as more predictive of reward that the S_D in that the necessary action has already been performed) reward predicting stimulus; the CS (P=1).

Fiorillo et al (2003) also report a sustained neural activity in the 2-s interval between the CS and reward, when the reward probability was set at P=0.5. It is therefore possible to compare such sustained activity in their sample, with the response pattern seen in our neurons. Namely, a somewhat weak response to the OS (P=0.5), leading up to increased activity at CS presentation, just before reward presentation was to occur.

However, such explanation presents a number of caveats, mostly because of methodological differences between the present study and Fiorillo's investigation. Firstly, the interval between CS presentation and reward in Fiorillo's work was 2-s, whereas in our experiment the interval was >10-s. A recent study has shown that, whilst not looking at increases in activity during the CS-reward interval, the dopaminergic response to conditioned stimuli is decreased with longer delays between CS's and rewards (Kobayashi & Schultz 2008). Secondly, whilst Fiorillo's study employed a Pavlovian procedure using a single conditioned stimulus, we recorded

dopaminergic activity in an operant conditioning paradigm using a configuration of stimuli. This means that:

A) obtaining the US is dependent upon the CR (bar-pressing) during the compound presentation of OS-S_D, making the precise timing from OS onset to US delivery somewhat variable

B) it is possible to speculate that due to the heavy attentional components required to obtain reward in our task, different neural substrates are therefore recruited.

Indeed, there is evidence to suggest that the prefrontal cortex (PFC) and the inferior parietal lobe (IPL) are involved in sustained attention whereby animals are required to track for extended periods of time the location of a cue that is predictive of reward presentation (Ciaramelli et al 2008). This could in turn provide an explanation for the relatively low dopaminergic response here reported to the OS, where sustained attention is required for a period of 10-s. Nevertheless, it could also be argued that given that the OS was always on, the rat may have only been processing the OS onset, and the change in stimulus contingencies (presentation of S_D, leading to reward presentation when bar pressing).

Another interpretation of our findings comes from the results of an operant conditioning task whilst dopaminergic neurons in the substantia nigra *pars compacta* were recorded (Schultz et al 1993). In one of the conditions, monkeys were presented with an instruction cue light followed by a 1-s trigger light. Lever pressing during the 1-s period when both lights were on, was followed by reward delivery. In the other conditions, the instruction cue light lasted 1-s and the trigger light would come up at a random variable interval of 2.5-3.5s after instruction onset. Lever pressing during the trigger light presentation was followed by reward. Interestingly, it was shown that the unpredictability of the trigger light presentation modulated the strength of the

dopaminergic response to that stimulus. That is, in the first condition where the interval between instruction cue and trigger light was fixed, 49% of the neurons responded during the instruction cue light, versus 9% were during the trigger light. In the second condition, where the interval between instruction cue and trigger light was variable, 38% of the neurons responded during the instruction cue light, versus 49% during the trigger light. In the present study, the interval between the OS and the S_D was fixed (10-s), albeit in only 50% of trials did the OS follow the S_D. Therefore, the weak enhancement of the responses to the S_D by the OS could be partially explained by the predictability of the interval between OS and S_D presentation.

Another interpretation of the present results deals with the rate of decay of the eligibility trace (ETP). As previously reviewed (See 1.2.2.7), the eligibility trace refers to the rate at which sensory events are forgotten by the system: the more of a trace that is forgotten, the less the system learns about it. In reinforcement learning models, this is a very important parameter that is used to handle delayed rewards. Pan et al (2005) showed that in a serial compound stimulus presentation using a high ETP value, the match between dopaminergic activity and modelling performance was fairly close. More specifically, two stimuli were presented in serial fashion, and separated by a 0.5s-2s interval, which were followed by reward. The results showed that dopaminergic responding occurred to both cues predictive of reward, as the model predicted. In our data, the neural eligibility trace of the onset of the OS might have decayed before reinforcement, so that dopaminergic responses would develop to the OS.

Therefore, the eligibility trace decay explanation (that is, the great temporal distance between OS onset and US resulting in low neural responding to the OS) I propose here is based on the idea that the event driving any association with the reinforcer (US) was the onset of the OS. Pan et al (2005) additionally reported that

dopaminergic cells responded in a differing manner in the early as opposed to late stages of training. Namely, they reported a slight decreased neural response to the second cue in the late stages of training, and an abolished response to the reward. In our sample, we report the neurophysiological data of fully trained animals only, and hence we are unable to comment on Pan's findings.

An alternative explanation for the maximal responding to the most proximal stimulus, the CS, is that such activation may be due to the saliency of the stimulus. However, whilst we know that dopaminergic cells respond to salient, attention-grabbing stimuli, the population of cells in our sample showed sustained response to the CS after thousands of trials, excluding an exclusively attentional component to the CS response. Therefore, our overall conclusion points out that there are constraints on dopaminergic neurons ability to influence conditioned responses to occasion setters. Such constraints, as discussed, appear to be derived from the OS being a 50% predictor of reward, as well as to the rate of decay of the eligibility trace. Given that at a behavioural level, rats were able to discriminate the compound presentation of OS-S_D, this raises the possibility, albeit very speculatively, that neural substrates in the PFC may have overcome such constraints, and modulated the conditioned behavioural response independently of the VTA/SNpc.

4 Chapter 4. Overexpectation in midbrain dopamine neurons

4.1 Abstract

In the overexpectation paradigm, two or more cues that have been first independently trained to predict reward are then presented in compound followed by the same reward. The overexpectation effect refers to reduced responding when the stimuli are retested individually after the compound presentation. Temporal difference learning models predict the effect on the basis of a negative prediction error. That is, reduced responding is said to occur as a result of the violation of summed expectations during compound conditioning when reward magnitude has been kept constant.

We recorded from midbrain dopaminergic neurons whilst thirsty rats (n= 5) performed a Pavlovian task for liquid reinforcement. Rats were presented with a reward magazine light (CS_A) or reward magazine tone (CS_B). These stimuli could occur either alone or in compound. Reward magnitude during the compound presentation was varied, so that it could either be a single reward ('+', as in the CS_A and CS_B alone) or a double reward delivery ('++'). Each of the possible six trials(CS_A+, CS_B+, CS_{AB}+, CS_{AB}++, CS_A-, CS_B-) were randomly selected within a block of 28 possibilities where CS_A+, and CS_B+ were present 24/28 times, and each of the other trials were present 1/28 times.

From a population of 29 dopaminergic cells, we found preferential responding to one of the two CS's, that is the CS_B+ (tone), contrary to previous expectations. Behaviourally, in fact, rats discriminated the tone and light equally well. However, we found reduced licking frequency during the compound stimuli presentation, which goes against predictions made by an elemental account of learning. In a minority of neurons, the response magnitude to the light and tone presented in compound

corresponded to the summed response magnitude to the light and tone presented individually, however, the effect was weak. Given the differential responding to the light and tone individually, we restrict our analysis to the effect of presenting the stimuli together and not to a possible decreased response to the individual stimuli after the compound presentation.

4.2 Introduction

In the previous chapter we showed dissociation between the neural responses in the dopaminergic midbrain and conditioned behavioural responses. The neurons responded weakly to the onset of the OS, and additionally, there was only a mild enhancement of the responses to the S_D by the OS. Behaviourally, however, the rats' rate of bar-pressing was increased when the S_D was preceded by the OS, demonstrating that they could make use of the predictive features of the OS. In contrast with our neurophysiological data, the qualitative prediction of temporal difference algorithms of reinforcement learning is that the onset of the earliest stimulus predicting reinforcement should evoke a strong response. Such interactions are not currently captured by models of reinforcement learning in the dopamine system, including those here used (Montague et al 1996; Pan et al 2005).

Temporal difference learning methods make specific predictions about another configuration of multiple stimuli, otherwise known to produce the overexpectation effect. Furthermore, given the dissociation previously found between behavioural and neurophysiological data, as well as the mismatch between the model's predictions, and the dopaminergic neurons responses to the CS and OS, it remains to be seen how the behaviour, model, and dopaminergic neurons will match with one another during the overexpectation paradigm.

The overexpectation effect refers to the prediction that the Rescorla-Wagner model makes regarding two separately trained conditioned stimuli (CS_A, CS_B) with the US, which are then paired together with the US, to finally be retested in isolation with the US (Rescorla & Wagner 1972). The model predicts that the associative strength that each stimulus gains, after being individually paired with the US, is summed once stimulus CS_A and CS_B are presented in compound with the US (Dawson & Spetch 2005). Because their associative strength exceeds that supported by the US,

this will lead to a partial loss of responding by each stimulus when retested individually with the US, until their associative strength returns to equilibrium with the US (Kehoe & White 2004). From a theoretical point of view, the overexpectation effect has been argued to exist only in so far as an elemental rather than a configural encoding of the elements occurs (Pearce 1987; Pearce & Bouton 2001).

However, from a configural account of the overexpectation effect, when a pattern of stimulation signals the US, both the single cue and the compound enter into association with the US (Collins & Shanks 2006). This results in the compound being treated as a separate stimulus, with no summation of associative strength from either stimulus CS_A or CS_B. In addition, as the associative strength of the compound matches that of the US, no negative prediction error ensues and no overexpectation effect is predicted when the two stimuli are retested individually (McNally et al 2004). The behavioural evidence for an overexpectation effect, indeed, seems to reflect the degree of theoretical divergence that exists between proponents of an elemental versus configural approach. In fact, the overexpectation effect is affected by factors such as the degree of similarity between two or more stimuli, as well as by how often the compound product is presented (Kehoe & White 2004).

In this experiment, we aim to particularly vary the ratio between compound and single elements presentation, whilst recording from dopaminergic cells *in vivo*, adopting the temporal difference models of reinforcement learning previously used (Montague et al 1996; Pan et al 2005).

4.3 Methods

4.3.1 Subjects

16 Listar Hooded adult male rats (Harlan, UK) were used and these weighted 300 to 360 g when training began. All additional information can be found in Chapter 3, (3.2.1).

4.3.2 Apparatus

4.3.2.1 Behaviour

All information can be found in Chapter 3 (3.3.2.1)

4.3.2.2 Neurophysiology

All information can be found in Chapter 3 (3.3.2.2)

4.3.3 Procedures

4.3.3.1 Stage 1: Reward magazine training

All information can be found in Chapter 3 (3.3.3.1)

4.3.3.2 Stage 2: Overexpectation training

The overexpectation training was carried out over the next ~5 sessions. Each session lasted 60 minutes. There were 6 kinds of trials. The rats were split pseudo randomly into two groups, either being presented with a reward magazine light as their CS_A or reward magazine tone as their CS_B associated with sodium saccharin delivery. The six trials consisted of: CS_A+, CS_B+, CS_A-, CS_B-, $CS_{AB}+$, $CS_{AB}++$. Each of the six trials was presented at random from a block of 28 possibilities where CS_A+ occurred (12/28), CS_B+ (12/28), CS_A- (1/28), CS_B- (1/28), $CS_{AB}+$ (1/28), $CS_{AB}++$ (1/28).

When stimuli where presented in compound (CS_{AB}), the rats could either receive a single (+) (0.05 mL/s) or double (++) (0.10 mL/s) reward amount (saccharin). Between each trial, a random inter-trial interval was selected (ranging

from 0.4s and 3.2 s). The number of licks from CS onset was recorded. Rats (16 out of 16) that achieved an above 80% rate discrimination during CS_A+, CS_B+, were allowed to go on to the next stage (surgery and neurophysiological recording).

4.3.4 Surgery

All information can be found in Chapter 3 (3.3.4)

4.3.5 Histology

All information can be found in Chapter 3 (3.3.5)

4.3.6 Data analysis

4.3.6.1 Behaviour

The behavioural analysis was enlarged to sessions in which neurons were estimated to be in dopaminergic and non dopaminergic areas. The average licking rate for each rat for each condition (CS_A, CS_B or CS_{AB}) was calculated and analyzed by a repeated-measure ANOVA in which the licking rate in 0.1 sec bins from 0-2 secs after CS onset constituted the dependent variable. Corrections for heterogeneity of variance were explored using a Levene's test but the variance across conditions was sufficiently similar that these were not used.

4.3.6.2 Neurophysiology

4.3.6.2.1 Spike sorting

All information can be found in Chapter 3 (3.3.6.2.1)

4.3.6.2.2 Measuring spike duration

All information can be found in Chapter 3 (3.3.6.2.2)

4.3.6.2.3 Windows for spike counts

All information can be found in Chapter 3 (3.3.6.2.3)

4.3.6.2.4 Classification of response type

We performed a mixed ANOVA design (repeated-measures factor, Epoch by CS) on the spike frequency (Hz) of all neurons during the 300ms pre stimulus onset and 300ms post stimulus onset. This was carried out for all stimuli (CS_A, CS_B, and CS_A+CS_B). Neurons were classified as exhibiting a response when there was a significant pairwise comparison ($p <= 0.05$) between baseline (300ms pre stimulus onset) and stimuli time windows (300ms post stimulus onset).

We also performed a repeated-measure ANOVA using Sidak corrections for multiple comparisons on the spike frequency of responsive dopaminergic neurons, pre and post stimuli onset, looking at which CS (CS_A, CS_B, CS_A+CS_B) was responsive (type of CS by Epoch interaction). Neurons were classified as exhibiting a response when there was a significant CS by Epoch interaction ($p \leq 0.05$).

4.4 Results

4.4.1 Behaviour

We measured the licking rate (Hz) of eight rats that underwent neurophysiological recording. We found that during the 0-2secs after CS_A+CS_B stimulus onset, rats' licking rate was detectably slower than during the individual presentation of CS_A and CS_B (See Fig. 1).

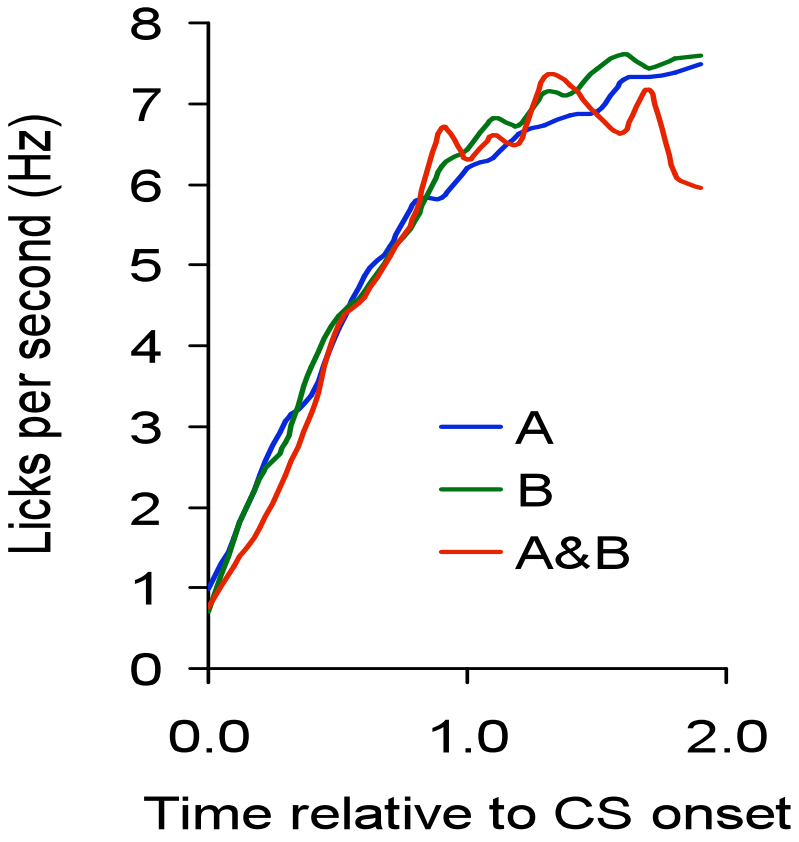


Figure 1: Licking rate frequency for eight rats under the three CS stimuli presentation (CS_A, CS_B, CS_{AB}). Responding was marginally slower during the compound presentation of CS_A and CS_B together (Stimulus x Time interaction). $F_{(2, 38)} = 2.636$, $p = 0.022$.$\eta^{2=}$.274

4.4.2 Neurophysiology

4.4.2.1 Electrophysiological characteristics of neurons

We successfully recorded from 45 neurons in 5 rats responding within the overexpectation task. In order to confirm that the neurons we recorded from were from midbrain dopaminergic areas, we looked at the location of the electrode tract, and estimated the dorsal-ventral distance from the first recording session to the last. On this basis, we counted 29 out of 45 neurons that were within tyrosine hydroxylase-stained areas of midbrain dopaminergic neurons (a vast majority from ventral tegmental area and a minority from substantia nigra *pars compacta*). The remaining 16 neurons were recorded from the red nucleus, which is located dorsal to the VTA.

The average firing rate and action potential duration were largely in line with those reported in the previous experiment. Hence, we will not here be reporting scatterplots of average firing rate/action potential duration as in Chapter 3. Of note, we found that dopaminergic neurons displayed higher signal to noise ratio than presumed non-dopaminergic cells, particularly so when these neurons were responsive to any of the CS stimuli presentation.

