The University Of Sheffield.

Access to Electronic Thesis

Author: Thomas Walton

Thesis title: The Discovery of Novel Actions

Qualification: PhD

The Discovery of Novel Actions

Thomas David Walton

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Department of Psychology

October 2011

Acknowledgements

The author would like to acknowledge the support of Professor Pete Redgrave, David Yates, Dr Martin Thirkettle and Cigir Kalfaoglu, who all offered advice and discussion throughout the course of this work.

Particular thanks are owed to my supervisor, Dr Tom Stafford, for giving me the opportunity to pursue a PhD and for his patience, encouragement and support throughout.

Most of all, I would like to thank my wife, Lisa Walton, for her unwavering support and belief in me.

Abstract

The aim of this research was to develop a behavioural paradigm capable of quantifying action acquisition. It takes the form of a series of experiments in which human participants learn to produce new actions with a joystick. Research questions were focussed on the behavioural implications of Redgrave and Gurney's (2006) theory that dopamine neurons in the ventral midbrain play a pivotal role in the reinforcement and reselection of motor output that is essential to action learning. The first study looked at the effect of delayed audio and visual reinforcement on the ability to learn stable hand positions. Delays of 100 ms were found to impair acquisition in both modalities. This degree of temporal sensitivity supports the idea that dopamine neurons fire at low latencies to reduce the reinforcement of non-contiguous motor output. The second study investigated the effect of delay on the learning of hand movements. The movements produced during the delay period were analysed to address the question of whether the quantity of non-contingent output would impact on learning over and above the mismatch in temporal alignment. The results revealed that this was not the case, thus suggesting that timing is of primary importance to learning. The final study utilised a task requiring more complex movements, in an attempt to reduce the contribution of high-level, conscious, learning in favour of low-level non-declarative learning. Performance was compared across conditions, which differed in the quantity of spatial information provided. No evidence was found that the type of movements produced during learning impacted on later performance, thus indicating a tendency to use high-level spatial guidance of movements. All findings are discussed in terms of the value of the current paradigm and the extent to which they support the theory that action learning is mediated by a time-stamping mechanism in the midbrain.

Contents

Ethics

All research described in this thesis was carried out with ethical approval from the departmental ethics committee of the department of psychology at the University of Sheffield.

Chapter 1: the discovery of novel actions

Redgrave and Gurney (2006) and Redgrave, Gurney and Reynolds (2008) offer an alternative explanation to the prevailing theory (Schultz, Dayan & Montague, 1997) regarding the function of the phasic activity of midbrain dopamine neurons. They argue that this activity does not serve as a reward prediction error signal, but is central to the reinforcement and reselection of motor output and the discovery of novel actions. If true, this dopaminergic activity could underpin some of the most basic aspects of learning and behaviour including agency detection and action acquisition. However, whilst the tools and methods for investigating the effects of reward prediction are refined and clearly calibrated across the research community in the form of operant chambers, shaping techniques and standardised schedules of reinforcement, the most popular model for investigating action acquisition – lever pressing – is limited to the extent it that it makes different levels of performance, and therefore different stages of acquisition, difficult to quantify. The motivation for the current research was to develop a behavioural paradigm, suitable for human participants, that is capable of testing hypotheses derived from Redgrave and Gurney's theory. The three chapters that follow chart the development of this paradigm, but first this introduction attempts to situate the topic of action acquisition amongst other similar fields in psychology, and then discusses possible neural correlates and the implications of neural processes at the behavioural level and finally consideration is given to some of the methodological requirements of a behavioural test of action acquisition.

Previous research into action acquisition

Thorndike: action acquisition in animals

Thorndike (1911) was a pioneer in the quantitative study of learning. He investigated action acquisition using an escape paradigm whereby animals were repeatedly exposed to situations in which a particular sequence of movements was required in order to bring about a desirable consequence, namely escape from an enclosure. The iterated nature of the task allowed Thorndike to record the change in an animal's behaviour over successive trials, both qualitatively: by means of general observations,

and quantitatively: by measuring the time between entry into the experimental chamber and the point at which its escape was completed. In perhaps the best known version of this task, a cat was enclosed in a puzzle-box and its behaviour observed as it learned, over successive trials, to escape by pressing a foot pedal linked to a door release mechanism. Initially, the animal would produce the escape behaviours natural to a cat in a confined space and, consequently, any depression of the foot pedal was a mere by-product of this. However, after many attempts, the normal escape response of the cat was gradually reduced and behaviour consistent with lever depression was increased. The cat's behaviour in the context had apparently become more purposeful and efficient. In other words, the cat had added a new action to its behavioural repertoire.

Thorndike's (1911) simple and intuitive account of the learning process was just as influential as the paradigm he used to record it. He emphasised the animal's ability to monitor the consequences of its own behaviour and observed that those responses which do not result in positive outcomes are gradually "stamped out" whilst all those resulting in reinforcement are gradually "stamped in" (p.74). He called this the law of effect and explained it as follows:

> The Law of Effect is that: Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond. (p.244).

Although somewhat simplistic, Thorndike's law of effect retains the durability of other process descriptions such as 'survival of the fittest', not because it is a complete explanation but because it provides a means of describing learning without reference to the unknown mechanisms on which learning depends. It also offers a neat summary of the process under investigation here.

For the current purposes, the term 'action acquisition' is used to describe the type of learning undertaken by Thorndike's animal subjects and here this is treated as distinct from action-outcome learning (Dickinson & Balleine, 2000). Typically, in action-outcome learning a stable sequence of movements has already been acquired and the learning in question involves the connection of new effects with this set of movements. Action acquisition, by contrast, is treated here as the reduction of behavioural variance from an initially large set of varied movements to a smaller, stable set of movements that reliably bring about a particular outcome. This does not necessarily involve the animal learning new movements that it was previously incapable of performing, rather it involves the animal discovering that a given chunk of behaviour has reliable consequences such that this chunk is now treated differently to the individual elements from which it is comprised. During this process, the animal is therefore forced to solve a credit assignment problem (Barto, Sutton & Anderson, 1983; Minsky, 1961), something that is less of an issue in action-outcome learning where the pre-learned sequence of movements can potentially be dealt with as a chunk. Despite the distinction drawn here, however, it is acknowledged that there is much crossover between the two types of learning.

Skinner: the maintenance of an action

With behaviourism, attention turned from action acquisition to the ways in which behaviour could be recorded and manipulated. This is exemplified by Ferster and Skinner (1957) who developed a technique to investigate the effect that different timings, magnitudes and probabilities of reinforcement could have on an animal's responses; they termed the technique operant conditioning. In a typical experiment, a rat or a pigeon would be placed inside a small cage or box (an operant chamber) containing one or more levers. Depression of a lever would be reinforced by delivery of food into a hopper. By varying the timing, quantity or likelihood of food being delivered, the rate at which the animal responded with depressions of the lever could be changed. Thus, psychologists had a means of distilling some of the complexities of learning down to discrete responses on a lever operandum.

Operant conditioning owes much to the work of Thorndike; however, it would be

wrong to think of it as a simple refinement of his work. Thorndike's escape paradigm was designed to investigate the acquisition of new actions and therefore experiments utilising this paradigm continue only until the point at which the relevant action has been fully learned. Operant conditioning, by contrast, was designed to investigate how reinforcement can change the rate at which animals produce fully formed (previously learned) actions. The relationship between the two approaches is explained by Skinner (1969) as follows:

> By thoroughly adapting the rat to the box before the lever is made available, most of the competing behaviour can be "stamped out" before the response to be learned is ever emitted. Thorndike's learning curve, showing the gradual disappearance of unsuccessful behaviour, then vanishes. In its place we are left with a conspicuous change in the successful response itself: an immediate, often quite abrupt, increase in rate. (pp.6-7).

In other words, Skinner's primary aim when employing operant conditioning was not to investigate how a rat learns to press a lever, but instead to record how the probability of a fully formed lever-press response changes over time.

The obvious question for anyone interested in action acquisition, is how do the animal subjects acquire the lever pressing (or key pecking) actions in the first place? To Skinner, the acquisition phase was essentially an obstacle to measuring rate of response (Lattal & Gleeson, 1990) and he developed a technique for quickly transforming simple components of an animal's behaviour into more elaborate actions and sequences of actions: a process known as shaping. From an experimental point of view, this avoids the need for a protracted period whereby the animal acquires the behaviour automatically. However, whilst it is easy to appreciate that researchers might want to skip straight to the process of interest, it is hard to escape the impression that Skinner's efficient techniques for investigating learning have resulted in questions of acquisition being left relatively ignored (Peterson, 2004).

What is learnt during action acquisition?

A somewhat confusing aspect of the study of action acquisition is the question of what constitutes an action. Redgrave et al. (2008) define action acquisition as "discovering a

movement that has a predicted outcome" (p.331). The animals in Thorndike's experiments were better able to effect an escape from the enclosures by the end of a testing session than they were at the beginning and yet it seems strange to suggest that they had learnt new movements during this time. So what exactly did they learn? One might argue that they learned that the pedal was the cause for the door opening and in a sense this is true, but with an important qualification. Thorndike found that demonstrating the task to the animals prior to a testing session did not increase their rate of learning. He also found that, once trained, the animals would attempt to press the door release pedal, even if that pedal was removed (they would 'press' thin air). In other words, it is likely that the cats' understanding of their behaviour was very limited. However, what we can say with confidence is that the cats had learned a stable sequence of movements that reliably resulted in a desirable outcome. The process of combining and refining movements would seem, therefore, to be central to action acquisition.

There is no simple relationship between the proximity of any given movement within an action and its contribution to bringing about a particular outcome. The final portion of an action may be no more or less important to the outcome than the earliest part of an action. Because of this, actions are often not decomposable into component parts, even if each of those parts is just a simple movement learned long before the whole action. Indeed, Ostlund, Winterbauer and Balleine (2009) have demonstrated that action sequences – that is to say, combinations of not just movements, but entire actions – can be treated as a single reinforceable entities or 'chunks'. They found that healthy rats were able to avoid a particular sequence of lever presses when the outcome of that sequence had been devalued through prefeeding. Importantly, they were able to demonstrate that the rats weren't merely avoiding sequences based on individual components within those sequences. Actions, then, are perhaps not only defined in terms of the movements of which they are composed, but also in terms of the outcome they bring about.

Action-outcome learning

Thorndike's work is clearly a very pure and intuitive approach to the study of action

acquisition, but the issue of action sequence chunking shows that we can take an entirely different perspective on the subject: it is possible to study action acquisition when the stable sequence of movements already exists within the animal's behavioural repertoire. In such circumstances, we can focus instead on what the animal learns about the consequences of those movements such as how the utility changes depending on the animal's inner state (e.g. level of satiety or nausea). This approach has come to be known as action-outcome learning (Dickinson & Balleine, 2000).

Action-outcome learning can be thought of as the formation of memories regarding the causal consequences of actions, such that the probability of invoking a learnt action at any given time is dependent on the current utility of that action's consequences. Adams and Dickinson (1981) pioneered the modern study of action-outcome learning with two experiments that tested "whether or not animals know about the consequences of their actions" (p.109). Rats were trained to press a lever in return for a food reinforcer. They were also fed on a different kind of food, not paired with lever pressing. Following this training phase, half of the rats had the contingent food devalued whilst the other half had the non-contingent food devalued; the devaluation was achieved by injecting them with lithium chloride (a toxic salt that causes sickness) following consumption of the relevant food. It was found that the number of lever presses performed during subsequent extinction trials was lower in the group of rats for whom the contingent food had been devalued. In other words, the rats' experience of the contingent food (i.e. sickness) outside of the normal lever-pressing context affected their behaviour when they returned to the lever pressing situation, even though they didn't experience the reinforcer again. Apparently, this difference in behaviour depends crucially on the rats' knowledge of what lever pressing does.

This study and other similar studies (e.g. Colwill & Rescorla, 1985) concern action acquisition insofar as they provide information about precisely what has been learnt by the animal. For example, in the Adams and Dickinson (1981) experiments, we find that learning the action involves more than just the connection of stimuli with responses because, after devaluation, the rate of response was measured during

extinction trials so there was no further exposure to a stimulus-response relationship that could have prompted the drop in rate of response. Therefore, to fully characterise the action that has been learnt by the rats, it isn't enough speak in terms of stable sequences of movements, we must also take into account what the rat has learnt about the consequences of these movements because this informs its choices as to whether or not to produce the action at a given time and allows it to make intelligent decisions which take into account its own internal state.

The focus of the current research concerns the reselection, repetition and reinforcement of recent movements as detailed by Redgrave and Gurney (2006) and Redgrave et al. (2008) and, therefore, it is considered that this puts it in the tradition of Thorndike's (1911) escape paradigm as opposed to the devaluation paradigms which are central to action-outcome learning research. In nonhuman animals, an understanding of utility is most directly expressed in terms of rate of response and is most informative when the response is a sequence of movements that already exists within the animal's behavioural repertoire. By contrast, the extent to which recent motor output has been reinforced is most directly expressed in terms of the efficiency with which a sequence of movements is performed. When studying this efficiency, we must assume that the outcome/consequence of the recent motor output has at least some utility to the agent be that through novelty, experimenter instructions or nutritional benefits. In other words, the animal's motivation to elicit the outcome is a requirement of the investigation, and the extent to which the outcome is useful to the animal is only a secondary concern.

Action-acquisition versus supervised learning

A basic consideration when developing a new experimental paradigm is to ascertain whether it is likely to be able to measure the process of interest. Specifically, there is a risk that a joystick paradigm might tell us more about motor control than it does about action acquisition. It has long been argued that the movements we make must, to some extent, be pre-programmed and, where necessary, be corrected during travel by the use of an internal model predicting the sensory outcome of the movement (Desmurget & Grafton, 2000; Jordan & Rumelhart, 1992). One of the most persuasive

reasons for believing this is that some movements are executed and corrected so rapidly that there wouldn't be time to guide the movement using sensory feedback, due to the latencies in conveying that information through the nervous system (Bard et al., 1999). Consequently, it seems certain that such movements are carried out based on a motor program that specifies the movement so, once planned, it can be carried out without external feedback and errors can be predicted in advance. The creation of these motor programs is, of course, closely related to the topic of action acquisition. Consequently, it is important to be clear on the differences between supervised learning and reinforcement learning, given that the latter is the primary focus of the current research.

A good example of a supervised learning paradigm is the task employed by Kitazawa, Kohno and Uka (1995). They projected targets onto a screen and required human participants to make reaching movements towards these targets. Participants wore prism spectacles, which shifted their view of the target by a fixed amount. The spectacles also served to render the target invisible during reaching movements by becoming opaque. This ensured that the only visual feedback received by the participants was on first sight of the target and then again once their finger had settled on the screen at the end of the reaching movement. By this method, the experimenters were able to measure the adaptation of the participants' reaching movements over several trials as they learned to correct for the effect of the prisms.

Whilst it is easy to identify archetypal examples of both supervised learning (e.g. Kitazawa et al., 1995) and reinforcement learning (e.g. operant conditioning or Thorndike, 1911), the key differences between the two aren't quite so apparent. Here we will take a methodological perspective, mindful that there are other technical ways of differentiating the two types of learning at the algorithmic level (e.g., Jordan & Rumelhart, 1992). Essentially the difference lies in how the results of behaviour are presented to the agent. In the case of supervised learning, the results are presented in such a way that the agent can compare their actual performance to a target performance. In the case of reinforcement learning, by contrast, the results are qualitative in nature, indicating whether performance was successful or not, and perhaps even differing degrees of success, but they can't be compared to a target

performance; the target performance must be discovered through practice. In a sense, then, supervised learning is learning how to execute a movement. The outcome will certainly be successful, but it is the extent to which it is successful that matters. Reinforcement learning involves discovering what the target level of performance is whilst simultaneously learning to execute the necessary movements. The outcome of any given attempt may or may not be successful and often a lack of success won't result in explicit feedback. Another way of looking at the difference is to consider what an absence of feedback means. In the case of reinforcement learning, an absence of feedback provides the agent with useful information: it means that it hasn't yet performed the relevant action, so it must continue to try to achieve the reinforcement. By contrast, with supervised learning, the absence of feedback is not informative as it simply means that the agent is no longer attempting the task.

In the case of Thorndike's (1911) escape paradigm, the animals had a clear goal which was to escape from a puzzle box and they had clear reinforcement in the form of the lever mechanism opening the door to allow the animals to escape, but they didn't have a way of monitoring performance at any given time relative to optimal performance. Instead, they were forced to extract this information from their own behavioural variance. In the experiment by Kitazawa et al. (1995), participants had a clear goal, which was to touch the target, and they had clear feedback in the form of the image of their finger relative to the target. This image provided participants with everything they needed to know in order to improve their performance; unlike Thorndike's cat, there was no need for them to gradually learn the target position/performance through several attempts at the task as it was explicitly presented to them. Interestingly, the Kitazawa et al. (1995) task appears to demonstrate that one of the features we often associate with reinforcement learning paradigms – the discrete presentation of reinforcement rather than continuously provided feedback on performance – is not something that can be used to differentiate supervised learning paradigms from reinforcement learning paradigms. By making the arm movements invisible to their participants Kitazawa et al. ensured that the feedback provided a single snapshot of performance once the attempt was complete.