In addition, we found in our population of neurons a tendency for shorter latency to stimulus onset than in the previous experiment. Since we were concerned about contamination of the data from artefacts, we excluded from data analysis all putative units that displayed very short latency to stimulus onset, absence of a refractory period, and a low signal-to-noise ratio.

4.4.2.2 Distinguishing neural responses in the overexpectation paradigm

Dopaminergic neurons in the VTA/SNpc were more likely to respond to any of the CS combination than non dopaminergic cells (20/29; 69% in the VTA/SNpc, 6/16; 38% in the red nucleus) (see Table 1). Within this dopaminergic population of

responsive neurons (20/29), there was a preferential encoding of CS_B (tone) than CS_A (light) (20/20; 100% tone- 5/20; 25% light) (see Table 2). The frequency of responses to the compound presentation of CS_A and CS_B was in between the total number of responses to the presentation of tone and light individually (11/20; 55%), however, a substantial number of these responses were modulated by the excitatory responses to the tone (see Table 2). Occasionally, nevertheless, we found that responses to light and tone individually correlated with the presentation of the compound stimuli (see Figure 2).

Stimulus	Neurons (%)[a]	Mean partial η^2
Unresponsive $_{(VTA)}$	9 (31%)	0.003
Unresponsive ($_{No\ VTA}$)	10 (63%)	0.005
Responsive $_{(VTA)}$	20 (69%)	0.046
Responsive $_{(No\ VTA)}$	6 (38%)	0.079

[A] Total=45

Table 1: Comparison of the responses neurons in the dopaminergic midbrain (top side of each row) versus neurons in adjacent overlying structures (bottom side of each row) at the CS onset (CS_A or CS_B) by the neurons in the sample. For each neuron, the firing rate in 300 msec time windows before and after a given stimulus onset were compared. Effect sizes are reported as in Chapter 3.

Stimulus	Neurons (%)[a]	Mean partial η^2
CS_A (Light)	5 (25%)	0.006
CS_B (Tone)	20 (100%)	0.083
CS_{AB} (Light+Tone)	11(55%)	0.018

[A] Total= 20 (restricted to responsive neurons only)

Table 2*:* For neurons in VTA/SNpc, responses to the CS_B (tone) were more prevalent and had greater strength than responses to CS_A and CS_A and CS_B together.

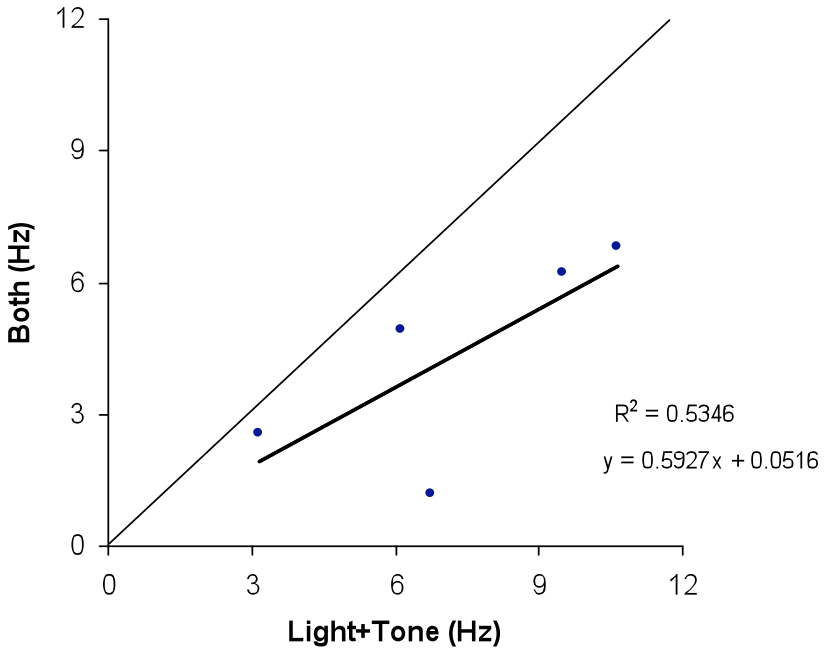


Figure 2: Scatterplot of the post minus pre firing rate (Hz) for dopaminergic cells (n=5) (where a significant difference from baseline was found) for the light and tone presented individually (Light+Tone), plotted against responses to the two stimuli presented simultaneously (Both).

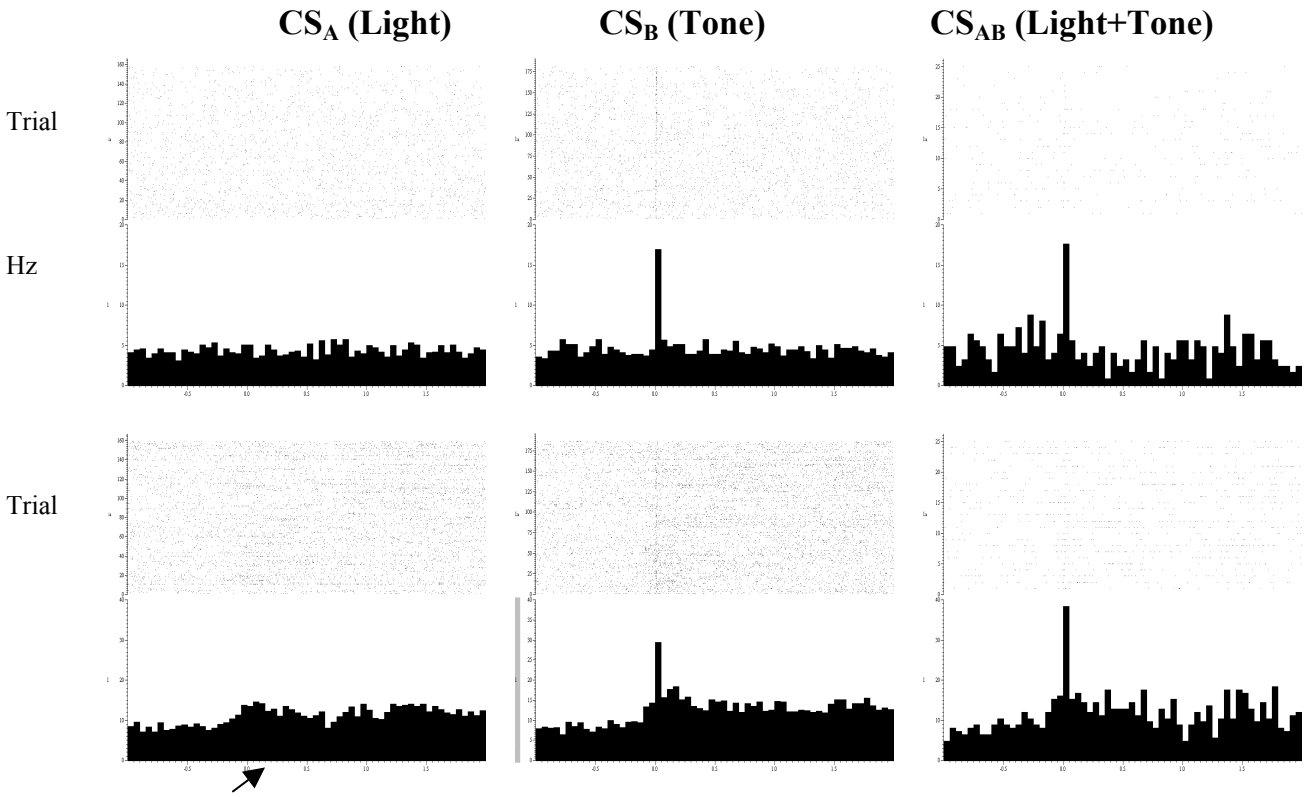


Figure 3: Example responses observed from 2 neurons in the VTA. Rasters and histograms show average firing rate (Hz) of the neurons relative to CSA, CSB, and CSAB onset on all trials (trial) (black arrows= stimulus onset). The top row shows the typical response of a neuron to stimuli presentation: most pronounced firing to CS_B (tone), where the properties of CS_A and CS_B combined do not add up to CS_{AB} presentation. The bottom row, on the other hand, shows that occasionally, neurons fired to CS_A (light) also, producing a pattern of response where CS_A and CS_B individually, add up to produce enhanced firing to CS_{AB} presented together.

4.5 Discussion

Our results show dissociation between the neural responses in the dopaminergic midbrain, conditioned behavioural responses, and the qualitative prediction of the temporal difference (TD) learning model here used. Namely, our TD model simulations (please see results in Chapter 2) predicted equal responding to CS_A (light) and CS_B (tone), and that the response to the two stimuli together should be greater than the response to the stimuli presented individually. However, in our population of dopaminergic neurons we found unequal responding to the two CS's. In addition, we found only a weak enhancement of the responses to the CS_A and CS_B together as modulated by an enhancement of response to CS_A and CS_B individually. At a behavioural level, we found that the animals used the information provided by both CS's, as licking frequency was similar during both CS_A and CS_B presentation.

This is indeed in accordance with previous behavioural literature (Khallad & Moore 1996). Interestingly, however, we also found that the presentation of the two stimuli together did not correspond with increased licking frequency as compared to when the stimuli are presented individually. Previous studies have demonstrated such effect (Takahashi et al 2009). We interpret such decreased conditioned response to the compound stimuli presentation due to the low number of trials that CS_{AB} occurred (2/28 in each block, as compared to 24/28 for CS_A and CS_B individually) (note that in Takahashi study, the CS_{AB} was presented on separate days).

Therefore, when CS_{AB} occurred, this may have been experienced as a novel stimulus which caused the animals to slow down their licking frequency. The reason for choosing only 2 presentations of the compound stimulus was based on the elemental account of TD learning which states that summation (that is, the associate strength of CS_{AB} is the result of the associative strength of CS_A plus CS_B) is optimally achieved when the compound stimulus has been made surprising (that is, infrequent

and unpredictable). In addition, an argument could be made that when CS_{AB} occurred, there was a 50% chance that this would be followed by the same reward magnitude as when the stimuli are presented individually, or by a doubled reward magnitude. As we have demonstrated in our previous experimental chapter (Chapter 3), when a stimulus is a 50% predictor of reward, both neural and behavioural responses are weak (onset of OS, see 3.4.1 and 3.4.2.3).

Although in both overexpectation and occasion setting experiments there was an equal probability of an outcome occurring, in the overexpectation paradigm, the occurrence of CS_{AB} indicated a 50% chance of a more desirable outcome, whereas in the occasion setting, the onset of the OS indicated a 50% chance of any outcome occurring. Therefore, this explanation can only apply insofar as the probability of an event is considered, regardless of the type of outcome.

Our sample of dopaminergic cells was most responsive to the presentation of the tone (CS_A) and not to the light (CS_B). At first sight, the interpretation of such results appears puzzling. Firstly, because the light stimulus (CS_B) is the same conditioned stimulus (CS) that was used in the previous experiment where the greatest neural response occurred (compared to OS, and S_D presentation, see 3.4.2.3). Secondly, because at a behavioural level, the animals licked in the presence of the light at a comparable rate to when the tone was presented (see 4.4.1). The neural characteristics of our population of dopaminergic neurons were largely similar to those reported in our previous experiment.

There was, however, a tendency for our populations of cells to have a shorter latency to stimulus onset. We excluded, therefore, from our final sample, those cells that responded within an implausibly short latency of stimulus onset, had very low signal to noise ratio, and displayed very little absolute refractory period. The remaining cells had a still shorter latency to stimulus onset than those reported in the

occasion setting paradigm; however, there are numerous reasons as to why we believe that we can exclude potential electrical artefacts.

Firstly, the cells had a high signal to noise ratio, and displayed an absolute refractory period. Secondly, electrical artefacts typically occur instantaneously after stimulus onset (whereas in the present sample, there was a >30msec latency to stimulus onset), and would be present in every trial where the stimulus is repeated; this did not occur in our sample. Thirdly, dopaminergic cells were more likely to respond to any of the CS combination (see Table 1, 4.4.2.2), and had a higher signal to noise ratio than our non dopaminergic sample (red nucleus), as we previously reported (see Table 1, 3.4.2.3).

The lower signal to noise ratio in the red nucleus than in the VTA/SNpc is suggestive that the signal increases as the bundle of stainless steel microwires moves away (ventrally) from the recording cannula (the red nucleus is located approximately 0.5mm from the cannula) (although a genuine difference between the RN and the VTA/SNpc cannot be excluded). The ability of VTA/SNpc cells to better discriminate amongst stimuli (OS, S_D, CS, see Chapter 3) compared to red nucleus cells, and their greater responsiveness to any CS combination in the present experiment than red nucleus cells, raises the possibility of two different neurotransmitters being sampled: a dopaminergic one in the VTA/SNpc, and a GABAergic/noradrenalinergic system in the red nucleus (Ciranna et al 2000). Therefore, having excluded possible artefacts as an interpretation of the greater responses to the tone but not the light, we can advance two possible propositions.

Firstly, one hypothesis is that the greater saliency of the tone overshadowed the presence of the light. Secondly, it may be advanced that there is a sensory map in the VTA. With regards to the first hypothesis, although behaviourally such effect did not occur, we can speculate that during the compound presentation of the stimuli, the

more salient stimulus (tone) suppressed the firing rate of the less salient stimulus (light). Whilst no neurophysiological evidence exists for this, we know from behavioural manipulations of the overshadowing paradigm that there are factors that are particularly likely to produce the overshadowing effect.

Namely, a low number of compound conditioning presentations can enhance the overshadowing effect to the less salient stimulus (and indeed, compound presentations of CS_{AB} occurred in only 2/28 cases per block), as well as a close temporal contiguity between the compound presentation and reward occurrence (in the present experiment, the CS remained on for the 2sec duration in which the rat was licking the reward spigot) (Stout et al 2003; Urcelay & Miller 2009).

A second possible interpretation of the data is that our sample of neurons in the VTA (with a minority of neurons sampled from the SNpc) responded in modality specific manner (preferentially encoding auditory versus visual sensory stimuli). The VTA receives sensory information mostly from the superior colliculus (SC) and the pedunculopontine tegmental neuron (PPTg). Although there is no direct evidence of VTA neurons encoding sensory specific information, we do know that the neurons of posterior side of the PPTg which project to the VTA respond to auditory but not visual stimuli and which respond to auditory stimuli with short mean latency (Winn 2008).

The SC, on the other hand, is constituted by a superficial layer that encodes visual stimuli, and a deeper layer that responds to auditory, tactile and visual stimuli which project to the VTA (King 2004). Although unlikely, we could speculate that the dopaminergic neurons in our sample where recorded from an area in the VTA that received auditory specific sensory information from the PPTg and the deep layers of the SC (however, the deep layers of the SC are multimodal). Alternatively, the findings raise the possibility of a sensory map in the VTA, with a visual part that is

partially separated from the auditory part; however, there is no evidence to back this up to the present date. To sum up, however, we believe that the overshadowing effect previously described a more plausible explanation for differential neural responding to the tone and light, although the simulations did not suggest this.

5 General discussion

5.1 Theoretical background and experimental summary

Reinforcement learning occurs when organisms adapt the propensities of given behaviours on the basis of associations with reward and punishment. Empirically, reinforcement learning models have largely been tested under conditions in which only 1-2 environmental stimuli are present as possible predictors of reward.

Therefore, we set out to understand how current models of reinforcement learning would respond under more complex conditions in which sequences of events are predictors of reward. An example of this type of situation is "occasion setting" in which reward is contingent on a given stimulus only when the stimulus has been preceded by another stimulus (the occasion setter). The other situation is the overexpectation effect, where two stimuli that have first been conditioned independently are then paired together, and are finally retested individually. At a behavioural level, the known effect is that provided that the reward magnitude during the compound presentation has been kept constant, there is diminished conditioned response to the two individual stimuli after the compound presentation.

In two experimental chapters of this thesis, we attempted to understand whether midbrain dopaminergic neurons would respond to occasion setters (Chapter 3), and to the overexpectation effect (Chapter 4). In addition, we ran simulations of the behavioural paradigms using temporal difference models of reinforcement learning (Chapter 2) and compared the predictions of the model with the behavioural and neurophysiological data.

In Chapter 3, by performing single-neuron recording from VTA and SNpc dopaminergic cells, we demonstrated that our population of neurons were most responsive to the latest predictor of reward (CS) and not the earliest (the OS). This is

in stark contrast with the predictions of the model (Chapter 2), where the greatest response is seen at the OS onset. We also showed at a neural level that there was only a weak enhancement of the response to the S_D when this was preceded by the OS. On the other hand, at a behavioural level, bar pressing was greatest when the S_D was preceded by the OS, demonstrating that rats could use the information provided by the OS, but that dopamine was not controlling the conditioned response.