Neural basis of action acquisition

The function of dopaminergic neurons

Over the past 25 years (beginning with Schultz, 1986), a large body of evidence has accumulated linking the activity of dopamine neurons in the ventral midbrain with reinforcement learning. It is now well documented that these neurons fire in response to the presentation of rewards and information predicting rewards (Schultz, 2000). In a typical experimental scenario, during which the activity of dopamine neurons is recorded, an animal learns that it will be presented with a reward (food or drink) should it perform a specific action such as pressing a button following the presentation of an arbitrary stimulus (a tone or a light). At first, the activity of the neurons correlates with the presentation of the reward itself but after repeated performance of the task the neural activity is no longer elicited by the reward and instead occurs in response to the conditioned stimulus (the tone or light). Furthermore, withholding the reward, and thus confounding the animal's prediction, results in the suppression of activity to below the baseline firing rate. This phenomenon has been formally described by Montague, Dayan and Sejnowski (1996) and Schultz et al. (1997) and has come to be known as the reward prediction error theory. As Bayer and Glimcher (2005) have put it, "the midbrain dopamine neurons are hypothesized to provide a physiological correlate of the reward prediction error signal" (p.129).

A general concern with any claim to have located the site of a particular signal in the brain is the possibility that a similar pattern of activity might exist in another unrecorded brain structure (Hellon, 1986). Furthermore, without the use of controls, it cannot be ruled out that the large numbers of responses required of animals under conditions of low sensory stimulation (e.g. Ljunberg, Apicella & Schultz, 1992; Schultz, Apicella & Ljungberg, 1993) might generate abnormal patterns of neural activity. That is to say, the unusual testing conditions (relative to natural conditions) might result in misleading data.

In the specific case of the reward prediction error theory, there are good reasons for having confidence about the involvement of dopamine neurons in instrumental learning. Firstly, increased activity is seen in response to stimuli that are correlated

with the delivery of reward (behaviourally significant) and not merely unpredicted stimuli (Ljunberg et al., 1992), so it isn't the case that the neurons are just responding to surprising stimuli. Secondly, and more convincingly, the experimental intervention in the activity of dopamine neurons has effects on operant behaviour. Tsai et al. (2009), for example, were able to bring about conditioning in mice through the deliberate activation of midbrain dopamine neurons by means of optogenetic tools. This along with lesions studies (e.g. Dowd et al., 2005) offer important evidence as to the causal role that these neurons play in instrumental learning and, consequently, add credence to theories of the function of dopamine neurons.

The reward prediction error is not the only interpretation of this neural activity. Redgrave and Gurney (2006) have called into question the observation that the activity of dopamine neurons is a response to rewards per se and not simply a response to stimuli that are novel. They argue that the neural response of these neurons to experimental stimuli occurs so soon – approximately 100 ms (Bayer & Glimcher, 2005; Guarraci & Kapp, 1999; Horvitz, Stewart & Jacobs, 1997; Ravel & Richmond 2006; Schultz, 1998; Takikawa, Kawagoe & Hikosaka, 2004) – after the occurrence of a stimulus as to be impossible for it to be a signal specifically associated with reward value. They submit that there would be insufficient time between the presentation of the stimulus and the response of the dopamine neurons for any rewarding characteristics of that stimulus to be assessed. The neurons react before a saccade can be executed at around 150 to 200 ms (Hikosaka & Wurtz, 1983; Jay & Sparks, 1987) and at a point that is at least coincident with, and therefore unlikely to be a consequence of, the time it takes for the earliest cortical identification of visual stimulus properties at around 80 to 100 ms (Rousselet, Thorpe & Fabre-Thorpe, 2004; Thorpe & Fabre-Thorpe, 2001). They point out that, in the visual domain, the only sensory signals available at such short latencies would arrive via the superior colliculus, a subcortical structure that is probably insensitive to the kind of visual information that could indicate reward value (Sumner, Adamjee & Mollon, 2002; Wurtz & Albano, 1980). This view is supported by other research, which suggests that the only visual structures with prominent connections to midbrain dopamine neurons and working at such short latencies are the superior colliculi (Comoli et al., 2003; Dommett et al, 2005; May et al., 2009).

Redgrave and Gurney (2006) offer an alternative theory, reasoning that dopamine neurons do not signal reward prediction errors but are instead involved in the job of linking novel events (irrespective of whether those events are rewarding or not) with the behavioural output that caused them. It is generally agreed that the phasic activity of these neurons changes in response to the predictability of events in a way that suggests that it is part of a learning process (Bayer & Glimcher, 2005). Rather than inferring what that learning process is from the behaviour of the animal, the approach of Redgrave and Gurney was to identify regions to which the dopamine neurons project and likely sources of input from other regions of the brain that would converge at similar times. They point out that sources of contextual (Apicella, Legallet & Trouche, 1997) and motor information (Reiner, Jiao, Del Mar, Laverghetta & Lei, 2003) should, in theory, converge on the striatum at a point in time that just precedes sensory input from the thalamus (Matsumoto, Minamimoto, Graybiel & Kimura, 2001) and the sensory evoked dopamine input from the ventral midbrain (Dommett et al., 2005), and that this indicates that the learning mechanism might be capable of both agency detection and the discovery of novel actions. In other words, their theory is that the dopaminergic activity is involved in helping an animal to learn whether or not it was responsible for the occurrence of a given stimulus and, if so, which aspect of its recent motor output was critical to eliciting that stimulus. They further argue that one of the reasons why the phasic response of dopamine neurons is of such short latency relative to the incidence of a novel stimulus, is that this reduces the opportunity for causally irrelevant movements to be reinforced and reselected (or 'stamped-in', to use Thorndike's terminology). In particular, the dopaminergic activity is of sufficiently short latency that it precludes the contribution of movements which are a direct response to the novel stimulus itself, thus removing contamination from necessarily non-contingent motor output.

The difference between the theory put forward by Redgrave and Gurney (2006) and the more well established theory offered by Montague et al. (1996) and Schultz et al. (1997) appears, initially, to be a subtle one at the neural level: reward-driven versus salience-driven learning. However, ultimately it is behaviour that matters, as behaviour is the only means through which the mechanism can be judged by evolutionary

pressures. As Hills (2006) notes:

> It is well argued that [dopamine's] role in extant species is associated with novelty or reward detection. However, this cannot be a complete definition with respect to evolution, because evolution cannot act on a detector that is not associated with some subsequent behavioural or physiological modulation. (p. 16).

At the behavioural level, the difference between the two theories is considerable, amounting to the difference between response maintenance (reward prediction error theory) and action acquisition (Redgrave et al.). Whether or not these two aspects of reinforcement learning ultimately amount to the same thing, the two theories tend to emphasise one over the other.

Here, the emphasis is firmly on action acquisition and, therefore, the theory offered by Redgrave and Gurney's (2006) and Redgrave et al. (2008) will be the primary theoretical reference point with regard to the neural correlates of action acquisition and their likely impact on behaviour. If this interpretation is correct then it means that phasic dopaminergic activity in the midbrain is doing a job very similar to the process of 'stamping-in' referred to by Thorndike (1911). For the present purposes, the term 'stamping-in' will be used to refer to the processes of reinforcement and reselection of motor output described by Redgrave and Gurney as being triggered by the phasic activity of dopamine neurons.

Temporal alignment and the eligibility period

A key component of Redgrave and Gurney's (2006) theory is the idea that a temporal alignment of sensory input and motor output is essential to the processes of action acquisition and credit assignment. They argue that phasic dopaminergic activity is triggered by salient stimuli and functions as a time stamp to indicate the portion of the motor record with which the stimulus is temporally aligned and, by virtue of this, the last part in the motor record that will be eligible for stamping-in (Redgrave et al., 2008). As to the learning and storage mechanisms, they point to long-term potentiation (LTP) and long-term depression (LTD) as the means by which motor output might be stamped-in and stamped-out respectively. If they are correct, then

there is a short-latency, short-duration timing signal to indicate when a sequence of movements is successful in eliciting a novel stimulus. There is also a process of LTP, triggered by the short-duration timing signal, which acts as the means by which successful sequences of behaviour are stored over an extended period of time (Staubli & Lynch, 1987). However, what is missing from the picture so far is a mechanism that can explain how contingent motor output and the consequences of actions are stored in the short term, from one attempt to the next and that can also determine how great a portion of the previous motor output should be eligible for stamping-in.

Histed, Pasupathy and Miller (2009) describe two possible explanations of how the consequences of an action might be stored in the short-term: either changes to synapses or "sustained firing patterns of neurons" (p.245). Their research has shown that sustained firing in the striatum and prefrontal cortex can last for several seconds, enough time to allow the temporary storage of the consequences of actions over the period between trials in their experimental testing sessions. They argue that sustained neuronal activity is the best candidate for short-term storage between learning attempts because it isn't subject to the delays inherent in synaptic changes that allow storage over comparable periods of time (2 to 6 s). Sustained neural activity is an excellent candidate mechanism for a process that involves an eligibility period. It seems plausible that such a mechanism might be responsible for sustaining the neural activity associated with motor output and there is evidence that preparatory activity in the human motor cortex (Mars, Coles, Hulstijn & Toni, 2008) can be sustained over similar periods of time

One of the implications of a timing based system of action acquisition is that it will learn indiscriminately. In a new learning situation, for example, an animal is likely to perform an action sub-optimally to begin with and yet the contiguous portion of the motor output, warts and all, will be stamped-in by virtue of its temporal proximity to a novel stimulus. In other words, if the system works by simply reinforcing recent motor output, it will be reliant on the imperfect repetition of this motor output during subsequent attempts in order to extract the common, task-relevant, elements of the behavioural variance. There is some evidence to support the existence of such a learning system. Thorndike's (1911) experiments, for example, revealed the kind of

gradual learning and apparent absence of insight that one would expect to result from such a mechanism. Skinner (1948), too, provides evidence in the form of the development of superstitious behaviour in pigeons: conditioning in the animals was produced by the coincidence of a movement with the non-contingent release of a food reinforcer. That is to say, Skinner found that movements could be stamped-in if the temporal alignment was right, even if those movements were functionally irrelevant to the process of acquiring food.

What is reward and what is reinforcement?

Throughout the present research, repeated reference will be made to the process of reinforcement so it is important to outline what is meant by this term. Much has been made of the distinction between rewards and reinforcers with some researchers placing great emphasis on the need to distinguish between the two. Salamone, Correa, Farrar, Nunes and Pardo (2009), for example, make a strong case for the need to define the concept of reward when undertaking research:

> In some papers, the word "reward" seems to be used as a rather monolithic, all- encompassing term that refers to any and all aspects of appetitive learning, motivation and emotion, whether conditioned or unconditioned. Used in this way, the term reward is a rather blunt instrument. These problems are not merely semantic, as it is difficult to test a hypothesis which maintains that a neurotransmitter mediates such an ill-defined set of functions. (p.1).

Certainly this is true when research depends on a theory that relates to things that are rewarding as opposed to merely reinforcing. However, with the greatest will in the world, authors can struggle to pin down the concept of reward. The following definition from Schultz et al. (1997) is as good an attempt as any and yet it is still tends towards the broad; it isn't clear, for example, whether it would meet the standards of Salamone et al. (2009):

> "Reward" is an operational concept for describing the positive value that a creature ascribes to an object, a behavioral act, or an internal physical state. The function of reward can be described according to the behavior elicited. For example, appetitive or rewarding stimuli induce approach behavior that permits an animal to consume. Rewards may also play the role of positive reinforcers where they increase the

> frequency of behavioral reactions during learning and maintain well-established appetitive behaviors after learning. The reward value associated with a stimulus is not a static, intrinsic property of the stimulus. Animals can assign different appetitive values to a stimulus as a function of their internal states at the time the stimulus is encountered and as a function of their experience with the stimulus. (p.1593).

The concept of reward would seem to be particularly fuzzy in the context of research into action acquisition and often the term 'reinforcer' is used in its place. Natural examples of action acquisition do not involve operant chambers and food hoppers. Indeed, it can be difficult even to identify what the reinforcer is in a given situation, let alone make a decision as to whether or not it is best described as a reward or reinforcer. It is during the early development of higher mammals that we can observe the fastest rate of action acquisition and yet the rewards in operation during intensive exploratory and play behaviour are particularly obscure. There are also a number of non-play examples of responses being maintained by outcomes that have no obvious relationship to basic survival or reproductive requirements of an animal. For example, light flashes which are delivered in a way that is temporally contiguous and contingent with bar touching in mice result in increased responding and extinction effects when the light is not longer provided (Kish, 1955). Presumably, the ultimate advantage of a learning system that does not necessarily require reinforcement in the form of economic rewards (or stimuli predicting economic rewards) is that the organism can build up a behavioural repertoire in advance of needing to use the actions for direct survival reasons. Singh, Lewis, Barto and Sorg (2010) capture this distant relationship between reinforcement and ultimate survival benefits in their account of intrinsically motivated learning: "there are no hard and fast features distinguishing intrinsic and extrinsic reward computationally. Rather, the directness of the relationship between rewarding behavior and evolutionary success varies along a continuum" (p.70). In other words, the more distal the apparent relationship between a behaviour and genetic fitness, the more we can describe the properties of any reward involved as being intrinsic as opposed to extrinsic.

The idea that some behaviour is intrinsically motivated and that the extent to which it is intrinsic can be placed on a continuum of how distal is its relationship to

evolutionary success, ties in with theories about changes in the way dopamine has mediated behaviour over evolutionary time. Hills (2006) speculates that the effect of dopamine on behaviour in higher mammals is not so different to its effects on foraging and spatial search for food rewards in much more primitive animals; what has changed is rather the range of things that modern animals are capable of treating as goals and rewards:

> Molecular machinery that initially evolved for the control of foraging and goal-directed behavior was co-opted over evolutionary time to modulate the control of goal-directed cognition. What was once foraging in a physical space for tangible resources became, over evolutionary time, foraging in cognitive space for information related to those resources. (p.4).

In other words, higher mammals are no longer restricted to motivation through primary rewards and are able, via the effects of, amongst other things, dopamine, to treat abstract objects, events and perhaps even memories as goals. And yet, responses to these abstract goals should be much the same as they are still mediated by the same underlying brain processes. All of this would seem to support Redgrave and Gurney's (2006) account of the function of dopamine neurons because it suggests that there has been a move away from purely appetitive stimuli towards stimuli that are interesting to an animal for other reasons. As Berridge and Robinson (1998) have suggested, the effect of dopamine might be to make animals 'want' something without necessarily 'liking' it, a particularly useful trait for any animal that is designed to add lots of actions to its behavioural repertoire.

The reinforcement that will feature in the experiments to follow tends to involve simple discrete sounds or visual stimuli. These will be referred to as instances of reinforcement and, following Redgrave and Gurney (2006) and Redgrave et al. (2008), it will be assumed that these novel events will be reinforcing for the purposes of action acquisition. This is in the tradition of behaviourism where Ferster and Skinner (1957), for example, defined a reinforcer as anything that increases the probability of a specific response being emitted. From this perspective, then, there is no sense in which we should talk about rewards and reinforcers as being separate things; they either result in reinforcement or they do not.

Methodological requirements for a behavioural test of action acquisition

Thorndike's (1911) escape paradigm, described above, is an excellent methodological precedent for the study of action acquisition. However, whilst it is an elegant approach, it has many limitations, both from a quantitative and a practical point of view. Consequently, it provides a useful starting point for identifying the characteristics we would ideally find (or avoid) in a good action acquisition paradigm. Perhaps the biggest single limitation of the puzzle-box paradigm is that it makes data collection difficult. Whilst it was possible to extract a useful one-dimensional performance metric in the form of escape time, it was not possible for Thorndike to quantitatively deal with the more complex aspects of an animal's behaviour, so metrics such as the distance travelled by an animal or the time spent in the proximity of the lever were not readily available. Furthermore, the task was not easy to automate: whilst it would not be difficult to automatically record the time from when the animal first entered the puzzle-box to the point at which it effected its escape, each new trial still required experimenter intervention in order to put the animal back into the puzzle box. A further limitation is that the puzzle box scenario is clearly not suited to testing human subjects: any such endeavour would be unwieldy, requiring a large and complex, yet safe, puzzle environment. Perhaps more importantly, though, is the problem of hiding the puzzle mechanism from a human participant. When using animal subjects it is possible to rely on the animal's lack of insight into the action-outcome mechanism as a way of forcing the animal to rely on the reinforcement of recent motor output rather than simply jumping to a solution. This is far more difficult with human participants. It is therefore important that any new paradigm aimed at investigating action acquisition in humans, involves some means of obscuring the method of eliciting the outcome so that participants are forced to learn and not merely perform the required action. Finally, the puzzle-box scenario was limited from the point of view of repeated testing of individual subjects: repeated measures for Thorndike meant physically changing the apparatus that the animals were required to escape from. Such a high maintenance approach to repeated measures investigations is clearly something to be avoided if at all possible.

Thorndike's line-drawing paradigm

Thorndike (1927) addressed some of the above limitations with an experiment that tested the ability of human participants to draw lines of particular lengths. Blindfolded participants were asked to draw lines as close to a target length as possible. In the first phase of a testing session all of the lines were drawn without any feedback. During the second phase, feedback was provided in the form of the verbal responses "right" and "wrong", depending on whether or not the length of the line was close enough to the target length. In the third and final phase, the participants were tested once again without any feedback. Thorndike was then able to calculate the improvement from phase one to three as a means of describing the amount of learning that had taken place. For the present purposes, the line-drawing paradigm solves some of the short-comings of the puzzle box approach. From a practical point of view, the approach lends itself to the testing of human participants and also to the automation of the learning procedures. Whilst Thorndike himself was not able to implement the experiment in an automated fashion, it is nonetheless easy to appreciate how we might translate his procedure into a computerised version that could be run with relatively little input from the experimenter. Furthermore, the drawing approach makes it simple to run repeated measures designs. Participants can, of course, be tested on different lengths of lines, but clearly they can also be tested on other shapes and symbols.