In Chapter 4, our population of dopaminergic neurons showed that they would preferentially respond to only one of the two conditioned stimuli (CS_A, CS_B) in the overexpectation paradigm. The predictions of the model (Chapter 2) suggested that when the two stimuli would be presented in compound, there would be an inhibitory response if the reward magnitude was kept constant and an excitatory response if the reward magnitude was doubled. The lack of neural firing to one of the two conditioned stimuli, however, does not make for easy interpretation of the data.

Perhaps, one of the conditioned stimuli acted as if it were overshadowing the other, resulting in no response to the second CS. Interestingly, at a behavioural level, we did not see increased licking frequency to the compound stimuli presentation, a result that is somewhat at odd with the previous literature.

5.2 Chapter 2, 3 and 4: Making sense of it all: Behavioural, neurophysiological and modelling results in the occasion setting and overexpectation paradigm

In Chapter 3 (see 3.5), we provided two possible interpretations for our occasion setting data. Firstly, responses to the OS were weak because the OS was only a 50% predictor of reward. Secondly, responses to the OS were weak because of the great temporal distance between the OS and the US (>10s). Indeed, previous investigations partially support this hypothesis (Fiorillo et al 2003; Kobayashi & Schultz 2008). The weak neural responses to the OS match the low bar pressing frequency we saw at a behavioural level (see 3.4.1). This is in line with previous behavioural literature on the occasion setting phenomenon (Holland 1989; Ross 1983). In Chapter 2, however, the results showed that temporal difference simulations of the occasion setting paradigm, predicted that the greatest response should occur to the OS (see 2.4.1), something that we did not see in our neural population. Therefore, I speculated that given that at neural level there is some evidence of differential encoding of discriminative versus conditioned stimuli and that we run simulations of the occasion setting paradigm in classical fashion, future TD models could consider including a parameter that reflects differential associative weights between CS's and S_D's (Wan & Peoples 2006).

With regards to which neural structures may be encoding the OS, I previously reviewed evidence (Chapter 3) for which the PFC is highly activated when an animal is required to track the location of a cue that is predictive of reward for extended periods of time (Ciaramelli et al, 2008). Similarly, single cell recording studies show that PFC neurons are recruited during a two-stimulus discrimination task (Romo et al 2002; Salinas et al 2000). This task requires a monkey to compare the vibrotactile frequency of two stimuli that occur at two different time intervals, (summary of task: frequency 1 for 500msec, followed by a 3 sec interval, followed by frequency 2 for

500msec, then the monkey is required to press one of two buttons to indicate whether frequency 2 was higher or lower than frequency one: a correct response is followed by the delivery of a juice drop) and in order to do so, the monkey has to retain the information of the first stimulus in short term memory and later compare it with the second stimulus to reach a decision (Chow et al, 2009). This would suggest that in the occasion setting paradigm, the PFC stores the information provided by the OS, and at the time that the S_D is presented, is involved in the decision making process as to whether bar press for reward or not (presumably comparing the presence/absence of the OS).

The second important finding from the occasion setting experiment is that we found a dissociation between conditioned response behaviour to the OS → S_D pattern (S_D preceded by OS) and dopaminergic firing. That is, whilst rats showed increased bar pressing to the S_D (preceded by the OS), at a neural level, we only found a weak enhancement of dopaminergic firing to the S_D (when comparing S_D alone versus S_D preceded by OS). Whilst we could advance an interpretation of the findings for the OS on the basis of temporal discount and probability, this is more difficult to achieve for the OS → S_D pattern.

This is because the S_D preceded by the OS was a 100% predictor of reward, but also, because the timing between S_D onset and reward (US) was in the order of 0.5-1sec. It is interesting here to note that whilst both S_D (preceded by OS) and CS were 100% predictors of reward, the strongest response occurred to the conditioned stimulus and not to the discriminative stimulus. It is hence conceivable to speculate once again that there is something inherently different about the way in which the neural system encodes CS's and S_D's.

Indeed, most of the knowledge we have about midbrain dopaminergic neurons' responses to reward predicting stimuli has been based on behavioural

paradigms where conditioned stimuli have been used and in a few simple instrumental scenarios (Fiorillo et al 2003; Pan et al 2005; Schultz 1998; Schultz et al 1993; Wilson & Bowman 2006). Although, there is very little single neuron recording data comparing neural responses to discriminative versus conditioned stimuli, there is extensive psychopharmacological work using outcome devaluation and contingency degradation tasks that suggest that the core neural system involved in instrumental conditioning may not necessarily be the VTA or substantia nigra.

In the outcome devaluation task, after bar pressing to a S_D predictive of reward has been established during training, the outcome value of the reward is diminished, typically through specific satiety or aversion learning (Balleine & Ostlund 2007). Bar pressing frequency is then measured during extinction, and hence performance is compared between the devalued action, and a nondevalued control. Reduced bar pressing during extinction test (and after devaluation treatment) is taken as a measure of sensitivity to outcome (Yin et al 2005). That is, the animal has learned that the relationship between action-outcome or S_D-US has changed, and that bar pressing during S_D is no longer adaptive.

In the contingency degradation task, on the other hand, the probability that each bar press leads to reward presentation is altered so that the likelihood of receiving reward whether responding appropriately or not is equal. Lesions of the dorsomedial striatum have been found to render performance insensitive to both outcome devaluation and contingency degradation suggesting that the dorsomedial striatum (and in particular, the posterior area) may play a crucial role in the acquisition and expression of action-outcome associations in instrumental learning (Balleine 2005). Indeed, the neural substrate underlying the expression of action-outcome associations has been extended to include the prelimbic and medial

prefrontal cortex, the orbitofrontal cortex, and parts of the amygdala (Dolan 2007; Killcross & Coutureau 2003).

We believe, therefore, that an understanding of the distinction between instrumental versus classical conditioning could be an important factor in explaining our present results not only from a neural perspective, but also from a behavioural and reinforcement learning model view. Hence, from a behavioural point of view, instrumental learning is concerned with optimal choice (given the information available to the organism), or a learned association between actions that improve the subject's goals, the neural substrate of which, we have seen, may be composed by the dorsomedial striatum, the frontal cortex, etc.

In classical conditioning, however, choice is inconsequential to the occurrence or omission of reward and the response is linked with the positive/negative valence of the outcome (Dayan & Niv 2008). Here, we have reviewed extensive evidence which suggests that midbrain dopaminergic neurons respond to the presentation of conditioned stimuli.

From a modelling perspective, temporal difference methods of reinforcement learning have successfully been used to simulate the phasic activity of dopaminergic neurons in response to the unpredicted presentation of rewards, and to conditioned stimuli predictive of reward (Schultz 1998; 2004). One of TD models' architectural hallmarks, however, is that it uses momentary inconsistencies or prediction errors between a previous and a present state to learn values that are more accurate and which lead to higher outcomes.

Nevertheless, in doing so, the model combines information from the environment not only from the most direct experience, but also from previous and potentially inaccurate estimates of state values (Daw et al 2005). In addition, the information is encoded in a unique scalar value, as opposed to a composite variable,

so that it is not possible to identify a particular state with a given reward. This scalar value, therefore, represents a summary of its long-run future value, and is independent of specific outcome information (Daw et al 2005). As a result of this, TD models are slow in adapting to changes in contingency and outcome devaluation (Dayan & Niv 2008).

Therefore, it has been recently advanced that TD models may be best suited in modelling habitual action as modulated by dopamine and the dorsolateral striatum (Daw et al 2005). In contrast, so called "model-based" methods of reinforcement learning, have been proposed to better model goal-directed actions as modulated by the dorsomedial striatum, the prefrontal cortex, the OFC and the amygdala (Dayan & Niv 2008). Model-based methods of reinforcement learning differ from TD models (or model-free) in that they estimate long-term reward probabilities by building a model of the environment and by selecting the optimal action that best describes that task (Daw & Doya 2006). TD models, in contrast, learn directly from experience, and estimate long-term reward probabilities from one guess to the next (Sutton & Barto 1998).

Model-based methods are sensitive to changes in circumstances, so that in behavioural tasks where the reward is devalued the model quickly adapts by decreasing the value of actions that would lead to a devalued outcome (Daw et al 2005). If, therefore, we are to believe that instrumental action is best described by model-based methods and Pavlovian conditioned behaviour by model-free (TD models), we can start to have a better picture of how our behavioural, neurophysiological and modelling data may ultimately be explained.

The neural substrate that controlled conditioned responses to the S_D -OS pattern may have encompassed the dorsomedial striatum and its connections to the prefrontal cortex and parts of the amygdala, which are known to be involved in

response selection. On the other hand, the strong neural responses to the most proximal stimulus in the occasion setting paradigm (the reward signal, the CS), and to the CS_B in the overexpectation effect (Chapter 4), were encoded by dopaminergic neurons in the VTA and SNpc. The weak neural responses to the onset of the OS may be explained by the fact that the OS was a 50% predictor of reward, and that the temporal distance between the OS and reward was too great.

In the overexpectation paradigm (Chapter 4), the greater neural response to the tone (CS_B) and not to the light (CS_A) may be due to an overshadowing effect by the more salient stimulus. In our overexpectation simulations, however, although we experimented with varying the saliency of each conditioned stimulus, this had little qualitative effect on the overall output (please see in the appendices, Figure 7).The preferential neural response to the tone did not allow us to directly test whether responses to individual conditioned stimuli (in the overexpectation paradigm) were decreased after these had been paired together. However, a prediction of the model was that the compound presentation of the cues, should elicit a response that is equal to that of each conditioned stimulus added up together.

In the few instances where neurons responded not only to light and tone individually, but also when paired together, we can conclude that responses from each conditioned stimulus did not add up when the two stimuli were presented together. Similarly, at a behavioural level, we did not see increased licking to the compound cue presentation.

These data, although very limited in nature, suggests a configural rather than elemental account of learning, where the associative strength of the compound stimuli (CS_{AB}) is independently represented from that of the individual stimuli (CS_A, CS_B). Here, it should be pointed out, nevertheless, that the learning rule used to make predictions by the Pan *et al* and Montague's TD model is based on the Rescorla-

Wagner equation. In the Rescorla-Wagner rule, learning is viewed as occurring in an elemental fashion. In such equation $\mathbf{\Delta V_A = a \cdot \beta (\lambda - \sum V)}$: for a given cue A: $\mathbf{\Delta V_A}$ stands for changes in associative strength in a given trial; **a** is the learning rate parameter for the saliency of the cue; **β** is the learning rate parameter for the saliency of the outcome, **λ** the maximum associate strength supported by the US, **∑V** the current total associative strength of all cues presented (Collins & Shanks 2006).

Given these parameters, the model makes the following predictions. Firstly, the associative strength of a compound (CS_{AB}) will be equal to the sum of the associative strength of its element (CS_A and CS_B), that is, responding to the compound will be greater than that of the individual stimuli. Secondly, because the associative strength of the compound will be higher than that supported by the US, as trials progress, the associative strength of the compound will diminish in a manner that when the individual stimuli are retested in isolation ($CS_{A\ alone}$ and $CS_{B\ alone}$), their individual associative strength will be lower than that of control elements being conditioned in isolation but without being paired together (Collins & Shanks 2006).

This is indeed the overexpectation effect predicted using Pan *et al.* parameters incorporated in the present simulations (that ultimately rely on the Rescorla-Wagner rule). However, different predictions arise if a TD model relies on a learning rule that is based on a configural account of learning, the most prominent of which can be derived by the work of Pearce (Pearce 1994). In Pearce's rule, the associative strength of the compound is determined by the degree of generalization that occurs amongst its elements.

Therefore, the greater the similarity between stimuli, the greater the generalization (Collins & Shanks 2006). In Pearce's equation: $\mathbf{e_{A'} = {}_{A}S_{A'} \cdot E_A}$: the degree of excitation of a given stimulus A (E_A) generalises to a similar stimulus

$e_{A'}$ as by the similarity S of the two stimuli. The degree of similarity of two stimuli $_{A}S_{A'}$ is computed using the following formula:

$_{A}\mathbf{S_{A'}} = \mathbf{N_{com}} / \mathbf{N_A} \mathbf{x} \mathbf{N_{com}}/\mathbf{N_{A'}}$: where N_{com} is defined as the number of elements shared by $_{A}$ and $_{A'}$, whereas N_A and $N_{A'}$ represent the total number of elements of $_{A}$ and $_{A'}$ (Collins & Shanks 2006).

Thus, in a case where CS_A and CS_B are then followed by CS_{AB} (provided equal salience), the compound CS_{AB} will acquire ½ of associative strength from CS_A and ½ from CS_B, that is, the CS_{AB} compound will have equal associative strength to CS_A and CSB. Moreover, because the CS_{AB} compound fully predicts (that is, the associative strength of CS_{AB} will be supported by the US) the US, the associative strength of the compound will remain unchanged and no overexpectation effect would occur when the individual stimuli are retested in isolation (Collins & Shanks 2006).

The predictions regarding the overexpectation effect are therefore in stark contrast with those made by Pan *et al* (2005), and that ultimately rely on the learning rule by Rescorla-Wagner used in the present simulations. Although the predictions using the Pearce's rule cannot account for some of the neurophysiological data where preferential responding to one of the two CS's was reported (CS_B) (Chapter 4), they do provide some intriguing match of the data in the few instances where both CS_A and CS_B were activated but there was a lack of summation when the compound CS_{AB} occurred (see figure 2, 4.4.2.2).

5.3 Conclusion

The results of our two experimental chapters suggest that the role that midbrain dopaminergic neurons play in reinforcement learning is more complex than that envisaged by previous investigations (Bayer & Glimcher 2005; Pan et al 2005; Schultz et al 1997; Tobler et al 2005). In particular, we have shown (Chapter 3) that the general finding that dopaminergic neurons should respond to the earliest predictor of reward not to be the case. We have argued that such response could be affected by the temporal distance between the onset of the conditioned stimulus and the reward, and by the reward predictability of such stimulus. In addition, we speculated that the neural system controlling conditioned responses to the S_D –OS pattern may be encompassed by the dorsomedial striatum and its connections to the prefrontal cortex and amygdala: that is, structures known to be important in response selection.

Critically, our TD model simulations of the occasion setting and overexpectation paradigm have been unable to account for the response of our population of neurons in several ways (Chapter 2). They predicted greatest response to the OS, no modulation to the S_D in the presence of the OS (although we reported a weak modulation), and a doubled response to the presentation of a compound stimulus (CS_{AB}) in the overexpectation experiment. The dissociation between conditioned response and dopaminergic firing (Chapter 3 and 4), suggests that the complexity of a task using multiple configuration of stimuli may determine a pattern of results whereby dopaminergic neurons and additional neural system interact in a way that is unaccounted for by current TD models of reinforcement learning.