However, whilst the line-drawing paradigm solves some of the problems associated with the puzzle box, it also has some limitations from the point of view of the current research. At first, it appears like Thorndike was able to obscure the method of eliciting the outcome simply by blindfolding the participants: the task demands are sufficiently straightforward that they can be achieved whilst wearing a blindfold so the blindfold can be employed as a means of forcing the participants to learn to draw the lines through 'feel' rather than simply remembering what the correct length of line looks like. However, whilst the blindfold certainly serves to make the task more difficult, the task is still primarily one of executing a pre-specified movement more accurately. This point is easier to appreciate when we consider the kind of behavioural variance that

was being produced in the experiment. In theory, Thorndike was able to extract an extremely rich data set from the line drawing paradigm: participants produced hundreds of two dimensional movement traces and, in a modern computerised version of the task, it would be easy to also collect accurate timing data. And yet, irrespective of how complete the record of behaviour is, the task is just too heavily constrained to make full use of the record. There is only one way to complete a trial; no degrees of freedom other than the length of the line remain open for the participant to express behavioural variance. Failure to draw a line is simply failure to perform the task and the relative quality of a line (e.g. its straightness) is not taken into account. If we consider the puzzle box scenario it is clear that whilst each trial must ultimately end with the cat falling upon the lever in order to open the door, the means by which the cat achieves this end are free to vary as the cat chooses. In the line drawing task, there is no sense in which the participant can achieve the drawing of a line in an efficient or an inefficient manner: efficiency is not a parameter that is free to vary. In this sense, the line-drawing paradigm is somewhat more similar to a supervised learning task as opposed to a reinforcement learning task. This point is particularly underlined when we consider that feedback, whether it indicates a correct or an incorrect response, is provided regardless of what the participant does: the presence or absence of feedback is not contingent on performance, rather it is a description (albeit a crude one) of performance.

The serial response time task

An alternative approach to the study of action learning in humans is to record sequence learning in the form of multiple button presses. The serial response time task (SRTT) (Nissen & Bullemer, 1987) is designed to measure motor sequence learning at different levels of attention. In a typical set-up, participants press one of four buttons on a keypad in response to visual cues that occur at one of four corresponding locations on a computer monitor. As the buttons are pressed, the time between the signal appearing on the monitor and the associated button being pressed is recorded. Depending on the particular methodological setup, repeating sequences of button presses are hidden from attention by asking participants to perform a concurrent secondary task or by embedding them within random, non-repeating sequences

(Willingham, Salidis & Gabrieli, 2002). The aim is therefore to investigate "whether it is necessary to attend to a stimulus event in order to remember it later" (Nissen & Bullemer, 1987, p.1) and the relationship between learning and attention is effectively distilled down to a single vector of response time.

Researchers who employ the SRTT tend to be interested in the question of whether or not learning is available to awareness rather than the particular type of learning that is taking place. As a result, the task is difficult to categorise from a theoretical point of view. It is similar to a supervised learning paradigm in the sense that all of the targets and movements are visible to participants, so it is possible for a participant to monitor their performance at all times. However, it differs from a typical supervised learning task in that learning can only be expressed in terms of speed. Accuracy of movement has little freedom to vary: clearly it is possible for a participant to miss a button or press the wrong button, but there is no sense in which a correct button-press can be either accurate or inaccurate. In this respect, it is rather more like an action acquisition paradigm in that learning can be expressed in terms of the speed at which an action is performed. But, there are some differences here too. For instance, the learning environment is so highly constrained that participants must always produce the correct sequence of button presses; the way that a sequence is executed, therefore, cannot vary in structure. Furthermore, nothing is contingent on improved performance at the sequence level. Loosely speaking, we could say that the action that is being learnt is the sequence of button-presses; however this isn't itself a unit of reinforcement. The event that is contingent on behaviour is the disappearance of the current set of cues on the monitor. These cues disappear when the correct button is pressed, but they are unaffected by previous button-presses in the sequence. Consequently, whilst the SRTT is an attractive research option from a practical point of view and is an example of action learning, it does not allow the freedom of movement that will be a focus of the present research.

The Morris water maze

The Morris water maze (Morris, 1981) was developed as a way of investigating spatial memory in rats and is an example of a learning paradigm that allows for high variations

in the efficiency with which trials are solved. The search environment consists of a circular pool of water rendered opaque by the addition of milk. Rats are introduced into the pool whereupon they begin to search for a means of escape. The key to the paradigm is a platform that can be introduced just beneath the surface of the water. This platform is hidden from view and yet provides the animals with somewhere to stand if they are able to find it, enabling Morris to hide the means of escape without the use of physical barriers. The behaviour of the rats is recorded using a camera placed over the pool of water. From this recording, it is possible to extract a two-dimensional movement trace of the animals' attempts at finding the hidden platform; all manner of metrics can then be extracted and submitted to analysis. Of particular interest to Morris was path directionality. Previous walled maze paradigms physically restricted the ability of animals to express the directionality of their movements to a few particular choice points. The Morris water maze removes these constraints so that animals are not forced into correct or incorrect choices but can instead display their understanding of the environment in terms of the efficiency of their movement traces.

From the point of view of the current research, the Morris water maze solves some of the limitations of Thorndike's (1911) puzzle box escape paradigm described earlier. The issue of rich data collection is solved by the use of a video camera, which can capture the two-dimensional movement trace of the animals. This trace is rich enough to provide the experimenter with a choice of metrics to investigate and yet it is also constrained enough that the data are not overly difficult to interpret. The advantage of rich data is that it allows for multiple descriptions of the data. Whilst escape time scores might only tell us about raw performance for a given trial, other aspects of the data can convey information about particular processes: reduced speed of movement, for example, can indicate impaired sensorimotor function (Vorhees & Williams, 2006). Importantly, the paradigm constrains behaviour to two dimensions in a credible way. It doesn't take much of a stretch to imagine how the task demands of a trial in the water maze might translate into wild behaviour. The data are constrained to two dimensions not because the movements are only recorded from one angle but because the environment provided to the animals only allows the expression of behaviour in a two dimensional plane. This was not the case with Thorndike's task because locating the lever and pressing the lever were two separate components of the overall action.

Whilst it might be possible to describe the cat's movement towards the lever in terms of a two-dimensional trace, the cat's efforts at operating the lever must be described in a different way.

A further advantage to the water maze is that it solves the problem of hiding the mechanism from the animal performing the task so that, during hidden trials, the animal is forced to discover the location of the platform through its own search behaviour; it can only solve the problem of how to move towards the platform by virtue of its memory for the location; it can never perceive the location directly. Because of this location finding structure, the water maze also makes it easy to employ repeated measures designs. The starting position of the animal and the position of the platforms can be moved easily with each new position presenting a new learning scenario for the animal, whilst keeping the structure of the task consistent.

A final advantage of the Morris water maze, from the present perspective, is that it provides a means of running control conditions that share most of the physical demands but fewer of the memory and cognition demands found in the experimental conditions. Whilst a hidden platform provides challenges to memory and cognition, a visible platform can also be made available which allows experimenters to run control tests to ensure that animals are capable, and willing, to perform the basic physical dimension of the task (D'Hooge & De Deyn, 2001). In this way, experimenters can check to make sure that any manipulation hasn't unduly affected the animals' motivation or basic motor abilities.

Whilst the water maze has many of the attributes that we might want from an action acquisition paradigm, it is also limited in some important respects. From a purely practical point of view, the task is not scalable from rodent subjects up to human participants and it requires substantial experimenter input. Another issue is the number of movements required in order to complete a trial. It is not the case that the rats were able to complete a trial from start to finish in a handful of continuous movements. Optimal performance in the water maze still requires that the animal swim for many seconds and produce all of the movements involved in this activity. Even in the absence of a strict definition of what constitutes an action, it seems clear

that a typical water maze trial does not so clearly represent what we normally mean by an action as, say, a lever press or the unscrewing of a cap. A problem that is also related to the length of the trials is the difficulty in moving away from simple location finding to more complex gestural movements that we might associate with action acquisition. Clearly the availability of the platform in the water maze could be made contingent on the rats moving through certain other portions of the pool first, but the time it would take to discover and perform such gestures through swimming would likely make this kind of learning extremely difficult. The water maze is therefore a useful starting point for developing an action acquisition paradigm. If we bear in mind some of these limitations this general concept will form the basis of the action acquisition paradigm that is central to the current investigation.

Putting the pieces together: the joystick task

With these theoretical, methodological and statistical considerations in mind, the following chapters will show how this information was used to design and develop a task with some of the theoretical and practical attributes of existing paradigms whilst also attempting to capture an intuitive sense of what constitutes an action. The first version of the task is detailed in chapter 2.

Chapter 2: the joystick task and reinforcement delay

Having set out the motivations for the current research and situated the topic of action acquisition relative to other research areas, the current chapter provides a description of the joystick task. Following this, the importance of delay as an experimental variable for the investigation of action acquisition is discussed and a study investigating the impact of delayed audio and visual reinforcement is described.

The joystick task

In an attempt to achieve a balance between freedom and constraints of movement, the current task deliberately emulates the Morris water maze (Morris, 1981) by providing a learning environment (albeit a virtual one) with similar characteristics. A computer defines a two-dimensional, square environment and a joystick is the means by which a user can search this space (figure 2.1). Whilst, technically speaking, any input device could have been used, a joystick was chosen for several reasons. Most important of these is that joysticks provide a good physical representation of a two-dimensional search space. The travel of the joystick is constrained by the aperture within which the moveable part of the apparatus is moved so that the user can feel the limits of the space that they have to explore and are physically restricted from going beyond it. The joystick also provides the benefit that the moveable part of the apparatus cannot be separated from the base (unlike a mouse or a stylus, for example) so the participant's understanding of the search space (and our ability to track their movements within it) are consistent because all movements are relative to the stable base of the joystick. The result is that the user has an intuitive understanding of which of their movements are task relevant because they can be sure that only movements of the joystick will translate into movements within the search environment. Finally, the joystick offers a great practical benefit to the researcher in that the spring allows it to return to a consistent central starting position automatically.

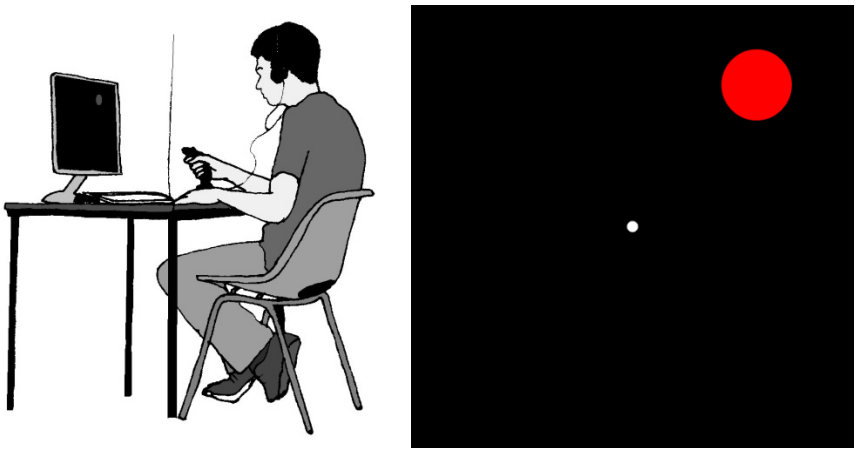

Figure 2.1 The basic design of the task was based on the Morris water maze, with a joystick being used to search a two-dimensional square space for a circular reinforced area (coloured red above). The small white circle marks the centre point to which the joystick returns when released.

A virtual search environment offers huge flexibility to the experimenter. The means of eliciting reinforcement in the joystick task was based on the idea of escape platforms in the Morris water maze (Morris, 1981). Circular portions of the search space, instead of providing a means of escape from a pool of water, simply cause a stimulus event (reinforcement) to occur when the joystick is moved into them. In this virtual environment, we might lose the ability to elicit primary motivational drives (such as a desire to escape from water) that align with our experimental goals, but the task provides a very intuitive game-like experience and we gain much in terms of flexibility with the ability to vary the goals easily. For instance, the range of reinforcement contingencies that can be set up is essentially limitless and the rules can be changed at any time, offering the additional benefit of allowing the experimenter to present the same individual with many different learning situations within a single experiment. Furthermore, it should also be possible to scale task difficulty. The benefit of this lies in the opportunity it presents for calibrating task difficulty when testing clinical populations or even non-human animals.

In terms of data collection, the joystick task provides 3 basic elements, x coordinates, y coordinates and time. From these data it is possible to calculate the distance, duration and speed of movement as well as potentially deriving a host of other higher level metrics that focus on the shape or style of movement. The focus of the current study is on the simpler performance metrics such as distance, however, even then, the

richness of the data still provide the researcher with useful information. For instance, it is possible to play back all of the movements made during a particular trial in the form of a continuous movement trace; it is also possible to generate still images featuring all or part of the total movement trace (figure 2.2). The advantage of this is that it alerts the researcher to problems with the task, enabling us to ask the question of whether participants are performing the task in the intended way and, if not, what might be done to develop the task in future. A further advantage is that the traces can provide information about which formal analysis might best represent the learning and behaviour that occurred. For example, and as we will see later in this chapter, the movement data can alert us to sources of variance that make one performance metric difficult to analyse, but that are largely absent from another performance metric and thus can allow us to make better choices as to how we describe the data quantitatively.

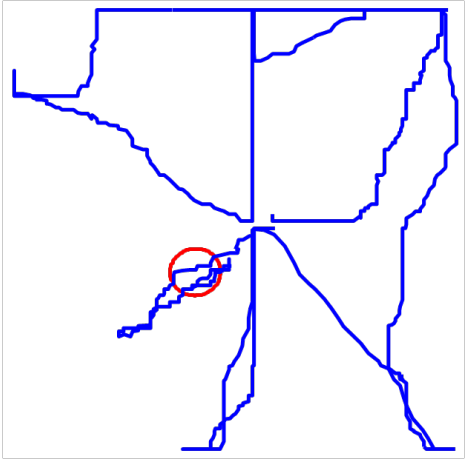

Figure 2.2 An example movement trace taken from the data for a trial in experiment 1. The ability to inspect still images and even play back trials to observe movement properties is a useful resource for both development of the task and for performing checks when conducting formal analyses of the data.

Reinforcement delay

Differing sensitivities to delay depending on the type of motor task

The effect of delay on reinforcement learning is dependent on a number of factors from the agent's experience of the learning environment (Dickinson, Watt & Griffiths, 1992) to the relative contribution of stimulus-response versus goal-directed learning

systems (Cardinal, 2006). It is, therefore, difficult to build up a picture of how learning responds to differing delay durations, partly because of these factors and partly because of how results are presented: researchers are free to report the point at which learning is abolished, the point at which an effect of delay is detected and everything in between. In other words, as Snycerski, Laraway and Poling (2005) have noted, the success criterion can have a huge effect on how results are interpreted.

Where there does appear to be a clear split, however, is in the overall sensitivity to delay seen in motor control and motor adaptation paradigms versus traditional reinforcement learning paradigms. It has been shown in human subjects, for example, that in visual tracking tasks, feedback delays as short as 300 ms can have a large impact on performance (Foulkes & Miall, 2000; Miall & Jackson, 2006; Miall, Weir, & Stein, 1985); tasks involving adaptation to visually displaced targets are found to be equally sensitive to delay (Held, Efstathiou, & Greene, 1966), with some demonstrating an effect at just 50 ms (Kitazawa et al., 1995). In reinforcement learning paradigms, learning and performance does not appear to be quite as sensitive to the effects of delay, with studies showing an effect at somewhere between 1 to 2 s (Elsner & Hommel, 2004; Shanks & Dickinson, 1991; Shanks, Pearson and Dickinson, 1989), although there is a notable absence of data to indicate the duration at which the effect of delay first appears.

However, whilst there is an apparent difference in the sensitivity to delay of reinforcement learning and supervised learning paradigms, there is also a difference in the type of experimental task employed. The former tend to employ simple button-pressing response mechanisms, whilst the latter rely on more complex reaching movements and target pursuit, which require precise motor control. The current experiments will investigate the effect of delay in a task that involves rich motor behaviour with delivery that is more typical of reinforcement learning.

Short latency dopaminergic activity and contamination

According to Redgrave and Gurney (2006), the advantage of time-stamping so quickly following the incidence of a novel stimulus (~100 ms) is that this reduces the

opportunity for causally irrelevant movements to be reinforced. Furthermore, such a rapid response effectively precludes the contribution to motor learning of any movements that are a direct consequence of the novel stimulus itself. The disadvantage of such low-latency activity, however, is that the response is necessarily an indiscriminate one. As discussed in chapter one, the signal occurs so quickly after an event that it happens before the stimulus can be brought onto the fovea and at the same time as cortical processing required for identifying stimulus characteristics.

These points raise the following question: why could the time-stamp not have evolved with a longer delay? Delaying the dopamine response for several hundred milliseconds would allow for a saccade to take place, plus some subsequent cortical processing of the event. Contamination could then be avoided simply by discounting the motor activity that occurred whilst this extra sensory processing was taking place. Similar discounting mechanisms for coping with internal biological transmission delays are commonly proposed in the supervised learning literature (e.g. Miall, Weir, Wolpert & Stein, 1993). However, the problem with this idea is that the time required for discounting would not be stable. Unlike simple internal transmission time delays, which change only gradually over time due to growth, ageing and other physiological factors, the length of time taken to identify a stimulus is necessarily highly variable from one instance to the next. It isn't clear how any neural mechanism could make a decision as to when the stimulus could be considered adequately identified, or how this information could be incorporated into such an evolutionarily ancient neural mechanism. It seems more likely, therefore, that an indiscriminate time-stamp evolved because it is sufficient for most purposes and has been selected in favour of a more resource-intensive 'sample and hold' discounting system.

If the dopamine time-stamp, as described by Redgrave and Gurney (2006) and Redgrave et al. (2008), is low latency specifically to avoid contamination, then the learning mechanism that it contributes to should be highly sensitive to delays of reinforcement. One way to investigate this is to test the sensitivity to temporal alignment of learning procedures that are believed to be dependent on the dopamine time-stamp. Whilst we can't experimentally manipulate the post-stimulus, pre-dopamine delay, it is a straightforward task to introduce post-movement, pre-stimulus

delays into the action-effect chain. In other words, temporal alignment of motor output and sensory feedback can be manipulated by interposing a delay between the point at which an animal causes an event to occur and the moment that the occurrence of this event is presented to the animal as a stimulus, thus approximating a learning system with a longer latency time-stamp. In theory, the temporally contiguous yet non-contingent motor activity that occurs during the delay period (that is to say, the artificial delay plus the 100 ms biological delay) would act as an unavoidable contaminant in the animal's learning system, making it more difficult to identify causally relevant portions of motor output. Indeed, Redgrave and Gurney (2006) and Redgrave et al. (2008) point to research demonstrating the detrimental effects of delay on reinforcement learning (Dickinson, 2001; Elsner & Hommel, 2004; Schultz, 2006) to support their explanation for the evolution of such a low latency system; however, the delay durations at which the effects are found in these studies start at 0.5 s (Schultz, 2006, original source: Hollerman & Schultz, 1998), much longer than the latency of the dopamine activity in question and sufficiently long that we might assume the delay would also interfere with post-saccadic (and high-level post-identification) learning processes.

If we are to believe that the phasic dopamine timestamp evolved as a means of reducing motor contamination and therefore reducing the credit assignment problem, then we would expect much shorter delays than the 0.5 s cited by Redgrave et al. to have a severe impact on learning. Specifically, there should be a measurable effect at shorter durations than that which would be required for identification of the stimulus to occur. If this were not the case, then it would be difficult to explain why the system doesn't simply delay the time-stamp for a similar duration, given the potential benefits this would bring as the dopamine response would not need to be an indiscriminate one. It seems reasonable to assume that for identification of the stimulus to occur, approximately 150 ms would be required for a saccade (Hikosaka & Wurtz, 1983; Jay & Sparks, 1987) to bring the stimulus on to the fovea, plus an additional 80 to 100 ms for basic cortical processing of the new image (Rousselet, Thorpe & Fabre-Thorpe, 2004; Thorpe & Fabre-Thorpe, 2001), giving a total of around 250 ms, i.e. 150 ms longer than the dopamine time stamp. In other words, a detrimental impact of delays less than 150 ms in duration would provide more convincing support for the theory than the 0.5

s durations referred to in the literature.

Experiment 1: within-trial repetition of audio reinforcement

Experiment 1 was developed to investigate the effects of reinforcement delay on the ability of people to home in on the location of a reinforced area within a two-dimensional search space. The first reason for the task was to assess where the sensitivity to delay lies in a rich motor task featuring discrete and qualitative instances of reinforcement. The results will provide insight into whether the key factor in the effect of delay has to do with the motor demands of the task, such as the speed and the accuracy of movements, or whether the type of reinforcement and the type of learning are the most important factors. In other words, does delay sensitivity lie beneath 0.5 s, as we might expect based on previous studies featuring tasks with similar physical demands, or does it lie somewhere in excess of 1 s, as we might expect based on physically very different reinforcement learning paradigms? The second question also concerns sensitivity to delay. If task performance ultimately relies on the activity of midbrain dopamine neurons and this activity is short latency specifically to avoid contamination with non-contingent motor output, then we would expect the system to be highly sensitive to artificial delays. It was, therefore, a goal of the current experiments to investigate the effect of delays of less than 0.5 s, including delays under the 150 ms that would be required for basic stimulus identification.

Method

Participants

27 people (25 female) participated in all conditions of this study. Ages ranged from 18 to 24 years with a mean age of 19 (*SD* = 1.3 years). Participants were all undergraduate students at the University of Sheffield who took part in return for credits in the department's research participation scheme. All subjects were naive to the purpose of the experiment.

Apparatus

The experiment program was written using Matlab (Version 2007) with the Psychophysics Toolbox extension. A commercial joystick (Logitech extreme 3D pro joystick, P/N: 863225-1000) was used as the input device. These tools were used for all experiments described in this thesis.

Defining the search environment and the reinforced area

The search space was defined as a square that was 1024 by 1024 units in size, which corresponded to the limits of the joystick's travel (the joystick movements were physically restricted by a square aperture at the base of the stick). Movements of the joystick mapped on to movements within the search space in a 1 to 1 fashion, with the joystick starting in the centre of the search space at the beginning of each trial. Once released from the grip of a participant, the joystick was able to return to the centre of the search space within a tolerance of 10 units, by virtue of a built-in spring mechanism.

Different sizes of reinforced area (hotspots) were tested during development and piloting of the task. The size was eventually set to occupy 0.91% of the overall search space based on finding a balance between making the task sufficiently difficult to provide useful data and the practical limitations of running multiple trials that were not time-limited; there was no theoretical reason for choosing this specific size of hotspot. At the beginning of every new trial, the centre of the hotspot was positioned randomly on an annulus placed centrally within the search space (figure 2.2). The inner edge of the annulus was exactly 1 times the diameter of the hotspot from the centre of the search space. At its closest point, the outer edge of the annulus equivalent to the radius of the hotspot from edge of the search space. The reason for these dimensions was to ensure that the hotspot never overlapped the central starting point or the outer edge of the search space.

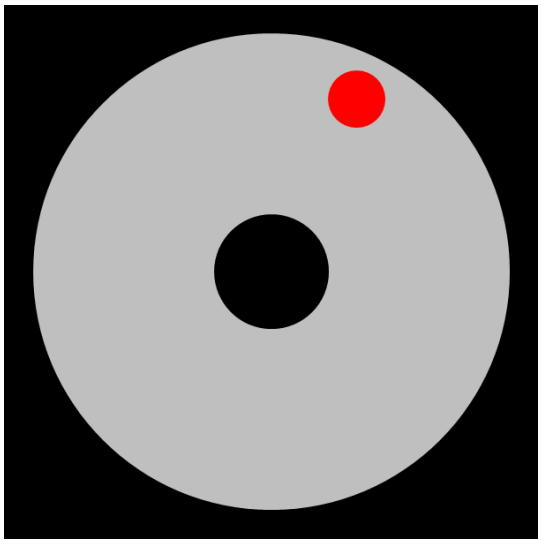

Figure 2.3 Experimental search space and hotspot positioning. The black area represents the search space and the grey annulus represents the area of the search space in which the centre of the hotspot (shown in red) could be randomly positioned at the start of each trial. The diagram is drawn to scale.

Defining reinforcement and the learning criterion

Any movement of the joystick into the hotspot region of the search space was defined as a hit and resulted in a short 'pip' sound of 10 ms duration (the reinforcement) – holding the joystick over the hotspot resulted in a rapid series of these discrete pip sounds (i.e. not a continuous sound). An audio stimulus was chosen for the first experiment in an attempt to reduce any focus the participants might have on the computer monitor and increase the feeling that the task was to find the correct position to hold the joystick in rather than a location on a screen; it was considered that if participants thought about the task in terms of points on the computer monitor, they would be more likely to take a cognitive/strategic approach to the task and their performance would then be less representative of the kind of learning under investigation. Another reason for choosing an audio stimulus was that such stimuli convey much less in the way of spatial information, misleading or otherwise. Generating a single hit was not sufficient to bring an end to a trial. Instead, a learning criterion was used to determine whether a participant had learnt the location of the hotspot. The learning criterion was defined as the number of hits required within 1 s in order to bring an end to a trial. Like hotspot size, the learning criterion was set using information gained from pilot tests in order to balance task difficulty (more hits per second meant the threshold was harder to meet) against better verification of learning (more hits per second requires a participant to demonstrate better learning of the

hotspot location). The criterion was set at 15 hits per second. From an individual participant's perspective the aim in a given trial was, therefore, to find the hotspot and try to maintain the position of the joystick over this region until having achieved 15 hits (and audio reinforcements) in a second.

Delay

Delayed reinforcement was achieved by interposing a delay between the point at which a participant moved into the reinforced area and the point at which a hit was recorded and the audio reinforcement delivered. In all, 6 delay conditions were chosen: 0-ms, 75-ms, 150-ms, 225-ms, 300-ms and 375-ms. Despite the findings from previous reinforcement learning paradigms showing a sensitivity to delay starting at around 500 ms, pilot test revealed that the experiment was extremely sensitive to delay and that delays of 500 ms would have made trials difficult to complete. Consequently the 6 delay conditions were chosen to provide a balance between task difficulty and the ability to discern the point at which an effect of delay started. It was necessary to include a refractory period of 25 ms after each instance of reinforcement during which another stimulus (and hit) could not occur; this was to allow sufficient time to prepare and play each audio stimulus and enabled us to ensure that we could provide discrete stimuli that didn't blend into one another and that would be repeated at regular intervals at their fastest rate.

Procedure

Participants sat at a desk in front of the joystick and a 19 inch computer monitor. Before starting the experimental program, the task was briefly described verbally with the task goal being phrased in terms of "finding the correct position to place the joystick in" rather than, say, "search for the correct location". In order to reduce the tendency of individuals to simply move the joystick around the very edges of its travel (something flagged up during pilot sessions), participants were encouraged to explore the whole range of joystick movements. Finally, the participants were told that the experiment involved no deception and that the correct position could always be found. This guidance was included because pilot tests revealed the task to be difficult and the lack of feedback might lead sceptical participants to believe that a given trial was

impossible, based on their experience of participating in other psychology experiments featuring deception. Following the verbal instructions, the experiment program was then started and the participants were asked to follow the onscreen instructions (see appendix 1). After reading the instructions, 3 practice trials commenced automatically. The practice trials involved no reinforcement delay and, as with all trials in the experiment, no feedback or screen graphics were provided during the trial (the monitor was kept black until the end of a trial). Once the practice trials were completed the experimental trials began and participants were left to complete all 18 trials.

Design

A repeated measures design was used. The independent variable was reinforcement delay and it included 6 conditions: 0-ms, 75-ms, 150-ms, 225-ms, 300-ms and 375-ms. Each experimental session was made up of 21 trials: 3 of which were practice trials (involving no delay); the remaining 18 were experimental trials. The experimental trials were presented in 3 batches of 6, such that all 6 conditions were experienced in each of the three batches. This was done to ensure that the 3 attempts at a particular condition were spread out over the full testing session. Each level of the independent variable was, therefore, experienced a total of 3 times with the order of presentation being randomly shuffled within each of the 3 batches.

Results

The post discovery period

For the purposes of analysis, each trial was treated as occurring in two phases: pre-discovery and post-discovery. The pre-discovery period lasts from the start of the trial to the first instance of reinforcement: it is the period during which the participant is naïve as to the position of the hotspot. The post-discovery period lasts from the first instance of reinforcement to the end of the trial; this period is of primary interest because it includes all instances of reinforcement and is, therefore, the period that is sensitive to the effects of the independent variable.

Distance travelled during the post-discovery period

Distance was identified as the baseline metric of performance and measured in search space units. Better performance should be reflected in a participant travelling a relatively short distance over the course of a trial. In other words, a top performing individual should be able to achieve the required hit rate with comparatively little movement of the joystick. Each participant had 3 attempts at each delay condition and it was the mean of these 3 distances that was submitted to analysis.

It was anticipated based on the recommendation of Keene (1995), that log-transformed data would provide the most representative picture of the learning and behaviour under investigation, when analysed using parametric tests. However, all data in this thesis also underwent tests of normality (Shapiro-Wilk before and after transformation) as well as checks on the relationship between standard deviations and means (i.e. whether standard deviations increased with higher mean values) and inspection of the frequency distributions in all cases to ensure that the data submitted to analysis were suitable for parametric tests. Due to the open ended nature of the tasks described, most of the data in this thesis were found to be positively skewed and were corrected using a log transformation (base-10). On each occasion throughout, this is indicated in the relevant results section.

A one-way repeated-measures analysis of variance (ANOVA), with 6 levels of delay, was conducted on the log-transformed data. Mauchly's test indicated that the assumption of sphericity had been violated ($\varkappa^2$ (14) = 27.54, $p < .05$; therefore degrees of freedom were corrected using a Greenhouse-Geisser estimate of sphericity (ε = .74). The results showed that there was a significant effect of reinforcement delay on the distance travelled during the post-discovery period, $F(3.7, 88.69) = 6.87$, $p < .001$. Figure 2.3 shows that the effect of delay was to increase the duration of the post-discovery period. Bonferroni corrected post hoc *t*-tests revealed that the 0-ms condition differed significantly from the 75-ms condition ($t(26) = 3.11$, $p < .05$) and that the 75-ms condition did not differ significantly from any other conditions ($p > .05$).

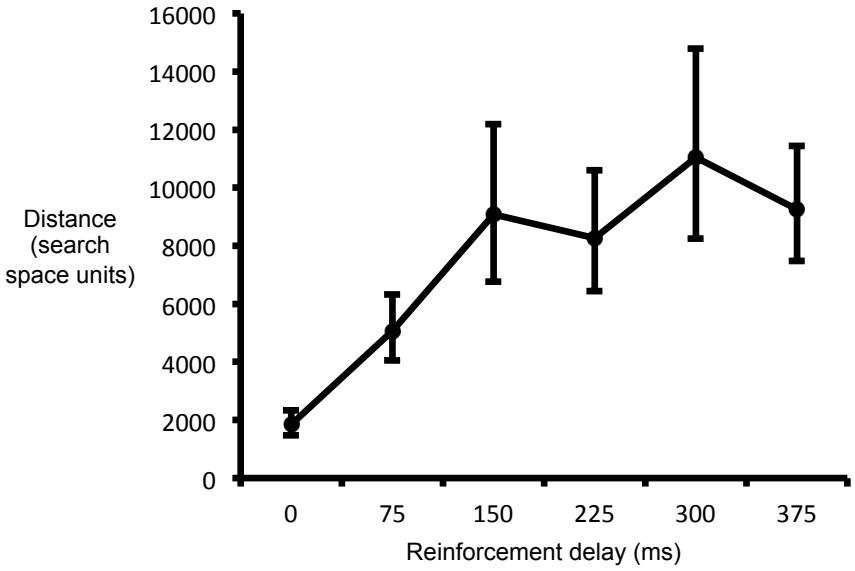

Figure 2.4 Mean distance travelled during the post-discovery period (and standard error) for the 6 levels of delayed audio reinforcement. Values are back-transformed from the log transformation.

Instances of reinforcement (hits) during the post-discovery period

As expected, the distance data were sensitive to an effect of delay but contained a large amount of variance, especially in the longer delay conditions. The ability to play back individual movement traces made it possible to look for potential sources of variance in the behaviour. One issue identified was that participants would often move into the reinforced area and receive reinforcement but then seemingly lose their way, resulting in large distance scores, despite the fact that on returning the correct area they were then able to home in on the hotspot relatively quickly. A way of reducing the impact of these events is to use the number of hits as a performance metric. An early hit followed by lots of searching does not impact the total number of hits as much as it would influence other metrics such as time or distance. Consequently, an analysis was conducted on the number of hits recorded during a trial: that is to say, the number of instances of reinforcement required for a participant to reach the learning criterion. With this metric, a top performing individual would require fewer hits (and, accordingly, fewer instances of reinforcement) in order to achieve the required hit rate.

Once again, the mean number of hits for each delay condition was calculated from the 3 attempts made by each participant. A one-way repeated-measures ANOVA was conducted on the log-transformed data. Mauchly's test indicated that the assumption of sphericity had been violated ($\varkappa^2$ (14) = 30.69, $p < .05$; therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .68). The results show that there was a significant effect of reinforcement delay on the number of hits during the post-discovery period, $F(3.4, 81.47) = 13.88$, $p < .001$. Figure 2.4 shows that for longer reinforcement delays, more instances of reinforcement were required in order to bring a trial to an end. Bonferroni corrected post hoc *t*-tests revealed that the number of hits for the 0-ms condition differed significantly from the 75-ms condition ($t(26) = 4.12$, $p < .05$) and that the 75-ms condition differed significantly from the 150-ms condition ($t(26) = 3.05$, $p < .05$). A further comparison was made between the 150-ms and the 300-ms conditions, which was found to be non-significant ($t(26) = 1.9$, $p > .05$).