6 References

Abi-Dargham A, Rodenhiser J, Printz D, Zea-Ponce Y, Gil R, et al. 2000. Increased baseline occupancy of D2 receptors by dopamine in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 97:8104-9

Adell A, Artigas F. 2004. The somatodendritic release of dopamine in the ventral tegmental area and its regulation by afferent transmitter systems. *Neuroscience and biobehavioral reviews* 28:415-31

Adell A, Celada P, Abellan MT, Artigas F. 2002. Origin and functional role of the extracellular serotonin in the midbrain raphe nuclei. *Brain Res Brain Res Rev* 39:154-80

Aghajanian GK BB. 1973. Central dopaminergic neurons – neurophysiological identification and responses to drugs. *Life sciences*:643-8

Albin RL, Makowiec RL, Hollingsworth ZR, Dure LSt, Penney JB, Young AB. 1992. Excitatory amino acid binding sites in the basal ganglia of the rat: a quantitative autoradiographic study. *Neuroscience* 46:35-48

Alcaro A, Huber R, Panksepp J. 2007. Behavioral functions of the mesolimbic dopaminergic system: an affective neuroethological perspective. *Brain Res Rev* 56:283-321

Amiro TW, Bitterman ME. 1980. Second-order appetitive conditioning in goldfish. *Journal of Experimental Psychology: Animal Behavior Processes* 6:41-8

Balleine BW. 2005. Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiology & behavior* 86:717-30

Balleine BW, Dickinson A. 2006. Motivational control of blocking. *Journal of experimental psychology* 32:33-43

Balleine BW, Ostlund SB. 2007. Still at the choice-point: action selection and initiation in instrumental conditioning. *Annals of the New York Academy of Sciences* 1104:147-71

Barto. 1995. Adaptive Critics and the Basal Ganglia. *In: Models of Information Processing in the Basal Ganglia*

Barto AG. 1994. Reinforcement learning control. *Curr Opin Neurobiol* 4:888-93

Baskfield CY, Martin BR, Wiley JL. 2004. Differential effects of delta9-tetrahydrocannabinol and methanandamide in CB1 knockout and wild-type mice. *The Journal of pharmacology and experimental therapeutics* 309:86-91

Bayer HM, Glimcher PW. 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129-41

Bellingham WP, Gillette K. 1981. Attenuation of overshadowing as a function of nondifferential compound conditioning trials. *Bulletin of the Psychonomic Society* 18:218-20

Berger B, Tassin JP, Blanc G, Moyne MA, Thierry AM. 1974. Histochemical confirmation for dopaminergic innervation of the rat cerebral cortex after destruction of the noradrenergic ascending pathways. *Brain research* 81:332-7

Berridge KC. 2007. The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* 191:391-431

Berridge KC, Robinson TE. 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 28:309-69

Bertsekas DP, Tsitsiklis JN. 1996. *Neurodynamic Programming* Nashua, NH Athena Scientific

Birkmayer W, Mentasti M. 1967. [Further experimental studies on the catecholamine metabolism in extrapyramidal diseases (Parkinson and chorea syndromes)]. *Archiv fur Psychiatrie und Nervenkrankheiten* 210:29-35

Bitterman ME, Menzel R, Fietz A, Schafer S. 1983. Classical conditioning of proboscis extension in honeybees (Apis mellifera). *Journal of Comparative Psychology* 97:107-19

Blaiss CA, Janak PH. 2008. The nucleus accumbens core and shell are critical for the expression, but not the consolidation, of Pavlovian conditioned approach. *Behavioural brain research*

Boakes RA, Westbrook RF, Elliott M, Swinbourne AL. 1997. Context dependency of conditioned aversions to water and sweet tastes. *Journal of Experimental Psychology: Animal Behavior Processes* 23:56-67

Bonardi C, Ward-Robinson J. 2001. Occasion Setters: Specificity to the US and the CS-US Association. *Learning and Motivation* 32:349-66

Bonardi C, Yann Ong S. 2003. Learned irrelevance: a contemporary overview. *The Quarterly journal of experimental psychology* 56:80-9

Bozarth MA, Wise RA. 1981. Intracranial self-administration of morphine into the ventral tegmental area in rats. *Life sciences* 28:551-5

Bowman EM, Brown VJ. 1998. Effects of excitotoxic lesions of the rat ventral striatum on the perception of reward cost. *Experimental brain research. Experimentelle Hirnforschung* 123:439-48

Bradberry CW. 2002. Dose-dependent effect of ethanol on extracellular dopamine in mesolimbic striatum of awake rhesus monkeys: comparison with cocaine across individuals. *Psychopharmacology (Berl)* 165:67-76

Bunney EB, Appel SB, Brodie MS. 2001. Electrophysiological effects of cocaethylene, cocaine, and ethanol on dopaminergic neurons of the ventral tegmental area. *The Journal of pharmacology and experimental therapeutics* 297:696-703

Burgdorf J, Panksepp J. 2006. The neurobiology of positive emotions. *Neuroscience and biobehavioral reviews* 30:173-87

Cahill L. 1997. The neurobiology of emotionally influenced memory. Implications for understanding traumatic memory. *Annals of the New York Academy of Sciences* 821:238-46

Caille S, Parsons LH. 2003. SR141716A reduces the reinforcing properties of heroin but not heroin-induced increases in nucleus accumbens dopamine in rats. *The European journal of neuroscience* 18:3145-9

Calabresi P, Pisani A, Mercuri NB, Bernardi G. 1992. Long-term Potentiation in the Striatum is Unmasked by Removing the Voltage-dependent Magnesium Block of NMDA Receptor Channels. *The European journal of neuroscience* 4:929-35

Calabresi P, Saiardi A, Pisani A, Baik JH, Centonze D, et al. 1997. Abnormal synaptic plasticity in the striatum of mice lacking dopamine D2 receptors. *J Neurosci* 17:4536-44

Calton JL, Mitchell KG, Schachtman TR. 1996. Conditioned inhibition produced by extinction of a conditioned stimulus. *Learning and Motivation* 27:335-61

Cameron DL, Williams JT. 1993. Dopamine D1 receptors facilitate transmitter release. *Nature* 366:344-7
Carelli RM, Deadwyler SA. 1994. A comparison of nucleus accumbens neuronal firing patterns during cocaine self-administration and water reinforcement in rats. *J Neurosci* 14:7735-46
Carlezon WA, Jr., Thome J, Olson VG, Lane-Ladd SB, Brodkin ES, et al. 1998. Regulation of cocaine reward by CREB. *Science (New York, N.Y* 282:2272-5
Carr DB, Sesack SR. 2000. GABA-containing neurons in the rat ventral tegmental area project to the prefrontal cortex. *Synapse (New York, N.Y* 38:114-23
Cass WA, Gerhardt GA. 1995. In vivo assessment of dopamine uptake in rat medial prefrontal cortex: comparison with dorsal striatum and nucleus accumbens. *Journal of neurochemistry* 65:201-7
Chen NH, Reith ME. 1994. Effects of locally applied cocaine, lidocaine, and various uptake blockers on monoamine transmission in the ventral tegmental area of freely moving rats: a microdialysis study on monoamine interrelationships. *Journal of neurochemistry* 63:1701-13
Chergui K, Charlety PJ, Akaoka H, Saunier CF, Brunet JL, et al. 1993. Tonic activation of NMDA receptors causes spontaneous burst discharge of rat midbrain dopamine neurons in vivo. *The European journal of neuroscience* 5:137-44
Chow SS, Romo R, Brody CD (2009). Context dependent modulation of functional connectivity: secondary somatosensory cortex to prefrontal cortex connections in two-stimulus-interval discrimination tasks. *Journal of Neuroscience* 29:7238-7245.
Churchill L, Dilts RP, Kalivas PW. 1992. Autoradiographic localization of gamma-aminobutyric acidA receptors within the ventral tegmental area. *Neurochemical research* 17:101-6
Ciaramelli E, Grady CL, Moscovitch M. 2008. Top-down and bottom-up attention to memory: a hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. *Neuropsychologia* 46:1828-51
Ciranna L, Licata F, Li Volsi G, Santangelo F. 2000. Neurotransmitter-mediated control of neuronal firing in the red nucleus of the rat: reciprocal modulation between noradrenaline and GABA. *Experimental neurology* 163:253-63
Clarke PB, Pert A. 1985. Autoradiographic evidence for nicotine receptors on nigrostriatal and mesolimbic dopaminergic neurons. *Brain research* 348:355-8
Clements JD. 1996. Transmitter timecourse in the synaptic cleft: its role in central synaptic function. *Trends in neurosciences* 19:163-71
Collins DJ, Shanks DR. 2006. Summation in causal learning: elemental processing or configural generalization? *Quarterly journal of experimental psychology (2006)* 59:1524-34
Contreras-Vidal JL, Schultz W. 1999. A predictive reinforcement model of dopamine neurons for learning approach behavior. *J Comput Neurosci* 6:191-214
Corbit LH, Ostlund SB, Balleine BW. 2002. Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *J Neurosci* 22:10976-84Corrigall WA, Franklin KB, Coen KM, Clarke PB. 1992. The mesolimbic dopaminergic system is implicated in the reinforcing effects of nicotine. *Psychopharmacology (Berl)* 107:285-9
Crawford LL, Domjan M. 1995. Second-order sexual conditioning in male Japanese quail (Coturnix japonica). *Animal Learning & Behavior* 23:327-34

Czoty PW, Justice JB, Jr., Howell LL. 2000. Cocaine-induced changes in extracellular dopamine determined by microdialysis in awake squirrel monkeys. *Psychopharmacology (Berl)* 148:299-306
Dackis CA, Gold MS. 1985. New concepts in cocaine addiction: the dopamine depletion hypothesis. *Neuroscience and biobehavioral reviews* 9:469-77
Dahan L, Astier B, Vautrelle N, Urbain N, Kocsis B, Chouvet G. 2006. Prominent Burst Firing of Dopaminergic Neurons in the Ventral Tegmental Area during Paradoxical Sleep. *Neuropsychopharmacology* 32:1232-41
Dahlstroem A, Fuxe K. 1964. Evidence for the Existence of Monoamine-Containing Neurons in the Central Nervous System. I. Demonstration of Monoamines in the Cell Bodies of Brain Stem Neurons. *Acta physiologica Scandinavica*:SUPPL 232:1-55
David V, Durkin TP, Cazala P. 2002. Differential effects of the dopamine D2/D3 receptor antagonist sulpiride on self-administration of morphine into the ventral tegmental area or the nucleus accumbens. *Psychopharmacology (Berl)* 160:307-17
Daw ND, Doya K. 2006. The computational neurobiology of learning and reward. *Current opinion in neurobiology* 16:199-204
Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience* 8:1704-11
Dawson MRW, Spetch ML. 2005. Traditional Perceptrons Do Not Produce the Overexpectation Effect. *Neural Information Processing- Letters and Reviews* 7:11-7
Dayan P, Balleine BW. 2002. Reward, motivation, and reinforcement learning. *Neuron* 36:285-98
Dayan P, Niv Y. 2008. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology* 18:185-96
Deisig N, Lachnit H, Giurfa M, Hellstern F. 2001. Configural olfactory learning in honeybees: negative and positive patterning discrimination. *Learning & memory (Cold Spring Harbor, N.Y* 8:70-8
Di Chiara G, Tanda G. 1997. Blunting of reactivity of dopamine transmission to palatable food: a biochemical marker of anhedonia in the CMS model? *Psychopharmacology* 134:351-3; discussion 71-7
Di Forti M, Lappin JM, Murray RM. 2007. Risk factors for schizophrenia -- All roads lead to dopamine. *European Neuropsychopharmacology* 17:S101-S7
Di Giovanni G, De Deurwaerdere P, Di Mascio M, Di Matteo V, Esposito E, Spampinato U. 1999. Selective blockade of serotonin-2C/2B receptors enhances mesolimbic and mesostriatal dopaminergic function: a combined in vivo electrophysiological and microdialysis study. *Neuroscience* 91:587-97
Dickinson A, Balleine B. 1990. Motivational control of instrumental performance following a shift from thirst to hunger. *The Quarterly journal of experimental psychology* 42:413-31
Dolan RJ. 2007. The human amygdala and orbital prefrontal cortex in behavioural regulation. *Philosophical transactions of the Royal Society of London* 362:787-99
Egelman DM, Person C, Montague PR. 1995. A predictive model for diffuse systems matches human choices in a simple decision-making task. *Soc Neurosci Abstr* 1087
Egelman DM, Person C, Montague PR. 1998. A computational role for dopamine delivery in human decision-making. *J Cogn Neurosci* 10:623-30

Everitt BJ, Wolf ME. 2002. Psychomotor stimulant addiction: a neural systems perspective. *J Neurosci* 22:3312-20
Ferrari R, Le Novere N, Picciotto MR, Changeux JP, Zoli M. 2002. Acute and long-term changes in the mesolimbic dopamine pathway after systemic or local single nicotine injections. *The European journal of neuroscience* 15:1810-8
Ferreira JG, Del-Fava F, Hasue RH, Shammah-Lagnado SJ. 2008. Organization of ventral tegmental area projections to the ventral tegmental area-nigral complex in the rat. *Neuroscience* 153:196-213
Fields HL, Hjelmstad GO, Margolis EB, Nicola SM. 2007. Ventral tegmental area neurons in learned appetitive behavior and positive reinforcement. *Annual review of neuroscience* 30:289-316
Fiorillo CD, Tobler PN, Schultz W. 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299:1898-902
Fiorillo CD, Newsome WT, Schultz W. 2008. The temporal precision of reward prediction in dopamine neurons. *Nature neuroscience*
Fisher H, Aron A, Brown LL. 2005. Romantic love: an fMRI study of a neural mechanism for mate choice. *The Journal of comparative neurology* 493:58-62
Freeman AS, Kelland MD, Rouillard C, Chiodo LA. 1989. Electrophysiological characteristics and pharmacological responsiveness of midbrain dopaminergic neurons of the aged rat. *The Journal of pharmacology and experimental therapeutics* 249:790-7
Friston KJ, Tononi G, Reeke GN, Jr., Sporns O, Edelman GM. 1994. Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* 59:229-43
Fu Y, Matta SG, Gao W, Sharp BM. 2000. Local alpha-bungarotoxin-sensitive nicotinic receptors in the nucleus accumbens modulate nicotine-stimulated dopamine secretion in vivo. *Neuroscience* 101:369-75
Garzon M, Vaughan RA, Uhl GR, Kuhar MJ, Pickel VM. 1999. Cholinergic axon terminals in the ventral tegmental area target a subpopulation of neurons expressing low levels of the dopamine transporter. *The Journal of comparative neurology* 410:197-210
Geisler S, Zahm DS. 2005. Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *The Journal of comparative neurology* 490:270-94
Georges F, Aston-Jones G. 2003. Prolonged activation of mesolimbic dopaminergic neurons by morphine withdrawal following clonidine: participation of imidazoline and norepinephrine receptors. *Neuropsychopharmacology* 28:1140-9
Gerrits MA, Van Ree JM. 1996. Effect of nucleus accumbens dopamine depletion on motivational aspects involved in initiation of cocaine and heroin self-administration in rats. *Brain Res* 713:114-24
Giegerich R. 2000. A systematic approach to dynamic programming in bioinformatics. *Bioinformatics (Oxford, England)* 16:665-77
Giorgetti M, Hotsenpiller G, Froestl W, Wolf ME. 2002. In vivo modulation of ventral tegmental area dopamine and glutamate efflux by local GABA(B) receptors is altered after repeated amphetamine treatment. *Neuroscience* 109:585-95
Grace AA, Bunney BS. 1980. Nigral dopamine neurons: intracellular recording and identification with L-dopa injection and histofluorescence. *Science (New York, N.Y* 210:654-6

Grace AA, Bunney BS. 1983. Intracellular and extracellular electrophysiology of nigral dopaminergic neurons--1. Identification and characterization. *Neuroscience* 10:301-15
Grace AA, Bunney BS. 1984. The control of firing pattern in nigral dopamine neurons: burst firing. *J Neurosci* 4:2877-90
Grace AA, Onn SP. 1989. Morphology and electrophysiological properties of immunocytochemically identified rat dopamine neurons recorded in vitro. *J Neurosci* 9:3463-81
Gratton A, Wise RA. 1985. Hypothalamic reward mechanism: two first-stage fiber populations with a cholinergic component. *Science* 227:545-8
Gronier B, Perry KW, Rasmussen K. 2000. Activation of the mesocorticolimbic dopaminergic system by stimulation of muscarinic cholinergic receptors in the ventral tegmental area. *Psychopharmacology* 147:347-55
Grossberg S, Bullock D, Dranias MR. 2008. Neural dynamics underlying impaired autonomic and conditioned responses following amygdala and orbitofrontal lesions. *Behavioral neuroscience* 122:1100-25
Guyenet PG, Aghajanian GK. 1978. Antidromic identification of dopaminergic and other output neurons of the rat substantia nigra. *Brain research* 150:69-84
Hakan RL, Henriksen SJ. 1989. Opiate influences on nucleus accumbens neuronal electrophysiology: dopamine and non-dopamine mechanisms. *J Neurosci* 9:3538-46
Hallam SC, Grahame NJ, Harris K, Miller RR. 1992. Associative Structures Underlying Enhanced Negative Summation Following Operational Extinction of a Pavlovian Inhibitor. *Learning and Motivation* 23:43-62
Hernandez L, Hoebel BG. 1988. Food reward and cocaine increase extracellular dopamine in the nucleus accumbens as measured by microdialysis. *Life sciences* 42:1705-12
Herve D, Blanc G, Glowinski J, Tassin JP. 1982. Reduction of dopamine utilization in the prefrontal cortex but not in the nucleus accumbens after selective destruction of noradrenergic fibers innervating the ventral tegmental area in the rat. *Brain research* 237:510-6
Herve D, Pickel VM, Joh TH, Beaudet A. 1987. Serotonin axon terminals in the ventral tegmental area of the rat: fine structure and synaptic input to dopaminergic neurons. *Brain research* 435:71-83
Holland PC. 1995. Transfer of occasion setting across stimulus and response in operant feature positive discriminations. *Learning and Motivation* 26:239-63
Holland PC. 1989. Occasion setting with simultaneous compounds in rats. *Journal of Experimental Psychology: Animal Behavior Processes* 15:183-93
Holland PC, Lamarre J. 1984. Transfer of inhibition after serial and simultaneous feature negative discriminative training. *Learning and Motivation* 15:219-43
Hollerman JR, Schultz W. 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature neuroscience* 1:304-9
Hollis KL. 1997. Contemporary research on Pavlovian conditioning: A "new" functional analysis. *American Psychologist* 52:956-65
Horvitz JC, Stewart T, Jacobs BL. 1997. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Res* 759:251-8
Hyland BI, Reynolds JNJ, Hay J, Perk CG, Miller R. 2002. Firing modes of midbrain dopamine cells in the freely moving rat. *Neuroscience* 114:475-92
Ikemoto S. 2007. Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. *Brain research reviews* 56:27-78

Ikemoto S, Panksepp J. 1996. Dissociations Between Appetitive and Consummatory Responses by Pharmacological Manipulations of Reward-Relevant Brain Regions. *Behavioral Neuroscience* 110:331-45
Ino T, Nakai R, Azuma T, Kimura T, Fukuyama H. 2009. Differential activation of the striatum for decision making and outcomes in a monetary task with gain and loss. *Cortex; a journal devoted to the study of the nervous system and behavior*
Johnson SW, North RA. 1992. Two types of neurone in the rat ventral tegmental area and their synaptic inputs. *The Journal of physiology* 450:455-68
Jones S, Kauer JA. 1999. Amphetamine depresses excitatory synaptic transmission via serotonin receptors in the ventral tegmental area. *J Neurosci* 19:9780-7
Julian MD, Martin AB, Cuellar B, Rodriguez De Fonseca F, Navarro M, et al. 2003. Neuroanatomical relationship between type 1 cannabinoid receptors and dopaminergic systems in the rat basal ganglia. *Neuroscience* 119:309-18
Kaczmarek HJ, Kiefer SW. 2000. Microinjections of dopaminergic agents in the nucleus accumbens affect ethanol consumption but not palatability. *Pharmacology, biochemistry, and behavior* 66:307-12
Kaelbling LP. 1993. *Learning in Embedded Systems*. Bradford: The MIT Press
Kalivas PW, Duffy P. 1995. D1 receptors modulate glutamate transmission in the ventral tegmental area. *J Neurosci* 15:5379-88
Kalos MH, Whitlock PA. 1986. *Monte Carlo Methods*: Wiley-YCH
Kamin, L.J. (1968). "Attention-like" processes in classical conditioning. In M.R. Jones (Ed.). Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation. Coral Gables, Florida: University of Miami Press (Pages 9-31).
Kampe KK, Frith CD, Dolan RJ, Frith U. 2001. Reward value of attractiveness and gaze. *Nature* 413:589
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
Kehoe EJ, White NE. 2004. Overexpectation: response loss during sustained stimulus compounding in the rabbit nictitating membrane preparation. *Learning & memory (Cold Spring Harbor, N.Y* 11:476-83
Kennedy RT, Jones SR, Wightman RM. 1992. Dynamic observation of dopamine autoreceptor effects in rat striatal slices. *Journal of neurochemistry* 59:449-55
Khallad Y, Moore J. 1996. Blocking, unblocking, and overexpectation in autoshaping with pigeons. *Journal of the experimental analysis of behavior* 65:575-91
Killcross S, Coutureau E. 2003. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb Cortex* 13:400-8
King AJ. 2004. The superior colliculus. *Curr Biol* 14:R335-8
Klink R, de Kerchove d'Exaerde A, Zoli M, Changeux JP. 2001. Molecular and physiological diversity of nicotinic acetylcholine receptors in the midbrain dopaminergic nuclei. *J Neurosci* 21:1452-63
Kobayashi S, Schultz W. 2008. Influence of reward delays on responses of dopamine neurons. *J Neurosci* 28:7837-46
Koga E, Momiyama T. 2000. Presynaptic dopamine D2-like receptors inhibit excitatory transmission onto rat ventral tegmental dopaminergic neurones. *The Journal of physiology* 523 Pt 1:163-73
Koob GF, Caine SB, Parsons L, Markou A, Weiss F. 1997. Opponent process model and psychostimulant addiction. *Pharmacology, biochemistry, and behavior* 57:513-21

Lacey MG. 1993. Neurotransmitter receptors and ionic conductances regulating the activity of neurones in substantia nigra pars compacta and ventral tegmental area. *Progress in brain research* 99:251-76
Lacey MG, Mercuri NB, North RA. 1989. Two cell types in rat substantia nigra zona compacta distinguished by membrane properties and the actions of dopamine and opioids. *J Neurosci* 9:1233-41
Lattal KM, Nakajima S. 1998. Overexpectation in appetitive Pavlovian and instrumental conditioning. *Animal Learning & Behavior* 26:351-60
Lavin A, Nogueira L, Lapish CC, Wightman RM, Phillips PE, Seamans JK. 2005. Mesocortical dopamine neurons operate in distinct temporal domains using multimodal signaling. *J Neurosci* 25:5013-23
Lee A, Wissekerke AE, Rosin DL, Lynch KR. 1998. Localization of alpha2C-adrenergic receptor immunoreactivity in catecholaminergic neurons in the rat central nervous system. *Neuroscience* 84:1085-96
Lennartz RC, Weinberger NM. 1992. Analysis of response systems in Pavlovian conditioning reveals rapidly versus slowly acquired conditioned responses: Support for two factors, implications for behavior and neurobiology. *Psychobiology* 20:93-119
Leyton M, Boileau I, Benkelfat C, Diksic M, Baker G, Dagher A. 2002. Amphetamine-induced increases in extracellular dopamine, drug wanting, and novelty seeking: a PET/[11C]raclopride study in healthy men. *Neuropsychopharmacology* 27:1027-35
Lind NM, Gjedde A, Moustgaard A, Olsen AK, Jensen SB, et al. 2005. Behavioral response to novelty correlates with dopamine receptor availability in striatum of Gottingen minipigs. *Behavioural brain research* 164:172-7
Linner L, Endersz H, Ohman D, Bengtsson F, Schalling M, Svensson TH. 2001. Reboxetine modulates the firing pattern of dopamine cells in the ventral tegmental area and selectively increases dopamine availability in the prefrontal cortex. *The Journal of pharmacology and experimental therapeutics* 297:540-6
Ljungberg T, Apicella P, Schultz W. 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of neurophysiology* 67:145-63
Lu XY, Churchill L, Kalivas PW. 1997. Expression of D1 receptor mRNA in projections from the forebrain to the ventral tegmental area. *Synapse (New York, N.Y* 25:205-14
Lubow RE, Moore AU. 1959. Latent inhibition: The effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of comparative and physiological psychology* 52:415-9
Lubow RE, Schnur P, Rifkin B. 1976. Latent inhibition and conditioned attention theory. *Journal of Experimental Psychology: Animal Behavior Processes* 2:163-74
Mackintosh NJ, Reese B. 1979. One-trial overshadowing. *Quarterly Journal of Experimental Psychology* 31:519-26
Margolis EB, Hjelmstad GO, Bonci A, Fields HL. 2005. Both kappa and mu opioid agonists inhibit glutamatergic input to ventral tegmental area neurons. *Journal of neurophysiology* 93:3086-93
Margolis EB, Lock H, Hjelmstad GO, Fields HL. 2006. The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *The Journal of physiology* 577:907-24

Martel P, Fantino M. 1996a. Influence of the amount of food ingested on mesolimbic dopaminergic system activity: a microdialysis study. *Pharmacology, biochemistry, and behavior* 55:297-302
Martel P, Fantino M. 1996b. Mesolimbic dopaminergic system activity as a function of food reward: a microdialysis study. *Pharmacology, biochemistry, and behavior* 53:221-6
Mathon DS, Kamal A, Smidt MP, Ramakers GM. 2003. Modulation of cellular activity and synaptic transmission in the ventral tegmental area. *European journal of pharmacology* 480:97-115
McClure SM, Daw ND, Montague PR. 2003. A computational substrate for incentive salience. *Trends Neurosci* 26:423-8
McNally GP, Pigg M, Weidemann G. 2004. Blocking, unblocking, and overexpectation of fear: a role for opioid receptors in the regulation of Pavlovian association formation. *Behavioral neuroscience* 118:111-20
McRitchie DA, Hardman CD, Halliday GM. 1996. Cytoarchitectural distribution of calcium binding proteins in midbrain dopaminergic regions of rats and humans. *The Journal of comparative neurology* 364:121-50
Menon V, Levitin DJ. 2005. The rewards of music listening: response and physiological connectivity of the mesolimbic system. *NeuroImage* 28:175-84
Millan MJ, Lejeune F, Gobert A, Brocco M, Auclair A, et al. 2000. S18616, a highly potent spiroimidazoline agonist at alpha(2)-adrenoceptors: II. Influence on monoaminergic transmission, motor function, and anxiety in comparison with dexmedetomidine and clonidine. *The Journal of pharmacology and experimental therapeutics* 295:1206-22
Miller RR, Barnet RC, Grahame NJ. 1995. Assessment of the Rescorla-Wagner model. *Psychological bulletin* 117:363-86
Mirenowicz J, Schultz W. 1996. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* 379:449-51
Miyata K, Kamato T, Yamano M, Nishida A, Ito H, et al. 1991. Serotonin (5-HT)3 receptor blocking activities of YM060, a novel 4,5,6,7-tetrahydrobenzimidazole derivative, and its enantiomer in anesthetized rats. *The Journal of pharmacology and experimental therapeutics* 259:815-9
Mobbs D, Greicius MD, Abdel-Azim E, Menon V, Reiss AL. 2003. Humor modulates the mesolimbic reward centers. *Neuron* 40:1041-8
Montague PR, Berns GS. 2002. Neural economics and the biological substrates of valuation. *Neuron* 36:265-84
Montague PR, Dayan P, Person C, Sejnowski TJ. 1995. Bee foraging in uncertain environments using predictive hebbian learning. *Nature* 377:725-8
Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936-47
Montague PR, Hyman SE, Cohen JD. 2004a. Computational roles for dopamine in behavioural control. *Nature* 431:760-7
Montague PR, McClure SM, Baldwin PR, Phillips PE, Budygin EA, et al. 2004b. Dynamic gain control of dopamine delivery in freely moving animals. *J Neurosci* 24:1754-9
Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. 2006. Midbrain dopamine neurons encode decisions for future action. *Nature neuroscience* 9:1057-63
Nauta WJ. 1958. Hippocampal projections and related neural pathways to the midbrain in the cat. *Brain* 81:319-40

Navarro M, Carrera MR, Fratta W, Valverde O, Cossu G, et al. 2001. Functional interaction between opioid and cannabinoid receptors in drug self-administration. *J Neurosci* 21:5344-50
Nestler EJ, Carlezon JWA. 2006. The Mesolimbic Dopamine Reward Circuit in Depression. *Biological Psychiatry* 59:1151-9
Numan M. 2007. Motivational systems and the neural circuitry of maternal behavior in the rat. *Developmental psychobiology* 49:12-21
O'Brien CP. 2001. Drug addicition and drug abuse. *Goodman and Gilman's. Pharmacol. Basis Ther.* :621–42
Oades RD, Halliday GM. 1987. Ventral tegmental (A10) system: neurobiology. 1. Anatomy and connectivity. *Brain research* 434:117-65
Oakman SA, Faris PL, Kerr PE, Cozzari C, Hartman BK. 1995. Distribution of pontomesencephalic cholinergic neurons projecting to substantia nigra differs significantly from those projecting to ventral tegmental area. *J Neurosci* 15:5859-69
Olds ME. 1982. Reinforcing effects of morphine in the nucleus accumbens. *Brain Res* 237:429-40
Omelchenko N, Sesack SR. 2006. Cholinergic axons in the rat ventral tegmental area synapse preferentially onto mesoaccumbens dopamine neurons. *The Journal of comparative neurology* 494:863-75
Palmerino CC, Rusiniak KW, Garcia J. 1980. Flavor-illness aversions: The peculiar roles of odor and taste in memory for poison. *Science* 208:753-5
Pan WX, Schmidt R, Wickens JR, Hyland BI. 2005. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J Neurosci* 25:6235-42
Pan WX, Schmidt R, Wickens JR, Hyland BI. 2008. Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *J Neurosci* 28:9619-31
Pavlov IP. 1927. *Conditioned Reflexes.* . London: Oxford Univ. Press
Pavlov IP, Gantt WH, Cannon WB. 1928. *Lectures on conditioned reflexes. Twenty-five years of objective study of the higher nervous activity (behavior) of animals*. Lectures on conditioned reflexes. Twenty-five years of objective study of the higher nervous activity (behavior) of animals. Pp. 414 pp. Oxford, England: International International Print
Pearce JM. 1987. A model for stimulus generalization in Pavlovian conditioning. *Psychological review* 94:61-73
Pearce JM. 1994. Similarity and discrimination: A selective review and a connectionist model. *Psychological review* 101:587-607
Pearce JM, Bouton ME. 2001. Theories of associative learning in animals. *Annual review of psychology* 52:111-39
Pecina S, Berridge KC, Parker LA. 1997. Pimozide does not shift palatability: separation of anhedonia from sensorimotor suppression by taste reactivity. *Pharmacology, biochemistry, and behavior* 58:801-11
Pecina S, Cagniard B, Berridge KC, Aldridge JW, Zhuang X. 2003. Hyperdopaminergic mutant mice have higher "wanting" but not "liking" for sweet rewards. *J Neurosci* 23:9395-402
Pessia M, Jiang ZG, North RA, Johnson SW. 1994. Actions of 5-hydroxytryptamine on ventral tegmental area neurons of the rat in vitro. *Brain research* 654:324-30
Phillips PE, Stuber GD, Heien ML, Wightman RM, Carelli RM. 2003. Subsecond dopamine release promotes cocaine seeking. *Nature* 422:614-8

Pickel VM, Chan J, Kash TL, Rodriguez JJ, MacKie K. 2004. Compartment-specific localization of cannabinoid 1 (CB1) and mu-opioid receptors in rat nucleus accumbens. *Neuroscience* 127:101-12

Pierce RC, Kumaresan V. 2006. The mesolimbic dopamine system: The final common pathway for the reinforcing effect of drugs of abuse? *Neuroscience & Biobehavioral Reviews* 30:215-38

Pillolla G, Melis M, Perra S, Muntoni AL, Gessa GL, Pistis M. 2007. Medial forebrain bundle stimulation evokes endocannabinoid-mediated modulation of ventral tegmental area dopamine neuron firing in vivo. *Psychopharmacology* 191:843-53

Pontieri FE, Tanda G, Orzi F, Di Chiara G. 1996. Effects of nicotine on the nucleus accumbens and similarity to those of addictive drugs. *Nature* 382:255-7

Prisco S, Pagannone S, Esposito E. 1994. Serotonin-dopamine interaction in the rat ventral tegmental area: an electrophysiological study in vivo. *The Journal of pharmacology and experimental therapeutics* 271:83-90

Quertemont E, De Witte P. 2001. Conditioned stimulus preference after acetaldehyde but not ethanol injections. *Pharmacology, biochemistry, and behavior* 68:449-54

Rada P, Moreno SA, Tucci S, Gonzalez LE, Harrison T, et al. 2003. Glutamate release in the nucleus accumbens is involved in behavioral depression during the PORSOLT swim test. *Neuroscience* 119:557-65

Rahman S, McBride WJ. 2001. D1-D2 dopamine receptor interaction within the nucleus accumbens mediates long-loop negative feedback to the ventral tegmental area (VTA). *Journal of neurochemistry* 77:1248-55

Redgrave P, Gurney K, Reynolds J. 2007. What is reinforced by phasic dopamine signals? *Brain Res Rev*

Rescorla RA. 1968. Probability of Shock in the Presence and Absence of Cs in Fear Conditioning. *Journal of comparative and physiological psychology* 66:1-5

Rescorla RA. 1973. Second-order conditioning: Implications for theories of learning. In *McGuigan, F. J; Lumsden, D. Barry. (1973). Contemporary approaches to conditioning and learning. xii, 321 pp. Oxford, England: V. H. Winston & Sons. V. H. Winston & Sons Print*

Rescorla RA. 1987. A Pavlovian analysis of goal-directed behavior. *American Psychologist* 42:119-29

Rescorla RA. 1988. Behavioral studies of Pavlovian conditioning. *Annual review of neuroscience* 11:329-52

Rescorla RA. 2004. Spontaneous recovery varies inversely with the training-extinction interval. *Learn Behav* 32:401-8

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. . In *Classical conditioning II*, ed. BWF Prokasy, pp. 64-99. New York: Appleton-Century-Crofts

Richards CD, Shiroyama T, Kitai ST. 1997. Electrophysiological and immunocytochemical characterization of GABA and dopamine neurons in the substantia nigra of the rat. *Neuroscience* 80:545-57

Riederer P, Gerlach M, Müller T, Reichmann H. 2007. Relating mode of action to clinical practice: Dopaminergic agents in Parkinson's disease. *Parkinsonism & Related Disorders* 13:466-79

Rodd ZA, Melendez RI, Bell RL, Kuc KA, Zhang Y, et al. 2004. Intracranial self-administration of ethanol within the ventral tegmental area of male Wistar rats: evidence for involvement of dopamine neurons. *J Neurosci* 24:1050-7

Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM. 2004. Dopamine operates as a subsecond modulator of food seeking. *J Neurosci* 24:1265-71
Roitman MF, Wheeler RA, Carelli RM. 2005. Nucleus accumbens neurons are innately tuned for rewarding and aversive taste stimuli, encode their predictors, and are linked to motor output. *Neuron* 45:587-97
Romo R, Herna´ndez A, Zainos A, Lemus L, Brody CD 2002. Neuronal correlates of decision- making in secondary somatosensory cortex. *Nat Neurosci* 5:1217–1225.
Ross RT. 1983. Relationships between the determinants of performance in serial feature-positive discriminations. *Journal of Experimental Psychology: Animal Behavior Processes* 9:349-73
Ross RT, LoLordo VM. 1986. Blocking during serial feature-positive discriminations: Associative versus occasion-setting functions. *Journal of Experimental Psychology: Animal Behavior Processes* 12:315-24
Rumbaugh DM, Richardson WK, Washburn DA, Savage-Rumbaugh ES, Hopkins WD. 1989. Rhesus monkeys (Macaca mulatta), video tasks, and implications for stimulus-response spatial contiguity. *Journal of Comparative Psychology* 103:32-8
Russo SJ, Sun WL, Minerly AC, Weierstall K, Nazarian A, et al. 2008. Progesterone attenuates cocaine-induced conditioned place preference in female rats. *Brain research* 1189:229-35
Salinas E, Hernandez A, Zainos A,RomoR 2000. Periodicity and firing rate as candidate neural codes for the frequency of vibrotactile stimuli. *J Neurosci* 20:5503–5515.
Sanchis-Segura C, Spanagel R. 2006. Behavioural assessment of drug reinforcement and addictive features in rodents: an overview. *Addiction biology* 11:2-38
Satoh T, Nakai S, Sato T, Kimura M. 2003. Correlated coding of motivation and outcome of decision by dopamine neurons. *J Neurosci* 23:9913-23
Schilstrom B, Fagerquist MV, Zhang X, Hertel P, Panagis G, et al. 2000. Putative role of presynaptic alpha7* nicotinic receptors in nicotine stimulated increases of extracellular levels of glutamate and aspartate in the ventral tegmental area. *Synapse (New York, N.Y* 38:375-83
Schmajuk NA, Buhusi CV. 1997. Stimulus configuration, occasion setting, and the hippocampus. *Behav Neurosci* 111:235-57; appendix 58
Schmajuk NA, Lamoureux JA, Holland PC. 1998. Occasion setting: a neural network approach. *Psychol Rev* 105:3-32
Schreurs BG, Kehoe EJ, Gormezano I. 1993. Concurrent associative transfer and competition in serial conditioning of the rabbit's nictitating membrane response. *Learning and Motivation* 24:395-412
Schultz W. 1998. Predictive reward signal of dopamine neurons. *Journal of neurophysiology* 80:1-27
Schultz W. 1999. The Reward Signal of Midbrain Dopamine Neurons. *News Physiol Sci* 14:249-55
Schultz W. 2001. Reward signaling by dopamine neurons. *Neuroscientist* 7:293-302
Schultz W. 2002. Getting formal with dopamine and reward. *Neuron* 36:241-63
Schultz W. 2004. Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current opinion in neurobiology* 14:139-47
Schultz W. 2007a. Behavioral dopamine signals. *Trends in neurosciences* 30:203-10
Schultz W. 2007b. Multiple Dopamine Functions at Different Time Courses. *Annual review of neuroscience* 30:259-88

Schultz W, Apicella P, Ljungberg T. 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J Neurosci* 13:900-13
Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science* 275:1593-9
Schultz W, Romo R. 1987. Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *Journal of neurophysiology* 57:201-17
Seeman P, Weinshenker D, Quirion R, Srivastava LK, Bhardwaj SK, et al. 2005. Dopamine supersensitivity correlates with D2High states, implying many paths to psychosis. *Proceedings of the National Academy of Sciences of the United States of America* 102:3513-8
Singer G, Wallace M. 1984. Effects of 6-OHDA lesions in the nucleus accumbens on the acquisition of self injection of heroin under schedule and non schedule conditions in rats. *Pharmacology, biochemistry, and behavior* 20:807-9
Singh SP, Sutton RS. 1996. Reinforcement learning with replacing eligibility traces. *Machine Learning* 22:123-58
Small DM, Jones-Gotman M, Dagher A. 2003. Feeding-induced dopamine release in dorsal striatum correlates with meal pleasantness ratings in healthy human volunteers. *Neuroimage* 19:1709-15
Staddon JE, Cerutti DT. 2003. Operant conditioning. *Annual review of psychology* 54:115-44
Stout S, Arcediano F, Escobar M, Miller RR. 2003. Overshadowing as a function of trial number: Dynamics of first- and second-order comparator effects. *Learning & Behavior* 31:85-97
Sulzer D, Joyce MP, Lin L, Geldwert D, Haber SN, et al. 1998. Dopamine neurons make glutamatergic synapses in vitro. *J Neurosci* 18:4588-602
Suri RE. 2002. TD models of reward predictive responses in dopamine neurons. *Neural Netw* 15:523-33
Suri RE, Schultz W. 1998a. Learning of sequential movements by neural network model with dopamine- like reinforcement signal. *Exp Brain Res* 121:350-4
Suri RE, Schultz W. 1998b. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental brain research. Experimentelle Hirnforschung* 121:350-4
Suri RE, Schultz W. 1999. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* 91:871-90
Suri RE, Schultz W. 2001a. Temporal difference model reproduces anticipatory neural activity. *Neural computation* 13:841-62
Suri RE, Schultz W. 2001b. Temporal difference model reproduces anticipatory neural activity. *Neural Comput* 13:841-62.
Sutton RS, Barto AG. 1998. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press
Swanson LW. 1982. The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain research bulletin* 9:321-53
Takahashi YK, Roesch MR, Stalnaker TA, Haney RZ, Calu DJ, et al. 2009. The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron* 62:269-80
Takahata R, Moghaddam B. 1998. Glutamatergic regulation of basal and stimulus-activated dopamine release in the prefrontal cortex. *Journal of neurochemistry* 71:1443-9

Takikawa Y, Kawagoe R, Hikosaka O. 2004. A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of neurophysiology* 92:2520-9

Tepper JM, Martin LP, Anderson DR. 1995. GABAA receptor-mediated inhibition of rat substantia nigra dopaminergic neurons by pars reticulata projection neurons. *J Neurosci* 15:3092-103

Tesauro G. 1992. Practical Issues in Temporal Difference Learning. *Advances in Neural Information Processing Systems 4* 4:259-66

Thompson L, Barraud P, Andersson E, Kirik D, Bjorklund A. 2005. Identification of dopaminergic neurons of nigral and ventral tegmental area subtypes in grafts of fetal ventral mesencephalon based on cell morphology, protein expression, and efferent projections. *J Neurosci* 25:6467-77

Thorndike E. 1898. Some Experiments on Animal Intelligence. *Science (New York, N.Y* 7:818-24

Tobler PN, Dickinson A, Schultz W. 2003. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *J Neurosci* 23:10402-10

Tobler PN, Fiorillo CD, Schultz W. 2005. Adaptive coding of reward value by dopamine neurons. *Science* 307:1642-5

Tsai C. 1925. The optic racts and centers of the opossum, didelphis virginiana. *J. Comp. Neurol.* :173-216

Tsitolovsky L, Babkina N, Shvedov A. 2004. A comparison of neuronal reactions during classical and instrumental conditioning under similar conditions. *Neurobiology of learning and memory* 81:82-95

Ungless MA, Magill PJ, Bolam JP. 2004. Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science (New York, N.Y* 303:2040-2

Urcelay GP, Miller RR. 2009. Potentiation and Overshadowing in Pavlovian Fear Conditioning. *Journal of Experimental Psychology-Animal Behavior Processes* 35:340-56

Volkow ND, Wang GJ, Fowler JS. 1997. Imaging studies of cocaine in the human brain and studies of the cocaine addict. *Annals of the New York Academy of Sciences* 820:41-54; discussion -5

Waelti P, Dickinson A, Schultz W. 2001. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412:43-8

Wagner AR, Rescorla RA. 1972. Inhibition in Pavlovian conditioning: appllication of a theory. In *Inhibition and learning*, ed. R Boakes, pp. 301-26. London: MS Halliday

Wallace DM, Magnuson DJ, Gray TS. 1992. Organization of amygdaloid projections to brainstem dopaminergic, noradrenergic, and adrenergic cell groups in the rat. *Brain research bulletin* 28:447-54

Wamsley JK, Gehlert DR, Filloux FM, Dawson TM. 1989. Comparison of the distribution of D-1 and D-2 dopamine receptors in the rat brain. *Journal of chemical neuroanatomy* 2:119-37

Wan X, Peoples LL. 2006. Firing patterns of accumbal neurons during a pavlovian-conditioned approach task. *Journal of neurophysiology* 96:652-60

Wasserman EA, Carr DL, Deich JD. 1978. Association of conditioned stimuli during serial conditioning by pigeons. Year of Publication 1978. *Animal Learning & Behavior* 6:52-6

Wasserman EA, Miller RR. 1997. What's elementary about associative learning? *Annual review of psychology* 48:573-607

Weiss F, Lorang MT, Bloom FE, Koob GF. 1993. Oral alcohol self-administration stimulates dopamine release in the rat nucleus accumbens: genetic and motivational determinants. *The Journal of pharmacology and experimental therapeutics* 267:250-8

Wilkenfield J, Nickel M, Blakely E, Poling A. 1992. Acquisition of lever-press responding in rats with delayed reinforcement: A comparison of three procedures. *J Exp Anal Behav* 58:431-43

Wilson CJ, Young SJ, Groves PM. 1977. Statistical properties of neuronal spike trains in the substantia nigra: cell types and their interactions. *Brain Res* 136:243-60

Wilson DI, Bowman EM. 2006. Neurons in dopamine-rich areas of the rat medial midbrain predominantly encode the outcome-related rather than behavioural switching properties of conditioned stimuli. *Eur J Neurosci* 23:205-18

Winn P. 2008. Experimental studies of pedunculopontine functions: are they motor, sensory or integrative? *Parkinsonism & related disorders* 14 Suppl 2:S194-8

Wise RA, Bozarth MA. 1985. Brain mechanisms of drug reward and euphoria. *Psychiatric medicine* 3:445-60

Wise RA, Newton P, Leeb K, Burnette B, Pocock D, Justice JB, Jr. 1995. Fluctuations in nucleus accumbens dopamine concentration during intravenous cocaine self-administration in rats. *Psychopharmacology (Berl)* 120:10-20

Wise RA, Spindler J, deWit H, Gerberg GJ. 1978. Neuroleptic-induced "anhedonia" in rats: pimozide blocks reward quality of food. *Science* 201:262-4

Wonnacott S, Sidhpura N, Balfour DJ. 2005. Nicotine: from molecular mechanisms to behaviour. *Current opinion in pharmacology* 5:53-9

Wyvell CL, Berridge KC. 2000. Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *J Neurosci* 20:8122-30

Wyvell CL, Berridge KC. 2001. Incentive sensitization by previous amphetamine exposure: increased cue-triggered "wanting" for sucrose reward. *J Neurosci* 21:7831-40

Yeomans J, Forster G, Blaha C. 2001. M5 muscarinic receptors are needed for slow activation of dopamine neurons and for rewarding brain stimulation. *Life sciences* 68:2449-56

Yin HH, Knowlton BJ, Balleine BW. 2004. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *The European journal of neuroscience* 19:181-9

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. 2005. The role of the dorsomedial striatum in instrumental conditioning. *The European journal of neuroscience* 22:513-23

Zheng F, Johnson SW. 2002. Group I metabotropic glutamate receptor-mediated enhancement of dopamine cell burst firing in rat ventral tegmental area in vitro. *Brain research* 948:171-4

Zheng H, Patterson LM, Berthoud HR. 2007. Orexin signaling in the ventral tegmental area is required for high-fat appetite induced by opioid stimulation of the nucleus accumbens. *J Neurosci* 27:11075-82

Project: Montague et al model IImacro.rbp
Date: Friday, 17 July 2009 17:02:35

Project Info:

Mac (Carbon PEF) App Name: Montague et al model
Mac (Carbon Mach-O) App Name: Montague et al model
Mac (Classic) App Name: Montague et al model
Windows App Name: Montague et al model.exe
Linux App Name: Montague et al model
Long Version:
Major Version: 2
Minor Version: 0
Sub Version: 0
Release: 3
Non-Release: 0
Mac Creator Code: thx
Windows MDI Caption: Montague model
Minimum Memory Size: 2048
Standard Memory Size: 4096

Class RLwindow

Inherits Window

RLwindow.EnableMenuItems:

```
Sub EnableMenuItems()
   ModelEditParameters.enabled=true
   ModelLoadparameters.enabled=true
End Sub
```

RLwindow.Open:

```
Sub Open()
   EnableMenuItems
   RLwindow.MaxHeight=Screen(0).width-100
   RLwindow.MaxHeight=Screen(0).height-100
   RLwindow.top=50
   RLwindow.left=50
   RLwindow.height=RLwindow.MaxHeight
   RLwindow.width=RLwindow.MaxWidth

   GraphTypePopup.Enabled=False
   GraphItemPopup.Enabled=False
```

```
  GraphButton.Enabled=False

  ResizeGraphStuff
End Sub
```

RLwindow.Resized:

```
Sub Resized()
  ResizeGraphStuff
End Sub
```

RLwindow.ModelEditParameters:

```
Function ModelEditParameters() As Boolean
  Pwindow.Show
End Function
```

RLwindow.ModelLoadparameters:

```
Function ModelLoadparameters() As Boolean
  DIm TextFile As FolderItem
  Dim TextStream As TextInputStream
  Dim TextType as New FileType
  Dim TextLine,ParsedText(-1),DelimiterString As String
  Dim FieldCount,i As Integer
  Dim RanOK As Boolean



  TextType.Name="Plain text"
  TextType.MacType="TEXT"
  TextType.Extensions="txt"

  DelimiterString=chr(9)

  TextFile= GetOpenFolderItem(TextType)
  If TextFile<> Nil then
    TextStream=TextFile.OpenAsTextFile
    Pwindow.ParametersPopup.ListIndex=3
    Pwindow.DefineSequence.DeleteAllRows
    Do
      TextLine=TextStream.ReadLine
      ParsedText=Split(TextLine,DelimiterString)
      FieldCount=CountFields(TextLine,DelimiterString)
      Select Case ParsedText(0)
      Case "ModelName"
        Pwindow.ModelName.Text=ParsedText(1)
      case "Plot"
        if ParsedText(1)="False" then
```

```
      RLwindow.SupressPlot=True
    end if
  Case "Trials"
    Pwindow.TrialN.Text=ParsedText(1)
  Case "Ticks"
    Pwindow.TrialLength.Text=ParsedText(1)
  Case "Stimuli"
    Pwindow.StimuliN.Text=ParsedText(1)
  Case "GradedInputs"
    If ParsedText(1)="True" then
      Pwindow.GradualOnOffCheckBox.Value=True
    else
      Pwindow.GradualOnOffCheckBox.Value=False
    end
  Case "LearningRate"
    Pwindow.LearningRate.Text=ParsedText(1)
  Case "TemporalDiscount"
    Pwindow.TemporalDiscount.Text=ParsedText(1)
  Case "TraceRise"
    Pwindow.InputOnsetConstant.Text=ParsedText(1)
  Case "TraceDecay"
    Pwindow.InputOffsetConstant.Text=ParsedText(1)
  Case "Length"
    Pwindow.TrialLength.Text=ParsedText(1)
  Case "DAminimum"
    Pwindow.DAminimum.Text=ParsedText(1)
  case "TrialShuffle"
    Pwindow.ShuffleTrialsCheckBox.Value=True
  Case "StartShuffle"
    Pwindow.StartTrialShuffleEditField.Text=ParsedText(1)
  Case "EndShuffle"
    Pwindow.EndTrialShuffleEditField.Text=ParsedText(1)
  Case "Run"
    App.UsedMacro=True
    App.MacroSaveDataDirectory=TextFile.Parent
    Pwindow.SaveOutput.Value=True
    Pwindow.OKclose=True
    Pwindow.Show
    Pwindow.AllDone
    RLwindow.Show
    RanOK=RunModel
    If RanOK=False then
      MsgBox("Error in running model from macro")
    end if
    Pwindow.ParametersPopup.ListIndex=3
    Pwindow.DefineSequence.DeleteAllRows
```

```
        RLwindow.SupressPlot=False
      Case else
        If FieldCount=6 then
          Pwindow.DefineSequence.AddRow("")
          for i=0 to 5
            Pwindow.DefineSequence.Cell(Pwindow.DefineSequence.LastIndex,i)
            =ParsedText(i)
          next
        End if
      End
    Loop Until TextStream.EOF
    TextStream.Close
    if App.UsedMacro=False then
      Pwindow.ModelName.Text=NthField(TextFile.Name,".",1)
      Pwindow.Show
    end if
  End if
End Function
```

RLwindow.ModelRun:

```
Function ModelRun() As Boolean
  Dim OK As Boolean

  OK=RunModel
  if OK=False then
    MsgBox("Error in running the model")
  end if

  return OK

End Function
```

RLwindow.ReferencesShowreferences:

```
Function ReferencesShowreferences() As Boolean
  MsgBox "The model in this program is derived from the following references:"+chr(13)
  +chr(13)+"Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic
  dopamine systems based on predictive Hebbian learning. J Neurosci 16: 1936-47"+chr
  (13)+chr(13)+"Egelman DM, Person C, Montague PR. 1998. A computational role for
  dopamine delivery in human decision-making. J Cogn Neurosci 10: 623-30"+chr(13)
  +chr(13)+"Pan WX, Schmidt R, Wickens JR, Hyland BI. 2005. Dopamine cells respond to
  predicted events during classical conditioning: evidence for eligibility traces in the
  reward-learning network. J Neurosci 25: 6235-42"
End Function
```