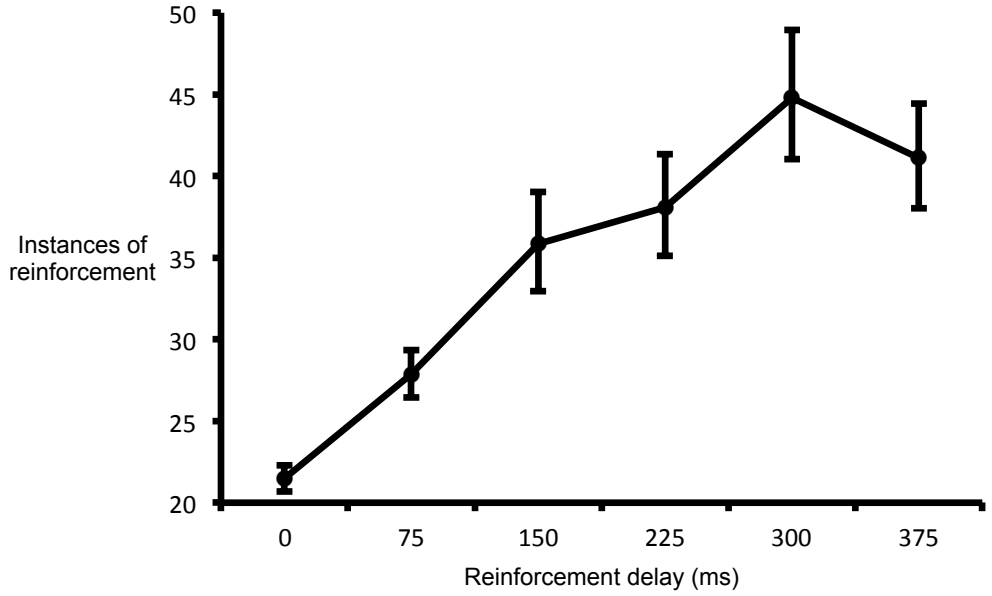

Figure 2.5 Mean number of hits (and standard error) for the 6 levels of delayed audio reinforcement. Values are back-transformed from the log transformation.

Speed during the post-discovery period

Whilst the length of the delay period was short in all conditions, the effects of the delay on the feel of the task were noticeable when performing the task as a non-naïve

participant. During development and piloting of the task, all of the delay conditions were tested and the general impression when experiencing a condition with delayed reinforcement was one of frustration, almost as if the target position was moving, particularly for delays in excess of 200 ms. This experience of frustration was echoed by many of the participants in the experiment, though none of them were able to guess what was being manipulated. The fact that the manipulation was noticeable, at least to some people, raises the possibility that participants might have adopted a strategy either consciously or unconsciously. If this was the case then a basic check is to test whether speed varies depending on the amount of delay experienced. To this end, speed was calculated in terms of screen units per second for the post-discovery period. A one-way repeated-measures ANOVA was conducted. The results (figure 2.5) showed that there was no effect of reinforcement delay on speed during the post-discovery period, $F(5, 120) = 1.23$, $p = .3$.

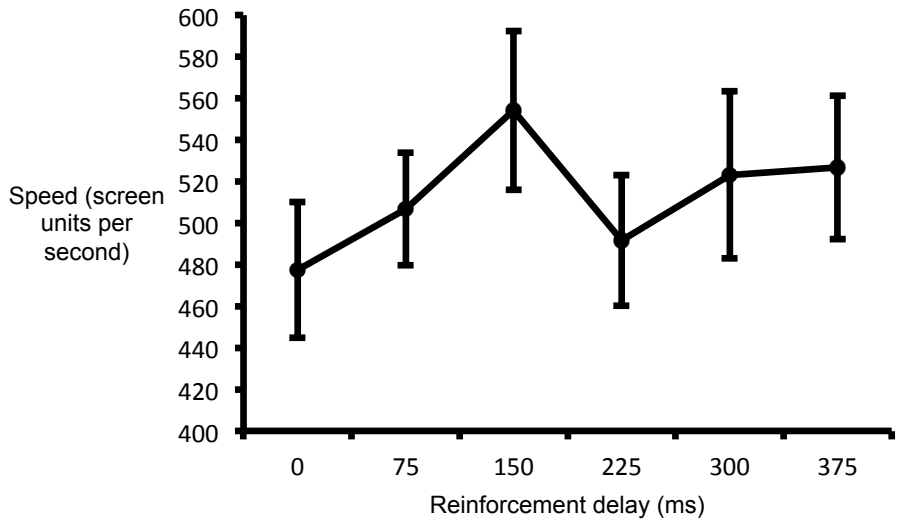

Figure 2.6 Mean speed (and standard error) for the 6 levels of delayed audio reinforcement.

Discussion

The results of experiment 1 revealed an effect of reinforcement delay on learning at just 75 ms. Even allowing for any additional delay caused by latencies in the apparatus, this still puts the effective delay at around 100 ms. If this type of learning is dependent on the short latency dopamine response, as argued by Redgrave and Gurney (2006) and Redgrave et al. (2008), then it is clear that there would be considerable costs to

having a longer latency dopamine response. The fact that this delay is shorter than the time it would take for a saccade and the necessary processing that would be required to identify the source suggests any costs associated with generating the dopamine response indiscriminately (i.e. without having first processed identifying characteristics of the stimulus) would be offset by the considerable costs associated with a more complex credit assignment problem. This finding, therefore, adds greater credibility to claims that the dopamine response is short latency as a means of minimising the amount of irrelevant contaminating motor and contextual input that is reinforced in the striatum following the occurrence of a novel stimulus (Redgrave & Gurney, 2006; Redgrave et al. 2008).

The sensitivity to delay in the current task is far higher than is generally found in reinforcement learning paradigms (Anderson & Elcoro, 2007; Black, Belluzzi & Stein, 1985; Elsner & Hommel, 2004; Hollerman & Schultz, 1998; Lattal & Gleeson, 1990; Okouchi, 2009; Renner, 1964; Shanks & Dickinson, 1991; Shanks et al. 1989; Sizemore & Lattal, 1977; Snycerski, Laraway, Huitema & Poling, 2004; Snycerski et al., 2005; Stubbs, 1969) and much more in line with what is reported for supervised learning paradigms (Foulkes & Miall, 2000; Held, Efstathiou, & Greene, 1966; Kitazawa, Kohno & Uka, 1995; Miall & Jackson, 2006; Miall, Weir, & Stein, 1985). Furthermore, comparisons between pairs of conditions revealed that the effect was not simply a delay versus no delay effect and that additional delays within the range tested had an added impact on performance. This is important because the distinction between the two types of learning tends to rely on the informational content of the feedback that is provided. However, the current results suggest that, in the case of delay, it is the specific task demands that are more important. This result widens the already large range of sensitivities to delay that have been observed in the reinforcement learning literature and calls into question the extent to which we can speak of the effects of delayed reinforcement on learning and response acquisition. The effects of delay are perhaps better described in terms of the task properties on which they are measured.

A useful property of the current paradigm was the resistance to the use of strategies on the part of the participants. The fact that there was no effect of delay on speed and that none of the participants guessed the independent variable suggests that people

weren't reacting systematically to delay by adopting strategies to cope with it. This is important because in supervised learning tasks, where targets are not hidden, researchers have found that people have a tendency to adopt strategies such as a 'move and wait' approach in order to cope with delay (Sheridan, 1993; Sheridan & Ferrell, 1963). In the current study, such strategies would have interfered with our ability to interpret the effect of delay in terms of the low-level neural mechanism offered by Redgrave and Gurney (2006) as an explanation for the acquisition of novel actions. However, it is nonetheless possible that the effect found in this experiment was associated with high-level cognitive/declarative processes and it is impossible to rule this out given the limitations of the current paradigm.

Having tested the paradigm with audio stimuli and finding it to be highly sensitive to the effects of delay, it was decided that the experiment should be repeated with visual stimuli to test whether the result was robust and also to ascertain whether the modality of the stimuli is important at such short delay durations. As will be explained in experiment 2, Redgrave and Gurney (2006) make particular reference to visual stimuli and the activity of the superior colliculus at less than 100 ms. Consequently, testing the effects with visual stimuli will provide further information relevant to this theory.

Experiment 2: within-trial repetition of visual reinforcement

Experiment 1 demonstrated a high sensitivity to short delays, consistent with Redgrave and Gurney's (2006) theory; however, it relied on audio stimuli for reinforcement. Redgrave and Gurney (2006) and Redgrave et al. (2008) make specific reference to the superior colliculus and visual stimuli as the source of sensory input on which the phasic dopamine response might depend. As noted by Redgrave et al., the claim is not that action learning is modality specific; rather the emphasis on visual input from the superior colliculus reflects the fact that most previous research has involved the use of visual stimuli. Response latencies in the superior colliculus are unique amongst visual areas in that they precede even the short latency dopamine response in the substantia nigra, indicating that they could provide the input that triggers this activity (Comoli et

al., 2003). Furthermore, it has been found that the superior colliculus has a greater ability to stimulate nigral neurons as compared to cortical visual areas (Comoli et al., 2003), thus adding support to the idea that these areas of the brain contribute to a single learning mechanism. In light of this, it was decided that a version of the task utilising visual stimuli should be run. This would not only provide data that is relevant to action learning mediated by input from the superior colliculus, but would allow for a comparison of learning dependent on stimuli from two different modalities. A final motivation for this version of the task was to pave the way for future research within this author's research group. Future experiments hope to compare action learning with visual stimuli to which the superior colliculus is minimally sensitive versus stimuli to which cortical visual areas are minimally sensitive.

Method

The method and design were identical to that employed in experiment 1 except that the reinforcement signal was visual instead of audio.

Participants

24 people (22 female) participated in all conditions of this experiment. Ages ranged from 18 to 23 years with a mean age of 19 (SD. = 1.4 years). Participants were all undergraduate students at the University of Sheffield who took part in return for credits in the department's research participation scheme. All subjects were naive to the purpose of the experiment.

Visual reinforcement stimuli

Any movement into the hotspot was reinforced by a short duration (17 ms) screen flash. A single flash consisted of the whole monitor area, which was black by default, turning completely white and then back to black again. Whilst a full screen flash, viewed straight on, was too large to be a stimulus to which collicular neurons would respond maximally, it was used to avoid any misleading (or, indeed, revealing) location information that might interfere with the task of finding the correct joystick position and the high change in luminance with each stimulus presentation ensured that it would be a stimulus to which the superior colliculus would be sensitive (Sparks, 1986).

This is particularly important when we consider the high sensitivity of the superior colliculus to stimulus location (Sparks, 1986; Wurtz & Albano, 1980). In practice, participants tended to adopt a similar attitude in this task to that adopted in the audio task and paid most attention to the movement of their hand. The screen flash was salient enough that participants did not feel the need to look directly at the computer monitor.

Results

Distance travelled during the post-discovery period

Once again, the data describing the distance travelled were the first to be investigated. A one-way repeated-measures ANOVA, with 6 levels, was conducted on the log-transformed data in order to investigate the effect of reinforcement delay. The analysis revealed that there was a significant effect of delay on the distance travelled during the post-discovery period, $F(5, 100) = 7.19$, $p < .001$. Figure 2.7 shows that the general effect of delay was to increase the duration of the post-discovery period, though the clearest effect was between the no delay and all other delay conditions. Bonferroni corrected post hoc *t*-tests revealed that the 0-ms condition differed significantly from all other conditions: 75-ms ($t(23) = 3.79$, $p < .05$); 150-ms ($t(23) = 3.34$, $p < .05$); 225-ms ($t(23) = 4.79$, $p < .05$); 300-ms ($t(23) = 5.14$, $p < .05$); 375-ms ($t(20) = 5.92$, $p < .05$). A further comparison between the 300-ms and the 375-ms conditions revealed no difference, though it approached significance ($t(20) = 2.22$, $p > .05$).

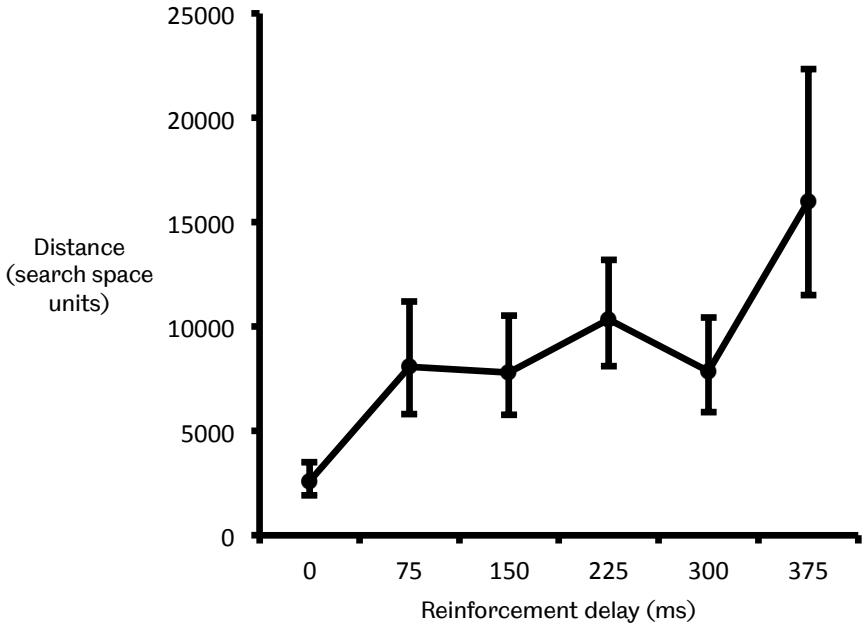

Figure 2.7 Mean distance travelled during the post-discovery period (and standard error) for the 6 levels of delayed visual reinforcement. Values are back-transformed from the log transformation.

Total number of hits recorded during the post-discovery period

Once again, the distance metric contained a considerable amount of variance, so an analysis of the hits metric was performed. A one-way repeated-measures ANOVA was conducted on the log-transformed data. The results show that there was a significant effect of reinforcement delay on the number of hits during the post-discovery period, $F(5, 100) = 14.84$, $p < .001$. Figure 2.8 shows the result from the audio task (experiment 1) plotted alongside the current visual results. Just as was the case with the audio task, the effect of the reinforcement delay was to increase the number of hits during the post-discovery period. Bonferroni corrected post hoc *t*-tests revealed that the 0-ms condition was significantly different to the 75-ms ($t(23) = 4.18$, $p < 0.05$), 150-ms ((23) = 4.56, $p < 0.05$) and the 225-ms ((23) = 6.14, $p < 0.05$) conditions. Further comparisons revealed that the 75-ms condition did not differ significantly from the 225-ms ($t(23) = 2.05$, $p > 0.05$) condition but did differ from both the 300-ms ($t(23) = 3.74$, $p < 0.05$) and 375-ms ($t(20) = 4.48$, $p < 0.05$) conditions. And finally, two further comparisons revealed that the 150-ms did not differ significantly from either the 300-ms ($t(23) = 2.78$, $p > 0.05$) or the 375-ms ($t(20) = 2.92$, $p > 0.05$) conditions. Thus, once again, the effect was not simply a delay versus no delay effect and the additional delays within the range tested had additional negative impacts on performance.

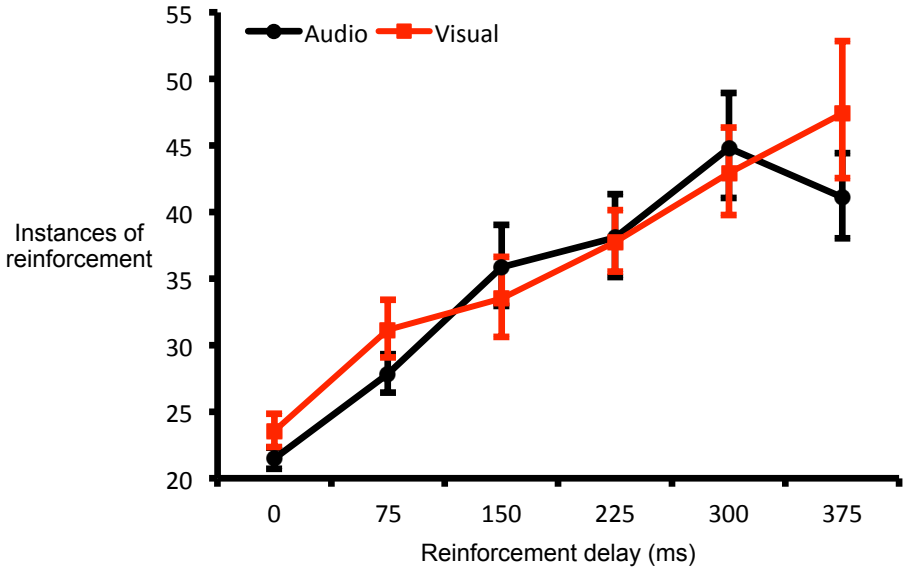

Figure 2.8 Mean number of hits during the post-discovery period (and standard error) for the 6 levels of reinforcement delay in experiments 1 (audio, shown in black) and 2 (visual, shown in red). Values are back-transformed from the log transformation.

Speed during the post-discovery period

Just as with experiment one, speed was calculated in terms of screen units per second for the post-discovery period. A one-way repeated-measures ANOVA was conducted. The results showed that there was no effect of reinforcement delay on speed during the post-discovery period, $F(5, 100) = 1.87$, $p = .12$ (means and standard error of the mean displayed in figure 2.9).

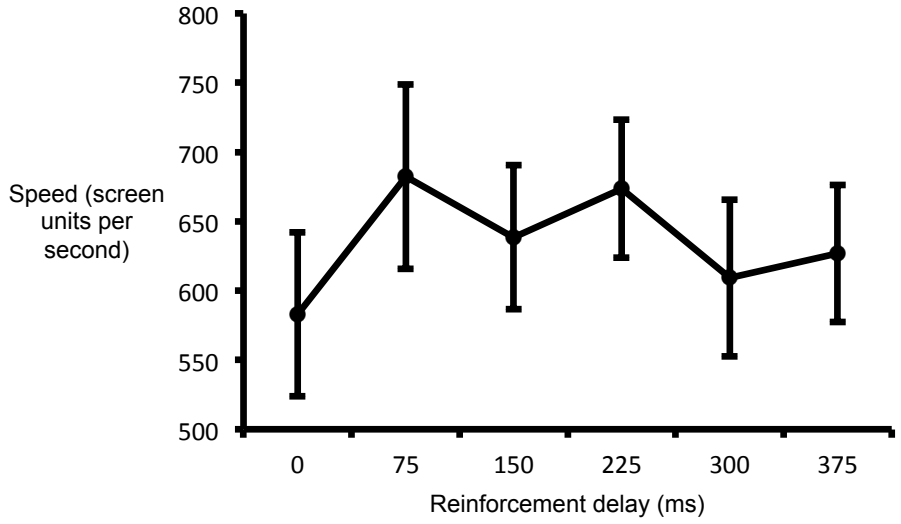

Figure 2.9 Mean speed during the post-discovery period (and standard error) for the 6 levels of delayed visual reinforcement.

Discussion

The results of experiment 2 were largely in agreement with the first experiment. Once again, there was an effect of delay at just 75 ms, which lends further support to Redgrave and Gurney's (2006) theory. The results also indicated that delay had an additional effect for durations in excess of 75 ms demonstrating that the effect was not simply a delay versus no delay contrast. There was no effect of delay on speed, once again, suggesting that participants (none of whom were able to guess the independent variable) were not using strategies to adjust their behaviour in order to cope with delay and therefore providing some assurance that the task was capable of measuring low-level, nondeclarative learning effects. Most importantly, the results showed that the sensitivity to delay at durations close to the latency of the phasic response of dopamine neurons held for visual stimuli (large changes in luminance) to which the superior colliculus would have been sensitive (Sparks, 1986). This, therefore, provides support for the specific account of an action acquisition mechanism that relies on the very short latency input from the superior colliculus.

General Discussion

There were three main reasons for conducting the experiments described in this chapter. Firstly, the task described was the first version of the joystick task and these

experiments allow for an initial assessment of it as a new behavioural paradigm for the investigation of action acquisition. Secondly, the experiments provided a means of investigating sensitivity to delay in a task that relies on feedback delivered in a fashion consistent with reinforcement learning and yet demanding of responses similar to the kind normally found in adaptation learning or supervised learning tasks. And thirdly, the experiments allowed for a test of an effect of delay at a duration of less than 100 ms as a way of investigating the idea that the phasic activity of dopamine neurons is short latency as an evolved means of reducing contamination when associating outcomes with the movements that caused them. These three issues are discussed below in reverse order.

High delay sensitivity and time-stamping mechanisms

It's easy to appreciate how a neural mechanism designed to stamp-in recent motor output, following a novel stimulus, might have evolved to function at extremely low latencies. The advantage of ignoring causally irrelevant motor output by promptly time-stamping behaviour that led up to a stimulus is conceivably so great that it outweighs any gains that might come from waiting for sensory input that has undergone a greater degree of processing. Consequently, the argument from Redgrave and Gurney (2006) and Redgrave et al. (2008), that short latency collicular input prompts short latency activity of dopamine neurons as part of a system involved in action acquisition, is a compelling one: it might be better to learn first and ask questions later.

Whilst we can't directly increase the latency of dopamine neurons in human subjects, it is possible to introduce a delay between an action and the subsequent stimulus. Even though the stimulus would not be contingent on the activity that took place during the delay period, there would be no way for an indiscriminate low-level learning mechanism to take this into account on its own. We might, therefore, assume that at least some of the irrelevant behaviour from the delay period will be stamped-in. So, if we introduce delay into a task that relies heavily enough on this kind of low-level learning mechanism, then we should expect it to interfere with learning even at very short durations. In both of the experiments detailed in this chapter, an effect was

found at delays in the region of 75 to 100 ms in duration. Although this doesn't prove that a particular brain region was involved in the effect, it shows that there is a learning mechanism, of some description, that is so sensitive to the timing of movements and sensory outcomes that it would benefit from short neural latencies of the kind discussed by Redgrave and Gurney (2006). Certainly, the effect found here offers more convincing support for this theory than the comparatively long duration delay effects found in previous experiments involving reinforcement learning cited by Redgrave and Gurney (e.g. Elsner & Hommel, 2004; Hollerman & Schultz, 1998). However, this in itself raises questions as to why the sensitivity in the current experiment should be so much greater than in previous studies of reinforcement learning.

Differing sensitivities to delay depending on the type of motor task

We hardly need to run an experiment to appreciate that the impact of delay on learning and performance is likely to be a negative one in most situations. However, predicting the relative impact of different delay durations depending on the type of learning or activity being undertaken is rather less straightforward. An apparent split in the overall sensitivity to delay seen in motor control and motor adaptation paradigms versus reinforcement learning paradigms is difficult to interpret due to major differences in the response mechanisms involved. Generally speaking, the former rely on tasks that demand complex or highly accurate responses such as precise reaching movements or target tracking. The latter, by contrast, tend to rely on simple discrete responses that place no emphasis on accuracy, such as button pressing. Consequently, we can't come to any conclusions as to whether the apparent difference in the effect of delay is due to the different learning mechanisms being employed or whether it is due to the response requirements of the tasks.

By employing a rich input device instead of button-press or lever-press responses, the task described in this chapter minimised one of the major differences between the experimental approaches. The results show that, in terms of overall sensitivity to delay, the task described in this chapter was more akin to tasks employed in motor

control (Foulkes & Miall, 2000; Miall & Jackson, 2006; Miall et al., 1985) or motor adaptation paradigms (Held et al., 1966; Kitazawa et al., 1995) than reinforcement learning paradigms (Elsner & Hommel, 2004; Shanks & Dickinson, 1991; Shanks et al., 1989), and yet it was designed to be a test of reinforcement learning, with stimulus characteristics that one would expect to find in such a task. For instance, all reinforcement signals were qualitative, indicating whether or not the correct movement had been produced but providing no information on the relative level of performance: the output only ever indicated whether or not the joystick had been moved into the correct position. Reinforcement was also contingent on behaviour: if an individual failed to move the joystick into the hotspot, they would have received no feedback at all. The question is: what can such a finding tell us about the distinction between these two kinds of learning?

There is much to be said for drawing a theoretical distinction between supervised learning and reinforcement learning (Jordan & Rumelhart, 1992; Wolpert, Ghahramani & Flanagan, 2001). However, in developing a task that involves rich movements with feedback of the kind found in reinforcement learning, there is a sense in which the boundaries between the two types of learning have been methodologically blurred in the current task. From a participant's perspective, the within-trial repetition of reinforcement found in these experiments demanded the kind of online readjustment of performance that we would not typically associate with reinforcement learning and yet the feedback provided retained characteristics associated with reinforcement learning. This version of the task, therefore, raises the question of whether the distinction between supervised and reinforcement learning carries much meaning at the behavioural level. It is possible that the distinctions we draw between different types of learning and their associated forms of feedback do not correspond to the way in which the brain deals with this information. Certainly, the current study suggests that, at the behavioural level, sensitivity to delay has more to do with whether or not reinforcement is delivered rapidly as an action is being produced as opposed to the informational properties of the reinforcement.

There is evidence that motor control with delayed visual feedback can improve with considerable practice, though performance always remains far below that achieved

with immediate feedback (Miall & Jackson, 2006; Sheridan, 1993). Such findings are primarily from supervised learning tasks, where the current position and the target are always visible to the agent. As a means of further understanding the type of learning that is being tested in the current task and the more general differences between supervised learning and reinforcement learning, it would be informative to measure the ability of people to improve their performance with delayed reinforcement over many trials and many experimental sessions.

Audio versus visual reinforcement

One possibility for the similarity in the effects of audio and visual stimuli is that the experiment simply wasn't calibrated to detect any such differences. Previous psychophysics research (Jaśkowski, Jaroszyk & Hojan-Jezierska, 1990), has found that reactions times to audio stimuli are approximately 40 ms faster than those to equivalent visual stimuli. On this basis, one might have expected to find a difference in the joystick task, especially in light of the large number of stimulus presentations that occurred in each trial. However, when audio and visual stimuli are well above threshold, as was the case here, there appear to be no such differences in reaction times (Kohfeld, 1971). Consequently, it is unlikely that different modalities could have produced a different effect unless the role they played in the learning mechanism being tested was substantially different.

Assessment of the task

The intention when developing the task was to measure learning based on discrete instances of qualitative reinforcement. This type of feedback could only provide information as to whether the correct action had been performed so it was a desirable means of investigating a learning mechanism based on the reinforcement of recent motor output. The learning criterion was introduced as a means of determining the point at which the correct position of the joystick had been sufficiently well learned. The combination of these two features made it possible to present participants with multiple new learning scenarios (in the form of new hotspot locations) and to extract a performance metric in the form of hits that was resistant to some sources of variance that are a problem for open ended trials.

However, efficient though the structure of the task was, it was nonetheless limited from the perspective of making use of the rich data that was collected. In essence, the action being learnt in this task was to hold the joystick steady in a particular position. Whilst the end position is well defined in this task, there is nothing to constrain the route that is taken towards the hotspot following each instance of reinforcement. In other words, the end point might be clearly spatially defined, but movement towards that point is bound to vary between instances of reinforcement as the participants overshoot the target and attempt to return. Figure 2.10 illustrates this point. We can see that the post discovery period is typified by a clustering of movements within a particular portion of the search space rather than by a particular shape of movement trace.

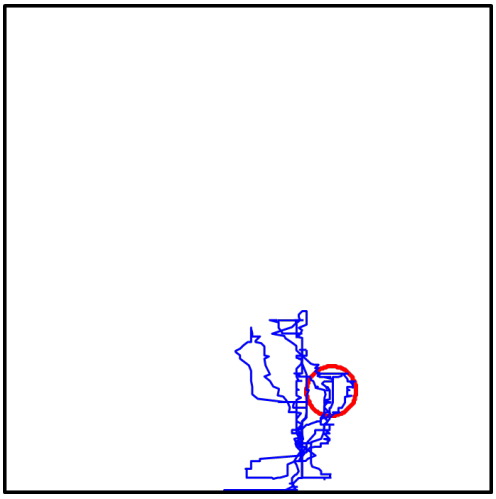

Figure 2.10 A trace of movements during the post-discovery period only. The task demands are such that they don't encourage particular shapes of movement. (The black frame depicts the boundaries of the search space.)

A further problem stemming from the task structure was that participants were never required to repeat a particular sequence of movements in its entirety. In Thorndike's (1911) puzzle-box paradigm, animals were required to perform a task until they happened upon the solution, at which point they were placed back into the puzzle-box and required to produce the whole sequence of movements again. If we could draw a movement trace of this activity, we would see that the trace would take on a clearer shape as trials progressed and movement became more efficient. In the current

version of the joystick task, however, it would have been difficult to learn the correct hand position through the reinforcement of particular muscle movements; instead, the consistent aspect of behaviour was the particular resting place of the hand, a higher level motor command. Whilst this level of representation is something that the nervous system is more than capable of encoding (Graziano, 2006; Graziano, Taylor & Moore, 2002), from an experimental point of view it makes the progression of learning much harder to decompose at the behavioural level. A potentially better situation would be to split learning into discrete trials (i.e. between-trial as opposed to within-trial repetition of reinforcement) and this will be the focus of the next chapter.

Chapter 3: contamination, delay and between-trial repetition of reinforcement

This chapter focuses on a change to the structure of the joystick task, whereby reinforcement is repeated between trials rather than within trials. Experiment 3 details the development of this task and its structure whilst also investigating the possibility that target size might be utilised as a way of scaling task difficulty for potential future investigations comparing populations of differing abilities. Experiment 4 details a study designed to investigate the effects of delay with this new task structure. The purpose of this was twofold. Firstly, it provides a measure of the sensitivity to delay of the joystick task where the delivery of reinforcement is more comparable to that used in other areas of research. Secondly, it allows the opportunity to more clearly investigate the activity that occurs during the delay period in an attempt to investigate hypotheses regarding the contamination of the motor record during action acquisition.

Reasons for employing between-trial repetition of reinforcement

Response acquisition in non-human animals is a subject that has received much research attention in fields associated with operant conditioning and reinforcement learning. At least some of this interest has to do with the practicalities of training animals to produce responses in the laboratory that they might not be naturally disposed to produce (Peterson, 2004). Many species of animal, faced with the need to produce a novel response, are not able to infer that response from the mechanism that confronts them, nor are they able to copy the behaviour from other animals or a human instructor (Thorndike, 1911). They are, therefore, forced to acquire responses through trial and error, extracting the important motor elements from their own behavioural variance. It is perhaps unsurprising, then, that a similar interest has not been taken in response acquisition in humans. The ability to acquire responses can, by and large, be taken for granted in humans: a person will either come to a new task with the response already in their behavioural repertoire or they will be able to produce the response immediately after having observed another person performing it. The joystick task is aimed specifically at recording response acquisition in humans by

removing access to a perfect model of performance and increasing the extent to which people must extract a response from their own behavioural variance.

Chapter 2 detailed a version of the joystick task that measured the ability of people to acquire novel responses based on discrete instances of qualitative feedback, but the structure of the task differed from the type of response acquisition that is typically required of laboratory animals (e.g. Lattal & Gleeson, 1990). During the acquisition of a lever-press response by a rat, for example, a stable set of movements, necessary for depression of the lever, must be produced before reinforcement can be delivered. This set of movements does not change from one lever-depression to the next and therefore it is this common element within the animal's behaviour that is learnt. The joystick task featuring within-trial repetition of reinforcement, described in chapter 2, required the learning of a stable hand position rather than a stable sequence of movements and therefore differed from the typical structure employed to investigate action acquisition in laboratory animals. In order that the joystick task can be considered on similar terms to such studies, it was decided that a version of the task featuring between-trial repetition of reinforcement should be developed.

By iterating the learning process between trials, with a stable starting position, we can ensure that the data for each trial represents a single attempt at finding the correct movement. Figure 3.1 shows what this looks like in practice. Trial one is necessarily a naïve trial and locating the correct position is a question of exploring the search space. The hotspot can then be kept in the same position for subsequent trials, allowing us to record performance following an individual instance of reinforcement. It is much easier to appreciate with this design how people have deviated from optimal performance on a given trial.

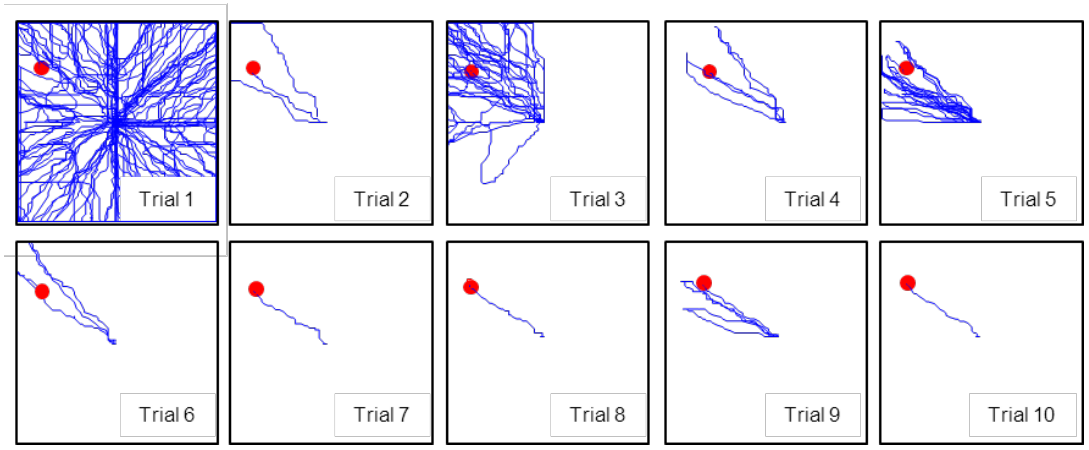

Figure 3.1 The progression of a movement trace over 10 trials with reinforcement occurring just once in a trial, on entry into the hotspot region of the search space (shown here in red).

A further advantage of adopting this task structure is that it should make it easier to draw comparisons concerning the effects of delayed reinforcement with the studies cited in the previous chapter (e.g. Anderson & Elcoro, 2007; Black et al., 1985; Elsner & Hommel, 2004; Hollerman & Schultz, 1998; Lattal & Gleeson, 1990; Okouchi, 2009; Renner, 1964; Shanks & Dickinson, 1991; Shanks et al., 1989; Sizemore & Lattal, 1977; Snycerski et al. 2004; Snycerski et al., 2005; Stubbs, 1969). It is possible that the high degree of sensitivity to delay found when employing within-trial repetition of reinforcement in experiments 1 and 2 would not occur with a task structure based on between-trial repetition of reinforcement. In other words, this change in task structure might bring the sensitivity to delay into the 0.5 to 1 s range, a duration that marks the maximal sensitivity of the experiments cited above.