RLwindow.InitRamp:

```
Sub InitRamp()
```

```
  Dim x As Integer, midramp As Integer, maxramp As Integer, stepsway As Integer

  maxramp=UBound(GraphRamp)
  midramp=maxramp/2
  for x=0 to maxramp
    if x<=midramp then
      stepsway=midramp-x
      GraphRamp(x)=RGB(255-stepsway,255-stepsway,255)
    else
      stepsway=x-midramp
      GraphRamp(x)=RGB(255,255-stepsway,255-stepsway)
    end if
  next

End Sub
```

RLwindow.ResizeGraphStuff:

```
Sub ResizeGraphStuff()
  GraphTypePopup.Left=20
  GraphTypePopup.Top=20
  GraphTypePopup.Height=20
  GraphTypePopup.Width=175

  GraphItemPopup.Left=GraphTypePopup.Left+GraphTypePopup.Width+20
  GraphItemPopup.Top=20
  GraphItemPopup.Height=20
  GraphItemPopup.Width=50

  GraphButton.Left=GraphItemPopup.Left+GraphItemPopup.Width+20
  GraphButton.Top=20
  GraphButton.Height=20
  GraphButton.Width=45

  GraphCanvas.Top=GraphTypePopup.Top+GraphTypePopup.Height+20
  GraphCanvas.Left=20
  GraphCanvas.Width=Me.Width-40
  GraphCanvas.Height=Me.Height-GraphCanvas.Top-20
  GraphCanvas.Refresh
End Sub
```

RLwindow.CalculateData:

```
Sub CalculateData()
  Dim trial,tick,totalticks As  Integer
  Dim datum As Double
```

```
  totalticks=App.MaxTrials*App.TrialLength
  for trial=1 to App.MaxTrials
    for tick=1 to App.TrialLength
      Presynaptic(trial,tick)
      TD(trial,tick)
      Dopamine(trial,tick)
      UpdateWeights(trial,tick)
    next
  next
End Sub
```

RLwindow.Presynaptic:

```
Sub Presynaptic(trial As Integer, tick As Integer)
  Dim i as integer

  PresynapticActivity(trial,tick)=0
  for i=1 to App.InputN
    WeightedInputsData(trial,tick,i)=EligibilityTrace(trial,tick,i)*WeightsData(trial,tick,i)
    PresynapticActivity(trial,tick)=PresynapticActivity(trial,tick)+WeightedInputsData
    (trial,tick,i)
  next
End Sub
```

RLwindow.TD:

```
Sub TD(trial as integer, tick as integer)
  if tick=1 then
    PostsynapticActivity(trial,tick)=0
  else
    PostsynapticActivity(trial,tick)=PresynapticActivity(trial,tick)-
    (App.TemporalDiscount*PresynapticActivity(trial,tick-1))
  end if

End Sub
```

RLwindow.Dopamine:

```
Sub Dopamine(trial as integer, tick as integer)
  dim threshold as double
  DopamineData(trial,tick)=RewardData(trial,tick)+PostsynapticActivity(trial,tick)
  if IsNumeric(App.MinDA) then
    threshold=val(App.MinDA)
    if DopamineData(trial,tick)<threshold then
      DopamineData(trial,tick)=threshold
    end
  end
End Sub
```

RLwindow.UpdateWeights:

```
Sub UpdateWeights(trial as integer, tick as integer)
  dim i,j as integer
  dim newWeight as double

  if tick>1 then
    for i=1 to App.InputN
      'newWeight=WeightsData(trial,tick-1,i)+(App.LearnRate*InputsData(trial,tick-1,i)
      *DopamineData(trial,tick))
      newWeight=WeightsData(trial,tick-1,i)+(App.LearnRate*EligibilityTrace
      (trial,tick-1,i)*DopamineData(trial,tick))
      DeltaWeightsData(trial,tick-1,i)=newWeight-WeightsData(trial,tick-1,i)
      WeightsData(trial+1,tick-1,i)=newWeight
    next
  end if

End Sub
```

RLwindow.SaveData:

```
Sub SaveData()
  Dim f as FolderItem
  Dim stream as TextOutputStream
  Dim FileName As string
  dim i,j,s As Integer

  FileName=App.ModelName
  Me.MouseCursor= System.Cursors.StandardPointer
  Me.UpdateNow
  if App.UsedMacro=False then
    f=GetSaveFolderItem("TEXT",FileName+"_output.txt")
  else
    f=App.MacroSaveDataDirectory.Child(App.ModelName+".txt")
  end if
  if f<> Nil then
    Me.MouseCursor= System.Cursors.Wait
    stream=f.CreateTextFile
    f.MacCreator="PSYr"
    'Header
    Stream.Write "Trial"+chr(9)
    Stream.Write "Tick"+chr(9)
    Stream.Write "LearningRate"+chr(9)
    Stream.Write "TemporalDiscount"+chr(9)
    If App.UseGradedTrace=True then
      stream.Write "TraceOnset"+chr(9)
      Stream.Write "TraceDecay"+chr(9)
```

```
    end
    if IsNumeric(App.MinDA)=True then Stream.Write "MinDA"+chr(9)
    for i=1 to App.InputN
      Stream.Write "S"+str(i)+chr(9)
    next
    for i=1 to App.InputN
      Stream.Write "T"+str(i)+chr(9)
    next
    Stream.Write "R"+chr(9)
    for i=1 to App.InputN
      Stream.Write "W"+str(i)+chr(9)
    next
    Stream.Write "V"+chr(9)
    Stream.Writeline "DA"
    'Data
    for i=1 to App.MaxTrials
      for j=1 to App.TrialLength
        Stream.Write str(i)+chr(9)
        Stream.Write str(j)+chr(9)
        Stream.Write str(App.LearnRate)+chr(9)
        Stream.Write str(App.TemporalDiscount)+chr(9)
        if App.UseGradedTrace=True then
          Stream.Write str(App.TraceRise)+chr(9)
          Stream.Write str(App.TraceDecay)+chr(9)
        end
        if IsNumeric(App.MinDA)=True then Stream.Write App.MinDA+chr(9)
        for s=1 to App.InputN
          Stream.Write str(InputsData(i,j,s))+chr(9)
        next
        for s=1 to App.InputN
          Stream.Write str(EligibilityTrace(i,j,s))+chr(9)
        next
        Stream.Write str(RewardData(i,j))+chr(9)
        for s=1 to App.InputN
          Stream.Write str(WeightsData(i,j,s))+chr(9)
        next
        Stream.Write str(PostsynapticActivity(i,j))+chr(9)
        Stream.Writeline str(DopamineData(i,j))
      next
    next
    Stream.Close
  end

  Me.MouseCursor= System.Cursors.Wait

End Sub
```

RLwindow.PlotGraph:

```
Sub PlotGraph()
  Dim trial, tick,stimulus As Integer

  Me.MouseCursor= System.Cursors.Wait
  AbsGraphZMax=0
  select case GraphTypePopup.ListIndex
  case 0 'Dopamine
    for trial=1 to App.MaxTrials
      for tick=1 to App.TrialLength
        GraphData(trial,tick)=DopamineData(trial,tick)
        If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
        (GraphData(trial,tick))
      next
    next
    GraphCanvas.Refresh
  case 1 'Inputs
    if IsNumeric(GraphItemPopup.Text) then
      stimulus=GraphItemPopup.Text.Val
      for trial=1 to App.MaxTrials
        for tick=1 to App.TrialLength
          GraphData(trial,tick)=InputsData(trial,tick,stimulus)
          If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
          (GraphData(trial,tick))
        next
      next
      GraphCanvas.Refresh
    else
      MsgBox "No input selected. Please select the number of the input and push the
      button again."
    end if
  case 2 'Weights
    if IsNumeric(GraphItemPopup.Text) then
      stimulus=GraphItemPopup.Text.Val
      for trial=1 to App.MaxTrials
        for tick=1 to App.TrialLength
          GraphData(trial,tick)=WeightsData(trial,tick,stimulus)
          If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
          (GraphData(trial,tick))
        next
      next
      GraphCanvas.Refresh
    else
      MsgBox "No input selected. Please select the number of the input and push the
      button again."
```

```
  end if
case 3 'Weighted input
  if IsNumeric(GraphItemPopup.Text) then
    stimulus=GraphItemPopup.Text.Val
    for trial=1 to App.MaxTrials
      for tick=1 to App.TrialLength
        GraphData(trial,tick)=WeightedInputsData(trial,tick,stimulus)
        If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
        (GraphData(trial,tick))
      next
    next
    GraphCanvas.Refresh
  else
    MsgBox "No input selected. Please select the number of the input and push the
    button again."
  end if
case 4 'Eligibility trace
  if IsNumeric(GraphItemPopup.Text) then
    stimulus=GraphItemPopup.Text.Val
    for trial=1 to App.MaxTrials
      for tick=1 to App.TrialLength
        GraphData(trial,tick)=EligibilityTrace(trial,tick,stimulus)
        If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
        (GraphData(trial,tick))
      next
    next
    GraphCanvas.Refresh
  else
    MsgBox "No input selected. Please select the number of the input and push the
    button again."
  end if
case 5 'Presynaptic activity on intermediate layer
  for trial=1 to App.MaxTrials
    for tick=1 to App.TrialLength
      GraphData(trial,tick)=PresynapticActivity(trial,tick)
      If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
      (GraphData(trial,tick))
    next
  next
  GraphCanvas.Refresh
case 6 'Postsynaptic activity of intermediate layer
  for trial=1 to App.MaxTrials
    for tick=1 to App.TrialLength
      GraphData(trial,tick)=PostsynapticActivity(trial,tick)
      If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
      (GraphData(trial,tick))
```

```
      next
    next
    GraphCanvas.Refresh
  case 7 'Reward
    for trial=1 to App.MaxTrials
      for tick=1 to App.TrialLength
        GraphData(trial,tick)=RewardData(trial,tick)
        If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
        (GraphData(trial,tick))
      next
    next
    GraphCanvas.Refresh
  case 8 'delta weights
    if IsNumeric(GraphItemPopup.Text) then
      stimulus=GraphItemPopup.Text.Val
      for trial=1 to App.MaxTrials
        for tick=1 to App.TrialLength
          GraphData(trial,tick)=DeltaWeightsData(trial,tick,stimulus)
          If Abs(GraphData(trial,tick))>AbsGraphZMax then AbsGraphZMax=Abs
          (GraphData(trial,tick))
        next
      next
      GraphCanvas.Refresh
    else
      MsgBox "No input selected. Please select the number of the input and push the
      button again."
    end if
  case else
  end select
  Me.MouseCursor= System.Cursors.StandardPointer
End Sub
```

RLwindow.RunModel:

```
Function RunModel() As Boolean
  Dim OK As Boolean

  Me.MouseCursor= System.Cursors.Wait
  GraphTypePopup.Enabled=False
  GraphItemPopup.Enabled=False

  CalculateData
  if App.SaveOutput=True then SaveData
  GraphDataFlag=True
  GraphTypePopup.Enabled=True
  GraphTypePopup.ListIndex=0
  GraphButton.Enabled=True
```

```
    if RLwindow.SupressPlot=False then
      PlotGraph
    end if
    Me.MouseCursor= System.Cursors.StandardPointer
    OK=True

    return OK
End Function
Protected AxisLabelSize As Integer

Protected GridToggle As Boolean

Protected GridColor As Color

Protected XTicEvery As Integer

Protected YTicEvery As Integer

Protected GraphRamp(512) As Color

Protected LegendWidth As Integer

Protected FrameWidth As Integer

Protected FrameHeight As Integer

Protected LegendMax As double

Protected LegendMin As double

Protected LegendMid As double

Protected GraphDataFlag As Boolean

GraphData(-1,-1) As Double

DopamineData(-1,-1) As Double

DeltaWeightsData(-1,-1,-1) As Double

RewardData(-1,-1) As Double

InputsData(-1,-1,-1) As Double

WeightedInputsData(-1,-1,-1) As Double
```

```
WeightsData(-1,-1,-1) As Double

PresynapticActivity(-1,-1) As Double

PostsynapticActivity(-1,-1) As double

AbsGraphZMax As Double

EligibilityTrace(-1,-1,-1) As Double

Protected SupressPlot As Boolean
```

RLwindow Control GraphTypePopup:

```
Sub Change()
  Dim i As Integer

  select case GraphTypePopup.ListIndex
  case 0,5,6,7 'Dopamine; intermediate layer input, intermediate layer output; reward
    GraphItemPopup.Enabled=False
  case 1,2,3,4,8 'Inputs; weights; weighted input; eligibility trace; change in weights
    GraphItemPopup.DeleteAllRows
    for i= 1 to App.InputN
      GraphItemPopup.AddRow str(i)
    next
    GraphItemPopup.Enabled=True
  case else
  end select
End Sub
```

RLwindow Control GraphCanvas:

```
Sub Open()



End Sub

Sub Paint(g As Graphics)
  Dim i as integer, j as integer, rampH as integer, rampX as integer, rampY as integer,
  midramp As Integer, maxramp As Integer
  Dim label as String, labelwidth as Integer, labelheight as Integer, labelevery as Integer
  Dim t As Integer, b As Integer, l As Integer, r As Integer, m As Integer
  Dim trial,tick,w,h,ramp,totalticks As Integer

  Me.MouseCursor= System.Cursors.Wait
```

```
g.TextSize=9
'Scale z-axis
LegendMax=AbsGraphZMax
LegendMid=0
LegendMin=AbsGraphZMax*-1
'Legend
maxramp=UBound(GraphRamp)
midramp=maxramp/2
InitRamp
rampH=GraphCanvas.Height/(maxramp+1)
rampX=GraphCanvas.Width-LegendWidth-30
for i=0 to maxramp
  rampY=GraphCanvas.Height-((i*GraphCanvas.Height)/(maxramp+1))
  g.ForeColor=GraphRamp(i)
  g.FillRect rampX, rampY, LegendWidth,rampH+1
  select case i
  case 0
    label=format(LegendMin, "+##.000")
    labelheight=g.StringHeight(label,30)
    labelwidth=g.StringWidth(label)
    g.ForeColor=RGB(0,0,0)
    g.DrawString label,GraphCanvas.Width-labelwidth-1,GraphCanvas.Height
  case midramp
    label=format(LegendMid, "##.000")
    labelheight=g.StringHeight(label,30)
    labelwidth=g.StringWidth(label)
    g.ForeColor=RGB(0,0,0)
    g.DrawString label,GraphCanvas.Width-labelwidth-1,rampY+(rampH/2)
  case maxramp
    label=format(LegendMax, "+##.000")
    labelheight=g.StringHeight(label,30)
    labelwidth=g.StringWidth(label)
    g.ForeColor=RGB(0,0,0)
    g.DrawString label,GraphCanvas.Width-labelwidth-1,12
  case else
  end select
next

'Graph frame
g.DrawRect 25,0,GraphCanvas.Width-Legendwidth-60,GraphCanvas.Height-15

'X-axis
labelEvery=Round(App.MaxTrials/10)
t=GraphCanvas.Height-15
b=t+3
j=GraphCanvas.Width-Legendwidth-61
```

```
for i=1 to App.MaxTrials
   m=25+(((i-1)*j)/App.MaxTrials)
   g.DrawLine m,t,m,b
   if ((i=1) or (i mod labelEvery =0)) then
      label=format(i, "###")
      labelheight=g.StringHeight(label,30)
      labelwidth=g.StringWidth(label)
      g.ForeColor=RGB(0,0,0)
      m=m+(j/(App.MaxTrials*2))-(labelwidth/2)
      g.DrawString label, m, GraphCanvas.Height
   end if
next

'Y-axis
r=25
l=r-3
j=GraphCanvas.Height-16
for i=1 to App.TrialLength
   m=GraphCanvas.Height-16-(((i-1)*j)/App.TrialLength)
   g.DrawLine l,m,r,m
   if ((i=1) or (i mod 5 =0)) then
      label=format(i, "###")
      labelheight=g.StringHeight(label,30)
      labelwidth=g.StringWidth(label)
      g.ForeColor=RGB(0,0,0)
      m=m-(j/(App.TrialLength*2))+(labelheight/2)
      g.DrawString label, 0, m
   end if
next

'Data
if GraphDataFlag=true then
   g.ForeColor=RGB(0,0,0)
   i=GraphCanvas.Width-Legendwidth-61
   j=GraphCanvas.Height-16
   w=(i/App.MaxTrials)+1
   h=(j/App.TrialLength)+1
   totalticks=App.MaxTrials*App.TrialLength
   for trial=1 to App.MaxTrials
      for tick=1 to App.TrialLength
         b=GraphCanvas.Height-16-(((tick-1)*j)/App.TrialLength)
         l=25+(((Trial-1)*i)/App.MaxTrials)+1
         if LegendMax-LegendMin>0 then
            ramp=((GraphData(trial,tick)-LegendMin)*UBound(GraphRamp))/(LegendMax-
            LegendMin)
         else
```

```
            ramp=0
          end
          g.Forecolor=GraphRamp(ramp)
          g.FillRect(l,b-h,w,h)
        next
      next
    end if

    Me.MouseCursor= System.Cursors.StandardPointer
End Sub
```