Scaling task difficulty

Before exploring the effects of delay, the new task structure was used to investigate the potential scalability in difficulty of the joystick task. Such manipulations are potentially of benefit when comparing across populations, for instance between clinical and control populations. Clearly, the difficulty of all experimental tasks can be varied to some extent but the advantage here is that the required precision of movement can potentially be calibrated to the abilities of a particular population whilst leaving the basic task rules (the contingencies on which reinforcement depends) stable. Such calibration could potentially provide a means of running more successful control groups when comparing healthy and clinical populations. This type of

manipulation is not possible in simple reaction time or lever pressing paradigms because in these instances the response itself never varies in terms of the difficulty of exploration. Experiment 3, therefore, details the structure of a joystick task featuring between-trial repetition of reinforcement and describes a test designed to investigate whether the joystick task paradigm would allow the effective scaling of task difficulty for such purposes.

Experiment 3: scaling task difficulty with variable target sizes

With a task based on location finding, the obvious candidate parameter for scaling difficulty is to change the size of that location. With larger targets it is easier to produce a sequence of movements that will take the joystick into the hotspot region of the search space and therefore learning and performance should improve with larger hotspot sizes. This experiment was designed to investigate the effect of hotspot size on performance and determine whether the change in task difficulty with hotspot size is a viable method for calibrating the task based on the abilities of a particular individual or population. A further benefit of running a formal test to investigate the effect of hotspot size is that it enables better informed choices of suitable hotspot sizes for future experiments including experiment 4 described later in this chapter.

Method

Unless otherwise stated, all apparatus and procedures were the same as those used in experiment 1 (see chapter 2).

Participants

29 people (26 female) participated in all conditions of this study. Ages ranged from 18 to 26 years with a mean age of 19 (*SD*. = 1.8 years). Participants were all undergraduate students at the University of Sheffield who took part in return for credits in the department's research participation scheme. All subjects were naive to the purpose of the experiment.

Apparatus

Defining the reinforced area

It was easy to determine from pilot studies that hotspots of the size used in experiment 1 (i.e. occupying 0.91% of the search space) would have made a task featuring between-trial repetition of reinforcement much too easy. Consequently, the largest hotspot in the current experiment was chosen to be substantially smaller than that used in experiments 1 and 2, in an attempt to cover as informative a range of hotspot sizes as possible. Ultimately, 4 different sizes were chosen. In order from smallest to largest, they occupied 0.07% (small), 0.14% (medium), 0.28% (large) and 0.56% (exlarge) of the search space, thus the figure doubled with each increase in size. Figure 3.2 is drawn to scale and gives a visual representation of the size of the hotspots relative to the overall search space.

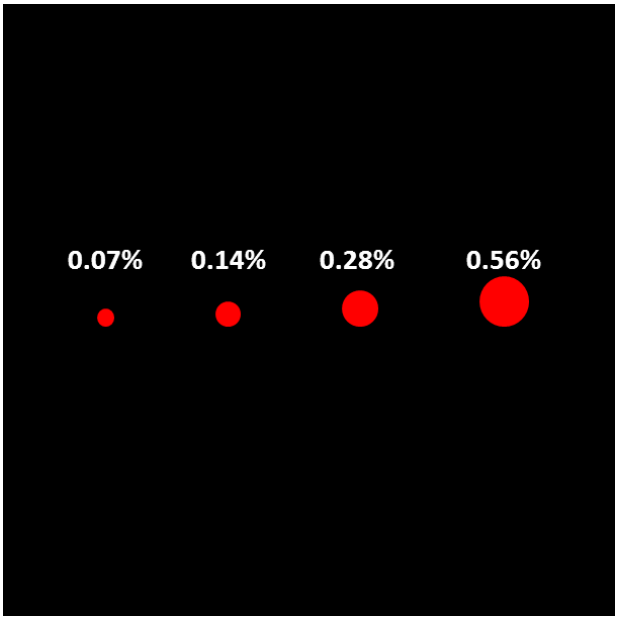

Figure 3.2 Hotspot sizes as a percentage of the overall search space.

Defining reinforcement

Just as in experiment 1, the screen was kept black throughout a trial and any movement into the hotspot resulted in a brief audio stimulus (10 ms 'pip' sound) but, unlike experiment 1, it also resulted in the immediate ending of that trial accompanied by white text on the black screen (see appendix 3 for full details of onscreen text).

Procedure

Participants were given verbal instructions as explained in experiment 1 (see chapter 2). Following this, the task program was started and the participants were asked to follow the onscreen instructions (see appendix 3). After reading the instructions, 3 practice trials commenced automatically. For the first practice trial, the participant was required to move the joystick around in search of the correct position and on finding that position was presented with a pip sound and the end of the trial. Following this they would then complete the second and then the third practice trials, with the hotspot remaining in the same position for all 3 practice trials. As with all the experiment trials, no feedback or screen graphics were provided during the practice trials and the monitor set to display a black screen. The practice trials differed from experimental trials in that the hotspot was larger than any of the experimental hotspots, occupying 0.75% of the search space. Once the practice trials were completed the experimental trials began. Participants were required to complete 10 trials at each hotspot size. For the first trial in a batch of 10 the hotspot was placed randomly within a limited region of the search space (as defined in experiment one), such that no part of it could overlap either the centre or the outer edges of the search space. During each batch of 10 trials, the size and position of the hotspot did not change.

Design

There was one independent variable with 4 levels: the 4 hotspot sizes detailed above. There were 43 trials in total: 3 practice trials and 40 experimental trials. Each experimental condition was therefore experienced in the form of 10 identical trials (i.e. same hotspot position and size). The order of presentation of the experimental conditions was counterbalanced.

Results

Irrelevant distance as a performance metric

As a result of the way that hotspot positions were determined, each new hotspot location could differ in terms of its distance from the centre of the screen. In other

words, the optimum distance between the starting position and the hotspot varied between trials. The reason for this was to keep the required action as variable as possible within the confines of the two-dimensional search space. However, it meant that 2 people achieving optimum performance on a given trial could differ in terms of the distance travelled. One potential alternative to using the raw distance is to take the ratio between the total distance and the optimum distance. This is intuitively appealing because it would seem to take into account the fact that for longer optimum distances there is more opportunity to make irrelevant movements, simply because more distance must be travelled in order to reach the hotspot.

However, the chief problem with ratio scores is that they have a disproportionately large effect on trials that involve even moderate amounts of movement. Whilst the optimum distance may only represent a small proportion of the overall distance travelled, it can greatly affect the performance score because it is the denominator in the fraction. This effect is easier to appreciate if we look at movement traces that display similar overall amounts of movement with different optimal distances. The actual distance travelled for the trace on the left hand side of figure 3.3 is 14 times the optimal distance whereas this ratio is just 6.5 for the trace on the right of the figure and yet it seems strange to suggest that performance is twice as efficient in the right hand trace.

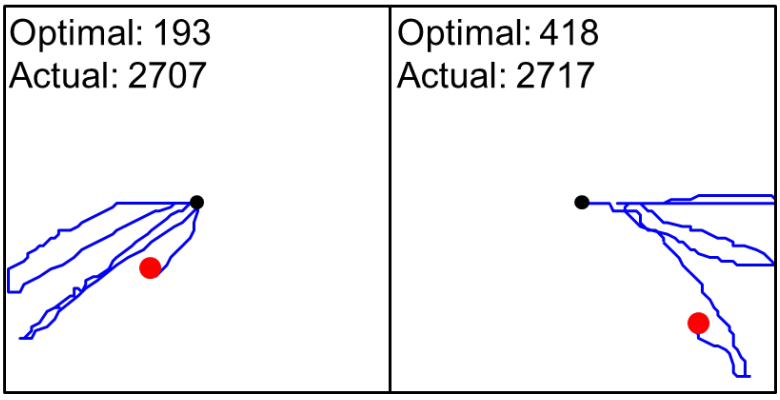

Figure 3.3 Examples of movement traces of similar length in trials with very different optimal distances.

In order to investigate this issue further, a test for a correlation between the optimum distance and the 'irrelevant distance' (total distance minus optimum distance) was run for each of the 4 hotspot sizes. If the optimum distance were having an important effect on the actual distance travelled then we would expect these two metrics to

correlate with one another. A Pearson product-moment correlation was run to test the relationship between the optimum and log-transformed irrelevant distances. This revealed that there was a significant (2-tailed) positive correlation for the small (r = 0.5, n = 29, p < .01) and medium (r = 0.52, n = 29, p < .01) conditions but no correlation for the large (r = 0.23, n = 29, p = .23) and the exlarge condition (r = 0.3, n = 29, p = .12). Longer optimal distances were therefore correlated with longer irrelevant distances for 2 of the 4 conditions.

Because the correlation differed across conditions, it was decided that optimum distance would not have made a consistent covariate and that the effect of calculating a ratio score might have interacted with the independent variable. It was therefore decided that irrelevant distance (total distance minus optimal distance) would be used as a conservative measure of performance. This metric avoids the issue of potentially overestimating the effect of the optimal distance whilst also taking into account the basic additional distance that some trials required over others by virtue of the hotspot being located further from the central starting position.

Scalable task difficulty and the effect of hotspot size

For each participant the mean irrelevant distance was calculated based on the last 9 trials of the 10 trial batches; the first trial was always excluded from the calculation as it represented naïve performance, indicative of search time rather than learning. A one-way repeated-measures ANOVA showed that there was a significant effect of hotspot size (F(3, 84) = 13.29, p < .001). Figure 3.4 shows that as hotspot size increases so does the mean irrelevant distance travelled; in other words, as one would expect, the task gets more difficult with small hotspots. Bonferroni corrected post hoc t-tests revealed that the mean distance travelled in the small condition differed significantly to both the large (t(28) = 4.6, p < .05) and the exlarge (t(28) = 4.94, p < .05) conditions and that the medium condition differed significantly from the exlarge condition (t(28) = 4.18, p < .05). No other comparisons reached significance, although the small-medium comparison approached significance (t(28) = 2.53, p > .05).

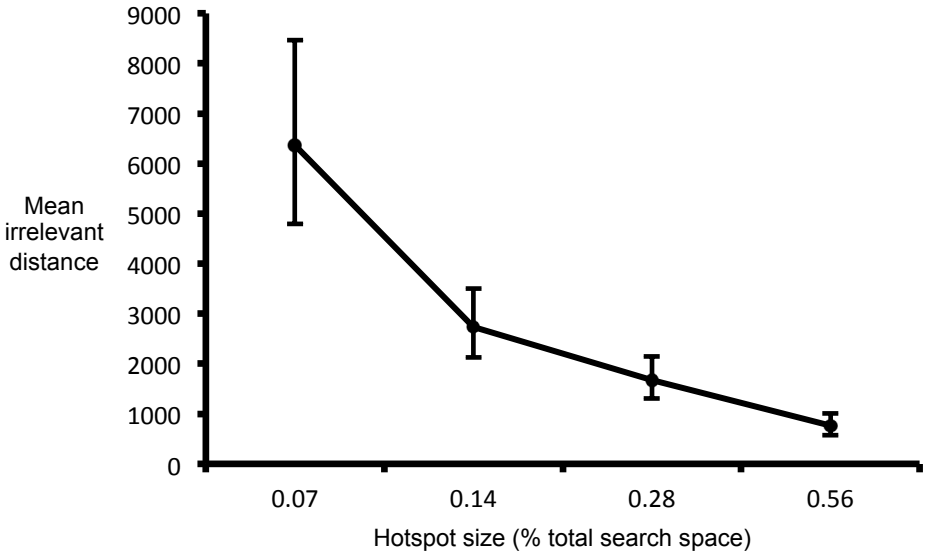

Figure 3.4 Mean irrelevant distance travelled for the 4 different hotspot sizes: small (0.07%), medium (0.14%), large (0.28%) and exlarge (0.56%). Values are back-transformed from the log transformation.

Because of the iterated structure of the task it is possible to plot the mean performance on each trial for the four different hotspot sizes. Figure 3.5 shows how the mean performance across participants improves with experience at a given hotspot size and location (the naïve trial 1 has been removed to make the comparison clearer). Figure 3.6 has the error bars removed and includes trial 1 to show where post-reinforcement performance lies relative to naïve searching. Overall learning is rapid for each of the 4 hotspot sizes, with the majority of the improvement occurring by trial 4.

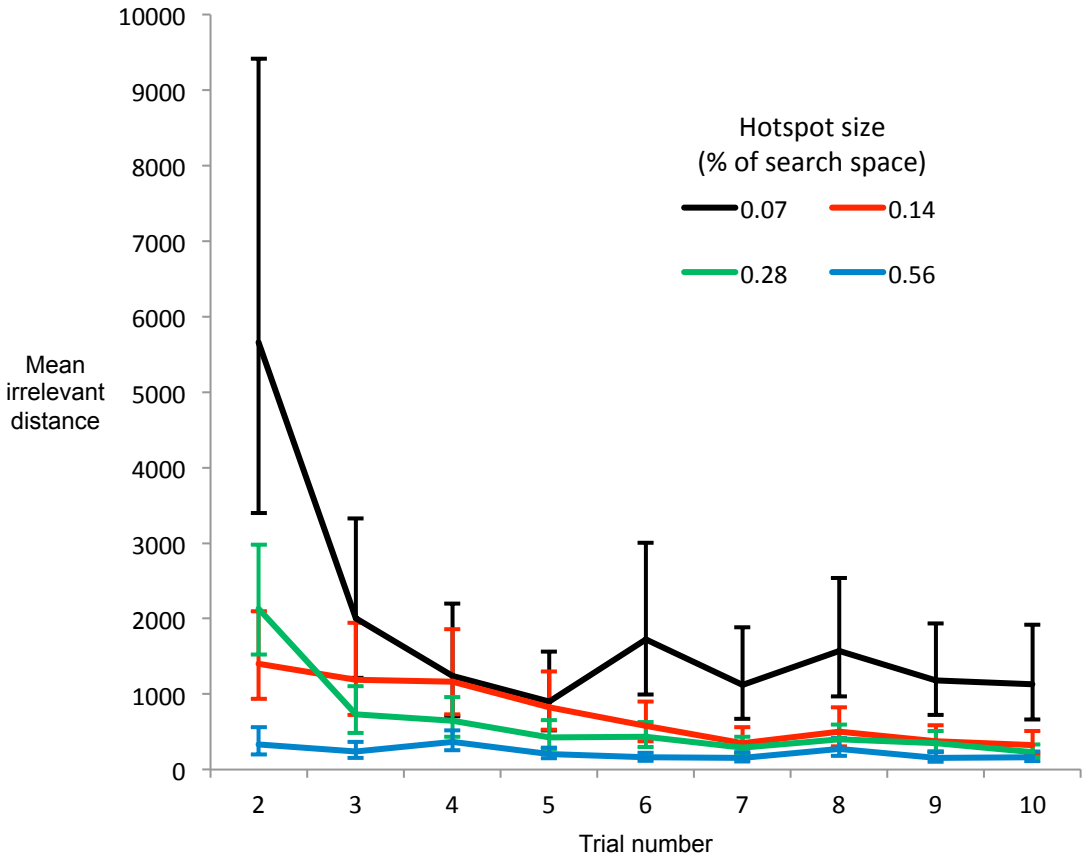

Figure 3.5 Mean irrelevant distance (and standard error) for each trial at the four different hotspot sizes. Trial 1 (naïve trial) is removed in order to display rate of learning more clearly. Values are back-transformed from the log transformation.

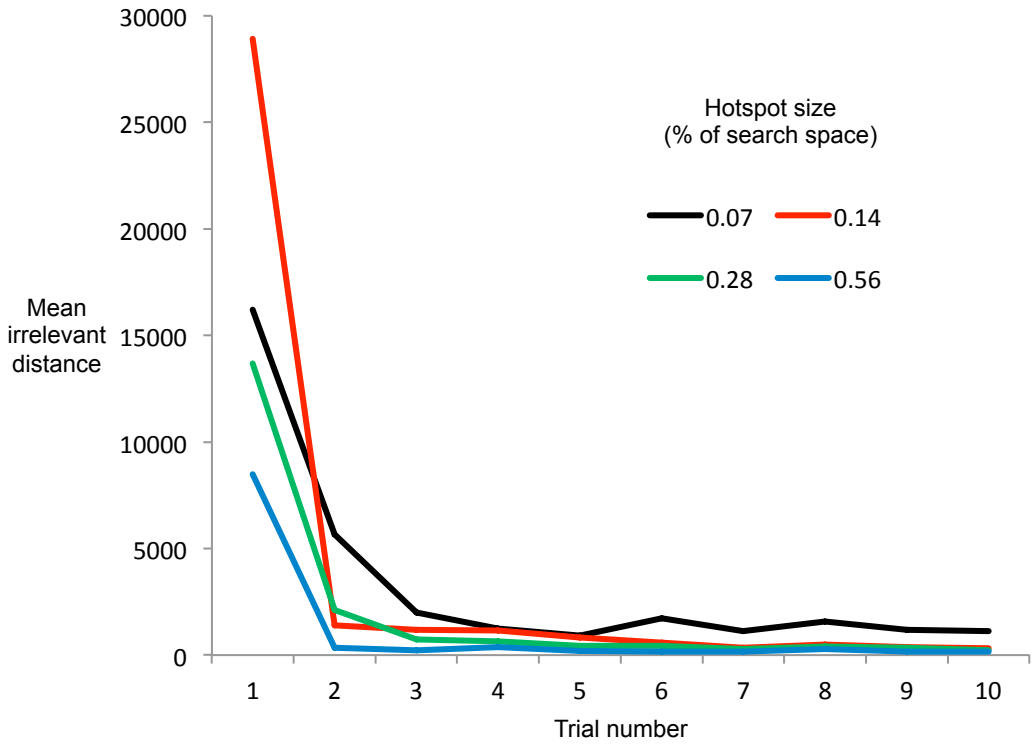

Figure 3.6 Mean irrelevant distance for each trial at the four different hotspot sizes. Trial 1 is included to show where post reinforcement performance lies relative to naïve searching. Error bars have been removed so that the approximate rate can be seen more clearly (see figure 3.5 for error bars on trials 1-9). Values are back-transformed from the log transformation.