RLwindow Control GraphButton:

```
Sub Action()
    PlotGraph
End Sub
End Class
```

Class App

```
Inherits Application

ModelName As string

SaveOutput As boolean

SaveParameters As boolean

TemporalDiscount As double

TraceDecay As double

TrialLength As integer

MaxTrials As integer

LearnRate As double

GraphicsSelections(6) As boolean

InputN As Integer

MinDA As String

TraceRise As Double
```

```
UseGradedTrace As Boolean

UsedMacro As Boolean

MacroSaveDataDirectory As FolderItem

End Class
```

Class Pwindow

Inherits Window

Pwindow.Open:

```
Sub Open()
  DefineSequence.Heading(-1)="First trial"+chr(9)+"Last trial"+chr(9)+"Onset"+chr(9)
  +"Offset"+chr(9)+"Event [r,1,2...]"+chr(9)+"Magnitude"
  SpecsDoneButton.Visible=True
  SpecsDoneButton.Enabled=True
  Pwindow.left=(Screen(0).width/2)-(Pwindow.width/2)
  Pwindow.top=(Screen(0).height/2)-(Pwindow.height/2)
  GradualOnOffCheckBox.Value=True
  LearningRate.Text="0.055"
  InputOnsetConstant.Text="-0.9"
  InputOffsetConstant.Text="-0.1054"
  TemporalDiscount.Text="0.98"
  DAminimum.Text="-0.05"
End Sub
```

Pwindow.AllDone:

```
Sub AllDone()
  Dim events As integer
  Dim i,j,k As integer
  Dim f as FolderItem
  Dim stream as TextOutputStream
  Dim FileName As string
  Dim StartTrial,EndTrial,StartTick,EndTick,Stimulus,Trial,Tick,DeltaTick As Integer
  Dim Magnitude,InputActivation as Double
  Dim StimulusASCII As String

  Me.MouseCursor= System.Cursors.Wait

  SpecsDoneButton.Enabled=False
  App.LearnRate=LearningRate.text.val
  App.TraceDecay=InputOffsetConstant.Text.Val
  App.TraceRise=InputOnsetConstant.Text.Val
```

```
App.TemporalDiscount=TemporalDiscount.Text.Val
App.ModelName=ModelName.text
App.SaveParameters=SaveSpecs.value
App.SaveOutput=SaveOutput.value
App.TrialLength=TrialLength.text.val
App.MaxTrials=TrialN.Text.Val
App.InputN=StimuliN.Text.Val
App.MinDA=DAminimum.text

Redim RLWindow.GraphData(App.MaxTrials+1,App.TrialLength+1)
Redim RLWindow.DopamineData(App.MaxTrials+1,App.TrialLength+1)
Redim RLWindow.RewardData(App.MaxTrials+1,App.TrialLength+1)
Redim RLWindow.InputsData(App.MaxTrials+1,App.TrialLength+1,App.InputN+1)
Redim RLwindow.EligibilityTrace(App.MaxTrials+1,App.TrialLength+1,App.InputN+1)
Redim RLWindow.WeightsData(App.MaxTrials+1,App.TrialLength+1,App.InputN+1)
Redim RLWindow.DeltaWeightsData(App.MaxTrials+1,App.TrialLength+1,App.InputN
+1)
Redim RLWindow.WeightedInputsData(App.MaxTrials+1,App.TrialLength+1,App.InputN
+1)
Redim RLwindow.PresynapticActivity(App.MaxTrials+1,App.TrialLength+1)
Redim RLwindow.PostsynapticActivity(App.MaxTrials+1,App.TrialLength+1)

for i=0 to App.MaxTrials
  for j=0 to App.TrialLength
    RLwindow.GraphData(i,j)=0
    RLwindow.DopamineData(i,j)=0
    RLwindow.RewardData(i,j)=0
    RLwindow.PresynapticActivity(i,j)=0
    RLwindow.PostsynapticActivity(i,j)=0
    for k=0 to App.InputN
      RLwindow.InputsData(i,j,k)=0
      RLwindow.EligibilityTrace(i,j,k)=0
      RLwindow.WeightsData(i,j,k)=0
      RLwindow.DeltaWeightsData(i,j,k)=0
      RLwindow.WeightedInputsData(i,j,k)=0
    next
  next
next

events=DefineSequence.ListCount

if SaveSpecs.Value=True then
  FileName=ModelName.text
  Me.MouseCursor= System.Cursors.StandardPointer
  f=GetSaveFolderItem("TEXT",FileName+".txt")
  if f<> Nil then
```

```
      Me.MouseCursor= System.Cursors.Wait
      stream=f.CreateTextFile
      f.MacCreator="PSYr"
      Stream.Writeline "Trials"+chr(9)+TrialN.Text
      Stream.Writeline "Ticks"+chr(9)+TrialLength.Text
      Stream.Writeline "Stimuli"+chr(9)+StimuliN.Text
      If GradualOnOffCheckBox.Value=True then
        Stream.WriteLine "GradedInputs"+chr(9)+"True"
      else
        Stream.WriteLine "GradedInputs"+chr(9)+"False"
      end
      Stream.Writeline "LearningRate"+chr(9)+LearningRate.Text
      Stream.Writeline "TemporalDiscount"+chr(9)+TemporalDiscount.Text
      if GradualOnOffCheckBox.Value=True then
        Stream.Writeline "TraceRise"+chr(9)+InputOnsetConstant.text
        Stream.Writeline "TraceDecay"+chr(9)+InputOffsetConstant.text
      end
      If IsNumeric(App.MinDA)=True then
        Stream.Writeline "DAminimum"+chr(9)+App.MinDA
      end
      if ShuffleTrialsCheckBox.Value=True then
        Stream.Writeline "TrialShuffle"+chr(9)+"True"
        stream.WriteLine "StartShuffle"+chr(9)+StartTrialShuffleEditField.Text
        stream.WriteLine "EndShuffle"+chr(9)+EndTrialShuffleEditField.Text
      end
      Stream.WriteLine "Structure"
      for i=0 to events-1
        for j=0 to 5
          select case j
          case 5
            Stream.Writeline DefineSequence.cell(i,j)
          else
            Stream.Write DefineSequence.cell(i,j)+chr(9)
          end select
        next
      next
      Stream.Close
    end
  end

  Me.MouseCursor= System.Cursors.Wait
  for i=0 to events-1
    'Setup arrays
    StartTrial=DefineSequence.cell(i,0).Val
    EndTrial=DefineSequence.cell(i,1).Val
    StartTick=DefineSequence.cell(i,2).Val
```

```
    EndTick=DefineSequence.cell(i,3).Val
    StimulusASCII=DefineSequence.cell(i,4)
    Magnitude=DefineSequence.cell(i,5).Val
    select case InStr(StimulusASCII,"Stimulus")
    case 0 'Reward
      for Trial=StartTrial to EndTrial
        for Tick=StartTick to EndTick
          RLwindow.RewardData(Trial,Tick)=Magnitude
        next
      next
    case else 'Stimulus
      Stimulus=val(right(StimulusASCII,1))
      for Trial=StartTrial to EndTrial
        for Tick=StartTick to EndTick
          RLwindow.InputsData(Trial,Tick,Stimulus)=Magnitude
          if GradualOnOffCheckBox.Value=true Then
            DeltaTick=Tick-StartTick+1
            InputActivation=Magnitude-(exp(App.TraceRise*DeltaTick)*Magnitude)
            if InputActivation>Magnitude then InputActivation=Magnitude
            RLwindow.EligibilityTrace(Trial,Tick,Stimulus)=InputActivation
            if Tick=EndTick then
              for j=EndTick+1 to TrialLength.Text.Val
                DeltaTick=j-EndTick
                RLwindow.EligibilityTrace(Trial,j,Stimulus)=exp
                (App.TraceDecay*DeltaTick)*Magnitude
              next
            end
          else
            RLwindow.EligibilityTrace(Trial,Tick,Stimulus)=Magnitude
          end
        next
      next
    end
  next

  if ShuffleTrialsCHeckBox.Value=True then DoShuffle

  Me.MouseCursor= System.Cursors.StandardPointer


End Sub
```

Pwindow.DoShuffle:

```
Sub DoShuffle()
  Dim i,j,k,source,destination as Integer
  Dim Temp as Double
```

```
  for i=StartTrialShuffleEditField.Text.Val to EndTrialShuffleEditField.Text.Val
    source=Round((rnd()*(EndTrialShuffleEditField.Text.Val-
    StartTrialShuffleEditField.Text.Val))+StartTrialShuffleEditField.Text.Val)
    destination=Round((rnd()*(EndTrialShuffleEditField.Text.Val-
    StartTrialShuffleEditField.Text.Val))+StartTrialShuffleEditField.Text.Val)
    for j=1 to App.TrialLength
      Temp=RLwindow.RewardData(destination,j)
      RLwindow.RewardData(destination,j)=RLwindow.RewardData(source,j)
      RLwindow.RewardData(source,j)=Temp
      for k=1 to App.InputN
        Temp=RLwindow.InputsData(destination,j,k)
        RLwindow.InputsData(destination,j,k)=RLwindow.InputsData(source,j,k)
        RLwindow.InputsData(source,j,k)=Temp
        Temp=RLwindow.EligibilityTrace(destination,j,k)
        RLwindow.EligibilityTrace(destination,j,k)=RLwindow.EligibilityTrace(source,j,k)
        RLwindow.EligibilityTrace(source,j,k)=Temp
      next
    next
  next


End Sub
OKclose As boolean
```

Pwindow Control DefineSequence:

```
Function CellKeyDown(row as Integer, column as Integer, key as String) As Boolean

  if key.asc=9 then 'Tab
    if column<DefineSequence.ColumnCount-1 then
      column=column+1
    else
      column=0
      if row=DefineSequence.LastIndex then row=0 else row=row+1
    end if
    DefineSequence.EditCell(row,column)
    return(True)
  else
    if column=4 then
      select case key
      case "r"
        DefineSequence.ActiveCell.text="Reward"
        return(True)
      case "s"
        DefineSequence.ActiveCell.text="Stimulus?"
```

```
        return(True)
      else
        if key.val<=9 and key.val>=1 then
          if key.val<=Pwindow.StimuliN.Text.Val Then
            DefineSequence.ActiveCell.text="Stimulus"+key
            return(True)
          else
            MsgBox "The stimulus you have specified here is greater than the maximum
            number of stimuli you specified on the 'Model' tab..."
            DefineSequence.ActiveCell.text=""
            return(True)
          end
        else
          DefineSequence.ActiveCell.text="???"
          return(True)
        end if
      end select
    end if
    return(False)
  end if
End Function

Function CellClick(row as Integer, column as Integer, x as Integer, y as Integer) As Boolean
  DefineSequence.EditCell(row,column)
End Function
```

Pwindow Control AddEvent:

```
Sub Action()
  DefineSequence.AddRow("")
  DefineSequence.Cell(DefineSequence.LastIndex,5)="1.00"
  DefineSequence.EditCell(DefineSequence.LastIndex,0)
End Sub
```

Pwindow Control DeleteEvent:

```
Sub Action()
  dim x as integer
  if DefineSequence.ListCount>=0 then
    x=DefineSequence.Listindex
    if x>=0 then
      DefineSequence.RemoveRow(x)
    else
      MsgBox "No event selected!"
    end if
  end if
End Sub
```

Pwindow Control SpecsDoneButton:

```
Sub Action()
  OKclose=True
  AllDone
  Pwindow.Close
End Sub
```

Pwindow Control InputOnsetConstant:

```
Sub TextChange()
  TraceCanvas.Refresh
End Sub
```

Pwindow Control ParametersPopup:

```
Sub Change()
  Select Case ParametersPopup.ListIndex
  Case -1 'Nothing selected
  Case 0 'EMB
    GradualOnOffCheckBox.Value=True
    LearningRate.Text="0.055"
    InputOnsetConstant.Text="-0.9"
    InputOffsetConstant.Text="-0.1054"
    TemporalDiscount.Text="0.98"
    DAminimum.Text="-0.05"
  Case 1 'Montague et al. 1996
    GradualOnOffCheckBox.Value=False
    LearningRate.Text="0.3"
    InputOnsetConstant.Text="-9999"
    InputOffsetConstant.Text="-9999"
    TemporalDiscount.Text="1"
    DAminimum.Text="none"
  Case 2 'User defined
  End Select

End Sub
```

Pwindow Control GradualOnOffCheckBox:

```
Sub Action()
  Select Case GradualOnOffCheckBox.Value
  Case True
    OffsetConstantLabel.Enabled=True
    OnsetConstantLabel.Enabled=True
    InputOffsetConstant.Enabled=True
    InputOnsetConstant.Enabled=True
    TraceCanvas.Enabled=True
```

```
    InputChartXLabel.Enabled=True
    InputChartYlabel.Enabled=True
    App.UseGradedTrace=True
  Case False
    OffsetConstantLabel.Enabled=False
    OnsetConstantLabel.Enabled=False
    InputOffsetConstant.Enabled=False
    InputOnsetConstant.Enabled=False
    TraceCanvas.Enabled=False
    InputChartXLabel.Enabled=False
    InputChartYlabel.Enabled=False
    App.UseGradedTrace=False
  End Select
End Sub
```

Pwindow Control InputOffsetConstant:

```
Sub TextChange()
  TraceCanvas.Refresh
End Sub
```

Pwindow Control TraceCanvas:

```
Sub Paint(g As Graphics)
  Dim t As Integer
  Dim steps As Integer
  Dim x As Integer
  Dim y As Integer
  Dim oldX As Integer
  Dim oldY As Integer
  Dim onset,offset,duration,deltaT As Integer
  Dim EligibilityTrace,maxTrace As Double

  onset=(Me.Width\6)*2
  offset=(Me.Width\6)*3
  duration=offset-onset
  Me.Graphics.ForeColor=RGB(200,255,200)
  Me.Graphics.FillRect(onset,0,duration,Me.Height)
  Me.Graphics.ForeColor=RGB(0,0,0)
  Me.Graphics.DrawRect(0,0,Me.Width,Me.Height)

  steps=Me.Width
  Me.Graphics.ForeColor=RGB(127,0,0)
  Me.Graphics.PenWidth=2
  Me.Graphics.PenHeight=2
  for t=0 to steps-1
    x=t
```

```
      Select Case t
      Case Is <onset
        y=Me.Height
      Case Is >offset
        deltaT=t-offset
        EligibilityTrace=exp(Pwindow.InputOffsetConstant.Text.Val*deltaT)*maxTrace
        y=(1.0-EligibilityTrace)*Me.Height
      Case else
        deltaT=(t-onset)+1
        EligibilityTrace=1-exp(Pwindow.InputOnsetConstant.Text.Val*deltaT)
        y=(1.0-EligibilityTrace)*Me.Height
        if t=offset then maxTrace=EligibilityTrace
      End Select
      if t>0 then Me.Graphics.DrawLine(oldX,oldY,x,y)
      oldX=x
      oldY=y
    next

    Me.Graphics.PenWidth=1
    Me.Graphics.PenHeight=1

End Sub
```

Pwindow Control ShuffleTrialsCheckBox:

```
Sub Action()
    Select Case ShuffleTrialsCheckBox.Value
    Case True
      StartShuffleLabel.Enabled=True
      EndShuffleLabel.Enabled=True
      StartTrialShuffleEditField.Enabled=True
      EndTrialShuffleEditField.Enabled=True
    Case False
      StartShuffleLabel.Enabled=False
      EndShuffleLabel.Enabled=False
      StartTrialShuffleEditField.Enabled=False
      EndTrialShuffleEditField.Enabled=False
    End Select
End Sub
```

Pwindow Control StartTrialShuffleEditField:

```
Sub LostFocus()
    if StartTrialShuffleEditField.Text.Val<1 then
      MsgBox "The starting trial for shuffling cannot be less than 1."
      StartTrialShuffleEditField.Text="1"
    end
```

```
End Sub
```

Pwindow Control EndTrialShuffleEditField:

```
Sub LostFocus()
  If EndTrialShuffleEditField.Text.Val<=StartTrialShuffleEditField.Text.Val then
    MsgBox "The last trial for shuffling must be greater than the first."
    if StartTrialShuffleEditField.Text.Val+1>App.MaxTrials then
      EndTrialShuffleEditField.Text=str(App.MaxTrials)
    else
      EndTrialShuffleEditField.Text=str(StartTrialShuffleEditField.Text.Val+1)
    end
  end

  If EndTrialShuffleEditField.Text.Val>TrialN.Text.Val then
    MsgBox "The last trial for shuffling must not be greater than the total number of trials
    you have specified"
    EndTrialShuffleEditField.Text=TrialN.Text
  end
End Sub
End Class
```