Discussion

The results of the overall comparison in performance (figure 3.4) indicate that task difficulty does significantly differ depending on hotspot size; however, the scope for calibrating the joystick task based on manipulations to this experimental parameter is likely to be limited. In terms of overall performance and rate of learning, the smallest hotspot size (0.07%) appears to mark a substantial shift in task difficulty relative to the other sizes. Despite the considerable variance in the performance scores, the difference to the other conditions was either significant or approached significance with a conservative post hoc correction. Furthermore, figure 3.5 indicates that this difference appears to be sustained across trials. In one sense, this is potentially useful as it indicates that with hotspots of this size we can be assured of avoiding ceiling effects. However, on playing back trials from this condition, there were signs that some of the variance and difficulty might have been caused by limitations in the apparatus. Although the tracking of the joystick was generally smooth, the smallest

hotspot occupied so little of the search space that irregularities in the movement of the joystick would have made it more difficult to locate the hotspots. In other words, it is more likely in the small condition that a participant could produce a movement that, under perfect conditions, would have entered the hotspot but due to the noise in the apparatus would actually result in a false-negative outcome. Furthermore, it is possible that this hotspot was presenting substantial challenges to the motor control of participants. Whilst it is potentially useful to make the task harder for some populations of individuals, the difficulty should depend on exploration of the search space rather than the motor control of the participant. At the other end of the scale, the largest hotspot (0.56%) produced the opposite problem, resulting in near ceiling levels of performance. Consequently, the scope for varying task difficulty by manipulating hotspot size appears to be very limited with this paradigm.

Experiment 4: between-trial repetition of delayed reinforcement

Temporal alignment and the eligibility period

In chapter 1, the idea of a short latency and relatively indiscriminate learning system was introduced and in chapter 2 the temporal sensitivity of such a system was discussed. It was also suggested in chapter 2 that the findings cited by Redgrave and Gurney (2006) as possibly indicating the negative effects of delay on such a learning system did not perhaps demonstrate the degree of sensitivity that their theory implied. However, it would be wrong to think that the delay durations employed in these studies were especially long in the general context of research into reinforcement learning. In fact, the range of delays over which reinforcement learning can apparently be sustained is surprisingly large. Several studies, for instance, have found that animals are capable of acquiring new responses, such as lever-press responses, with reinforcement delays in excess of 10 s (Lattal & Gleeson, 1990; Snycerski et al., 2004; Snycerski et al., 2005). A similar result has been reported for acquisition of response sequences by humans with delays also in excess of 10 s (Okouchi, 2009).

If it is the case that these subjects were relying on reinforcement learning and that this system is indiscriminate with respect to everything but the temporal alignment of action, outcome and context, then the apparent resistance to delay is somewhat surprising. As discussed in earlier chapters, Redgrave and Gurney (2006) suggest that the short latency phasic response of dopamine neurons to salient stimuli acts as a time stamp that reinforces and promotes the reselection of motor output that immediately preceded a novel stimulus and that the short latency nature of this response limits the quantity of non-contingent motor output that could be associated with the stimulus. Whilst the short latency of the dopamine response suggests that this learning system is likely to be sensitive to the effects of delay (as discussed in chapter 2), the limits of this system's tolerance to delay are unclear. According to Redgrave and Gurney, the converging input of sensory, motor and contextual information in the striatum is the target with which the phasic dopaminergic output interacts. Consequently, the limits of the duration over which learning in such a system could occur should ultimately be dictated by the length of time that the convergent signals in the striatum remain eligible for reinforcement.

This concept of an eligibility trace has proven useful in models of reinforcement learning for explaining how it is that a reinforcement learning mechanism might cope with delays of reinforcement (Singh & Sutton, 1996). If reinforcement occurs immediately, it will interact with a trace that contains a strong representation of the behaviours on which that reinforcement was contingent; however, with increasing delay between the contingent response and the reinforcement, the representation of that response weakens (decays) and the overall effect on learning is that less weight will be placed on these older responses and thus more repetitions would be required in order to learn under conditions of long delay duration. An eligibility period caused by the gradual decay of behavioural representations provides an explanation as to how it is that reinforcement learning could occur with delays in excess of 10 s, provided there is sufficient opportunity to repeat the behaviour many times. It would also provide an explanation as to how it is that humans and animals are able to learn sequences of actions that take place over many seconds because only the final portion of these sequences would be temporally contiguous with the outcome despite the fact that the whole sequence is what the outcome is contingent on.

However, one alternative explanation is that the eligibility period concerns not so much the length of time that has elapsed but the number of task-relevant movements that have occurred during the delay period between a contingent response and a given outcome. This is conceptually similar to the idea of working memory (Baddeley, 2003), with the learning mechanism being limited not only by time but also by the number of units of relevant information present in the trace. The more output that occurs during the delay period, the more potential there is for this non-contingent output to be assigned the credit of having caused the reinforcing event. By extension, the more similar the contingent and non-contingent output components of the motor record are to one another, the harder it might be to pick out the signal from the noise. If this is correct, then the sensitivity that was found in the version of the joystick task detailed in chapter 2 might be a consequence of the low contrast between task relevant and task irrelevant motor record that occurred during the delay period. In other words, the difference between relevant and irrelevant joystick movements is small as compared to the difference between lever-presses and cage exploration behaviour (the response opportunities in the extended delay studies cited above) and therefore non-contingent behaviour in the joystick task might have a more contaminating effect during learning, rendering the task more susceptible to the effects of delay.

The employment of within-trial repetition of reinforcement in the experiments described in chapter 2 made it difficult to explore the idea of contamination. In theory, learning should be less efficient if more movements occur during the delay period because it is then harder to identify the portion of movement activity on which the outcome was dependent; that is to say, the credit assignment problem becomes more difficult (Redgrave et al., 2008). However, measuring this activity was not straightforward because each trial contained many instances of delayed reinforcement and therefore many separate instances of contaminating movement activity. A solution to this is to have a single instance of reinforcement in each trial as outlined above. In this way, for any given trial, it would be straightforward to separate the task relevant portion of the movement trace versus the movements that occurred during the delay period. Consequently the contents of the delay period can be investigated, thus providing insight into how contamination might work to impede performance

under conditions of delay.

The problem with addressing this research question experimentally is that it would involve the manipulation of the amount of activity that occurs during the delay period. This could potentially be achieved by intervening in ways that make participants move the joystick faster or in more complex ways. However, with this type of manipulation, there is no obvious control condition that could tell us whether any effect was due to contamination per se or whether the intervention had simply made the task more difficult. Consequently, one of the aims of experiment 4 was to use the natural variation in the movements during the delay period in order to investigate whether contamination could be contributing to any drop in performance caused by reinforcement delay. Specifically, if contamination is important, the complexity and overall distance of movement during the delay period should be negatively correlated with performance.

The second aim of experiment 4 was to determine the sensitivity to reinforcement delay of a version of the joystick task employing between-trial repetition of reinforcement. The structure of the task is much more similar to that of traditional reinforcement learning and response acquisition paradigms. Once again, if the task relies on the learning mechanism described by Redgrave and Gurney (2006) then we would expect an effect of delay at the kind of short durations found in experiments 1 and 2. By contrast, if the effect of delay is primarily affected by specific task demands then we would expect that the sensitivity to delay should lower and much more in line with previous experiments featuring delayed reinforcement.

Method

Unless otherwise stated, all apparatus and procedures were the same as those used in experiment 3.

Participants

30 people (25 female) participated in all conditions of this study. Ages ranged from 18 to 22 years with a mean age of 19 (*SD*. = 1 year). Participants were all undergraduate

students at the University of Sheffield who took part in return for credits in the department's research participation scheme. All subjects were naive to the purpose of the experiment.

Apparatus

Hotspot size

The hotspot size was chosen based on the results from experiment 3. The 0.28% hotspot size employed in experiment three was selected in order to avoid the possible ceiling effects associated with the exlarge (0.56%) condition and the large amounts of error present in the small (0.07%) condition. The 0.28% (large) size was selected in preference to the 0.14% (medium) size on the basis that the introduction of delay would increase the difficulty of the task and it was important that the delay conditions shouldn't be overly difficult for the purposes of data collection. Furthermore, the results from the correlations in experiment 3 indicated that there was no relationship between optimum and irrelevant distance for this hotspot size so by choosing this size it was hoped that noise from this potential source of variance would be reduced.

Delay

Just as with experiments 1 and 2, delayed reinforcement was achieved by interposing a delay between the point at which the joystick moved into the hotspot and the point at which reinforcement was delivered. However, the structure of the task was the same as that employed in experiment 3 so the reinforcement stimulus (short duration pip sound) also signalled the end of the current trial. The range of durations chosen was: 0-ms, 150-ms, 300-ms and 450-ms. These durations therefore covered a slightly larger range than those in experiments one and two but with an increased increment between conditions. This decision was based on the fact that paired comparisons for delays in excess of 75 ms revealed no effects with a single increment in delay duration in experiments 1 and 2; furthermore, pilot tests suggested that the task was much less sensitive to the effects of delay so increasing the range with fewer durations would provide more informative data. Finally, it was important that at least one of the conditions was close to the 0.5 s delay, which was the shortest effective delay duration reported in the experiments cited earlier (e.g. Hollerman & Schultz, 1998).

Procedure

The procedure was identical to that employed in experiment three. However, participants were asked an additional question at the end of the experiment

Design

A repeated measures design was used. The independent variable was reinforcement delay, with 4 conditions: 0-ms, 150-ms, 300-ms and 450-ms. Each experimental session was comprised of 83 trials, the first 3 of which were practice trials (involving no delay) and the remaining 80 were experimental trials. The trials were presented as 8 batches of 10, where each batch represented a new hotspot location. Each delay condition was experienced twice. In other words, participants experienced 2 batches of 10 trials at each delay condition. Batch order was counterbalanced.

Results

Defining pre-discovery and post-discovery periods

Figure 3.7 displays movement traces that depict all of the movements made during a 10 trial batch at 450-ms delay. We can see that the movement trace doesn't stop at the hotspot, but carries on as a short tail representing the movements made during the delay period. It was therefore possible, for all trials in which reinforcement had been delayed, to distinguish between the pre-discovery and post-discovery periods. Just as with experiments 1 and 2, the pre-discovery period included all movements made from the start of the trial to the point at which the joystick moved into the hotspot. Figure 3.8 depicts this clearly by displaying only those movements that occurred during the pre-discovery period. The post-discovery period included all movements made from the moment the joystick moved into the hotspot until the end of the trial. Figure 3.9 depicts this by showing only those movements that occurred during the post-discovery period. Unlike experiments 1 and 2, both the pre-discovery and post-discovery periods are of interest. The distance travelled during the pre-discovery period gives us a metric of performance: shorter distances indicate that the participant has learned to move into the correct position more efficiently. The

behaviour during the post-discovery period is also of interest as it is indicative of the necessarily non-contingent behaviour that occurred during a trial. In other words, this design allowed us to measure a participant's performance and also to investigate how behaviour in the post-discovery period might impact on behaviour during the pre-discovery period.

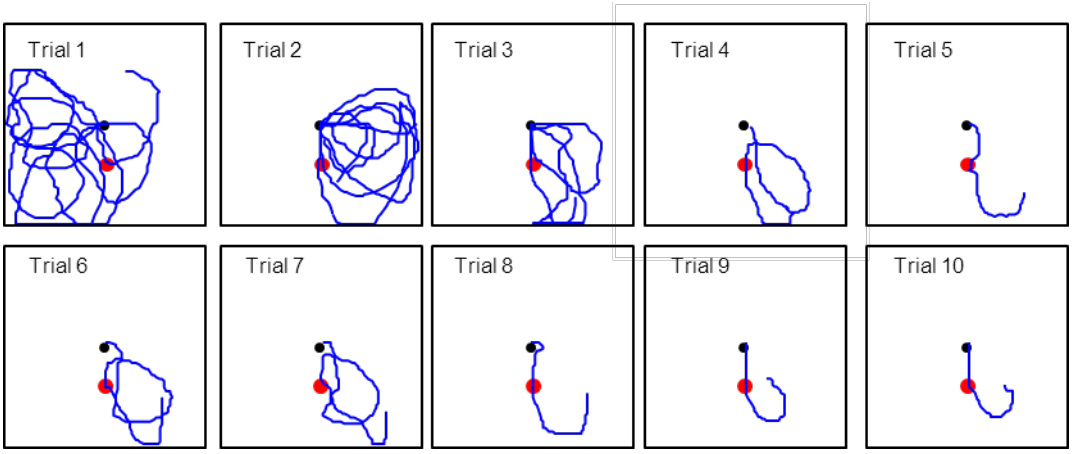

Figure 3.7 Movement traces showing all movements made during a batch of trials under the 450-ms delay condition. The black dot represents the centre of the search space and starting point on each trial. The red dot is the hotspot.

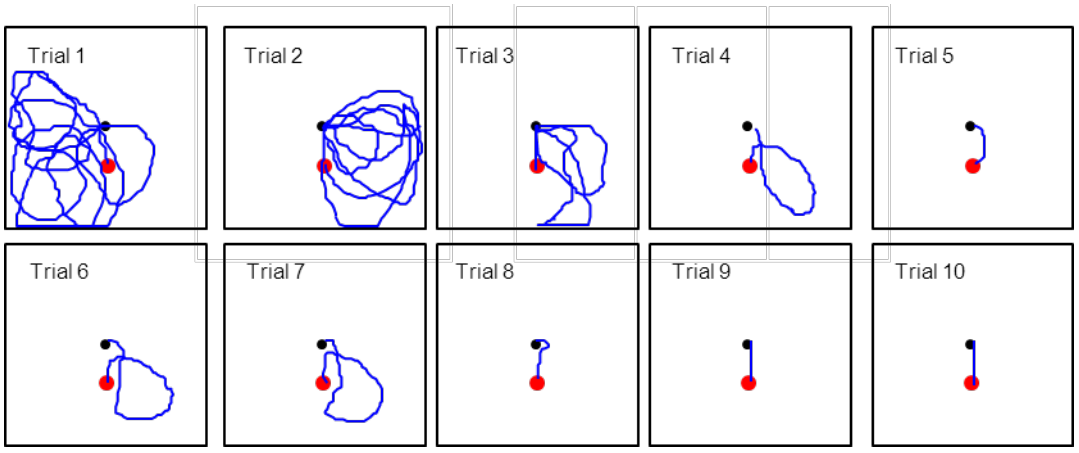

Figure 3.8 Movement traces showing only pre-discovery movements during a batch of trials under the 450-ms delay condition. The black dot represents the centre of the search space and starting point on each trial. The red dot is the hotspot.

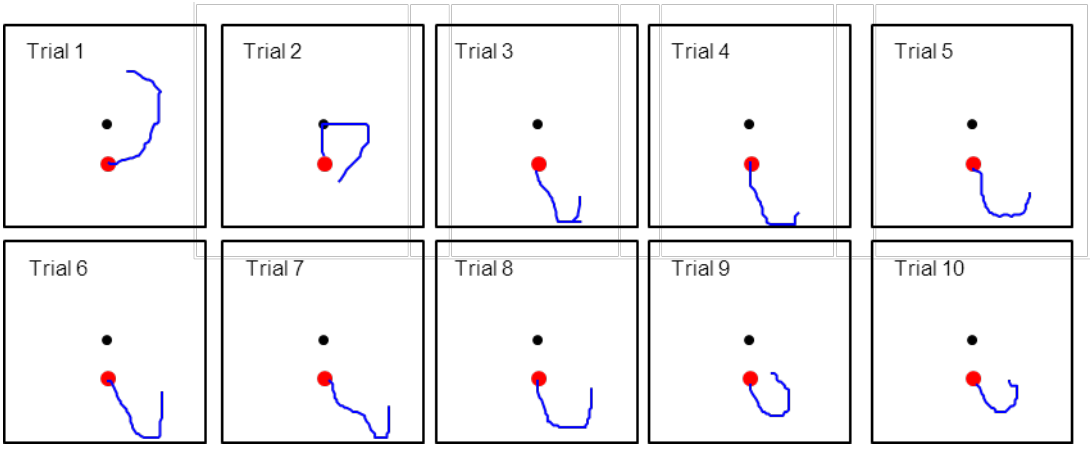

Figure 3.9 Movement traces showing only post-discovery movements during a batch of trials under the 450-ms delay condition. The black dot represents the centre of the search space and starting point on each trial. The red dot is the hotspot.

Performance and pre-discovery distance

Before running the analysis to determine the effect of delay, a test was run to check for a relationship between the optimum distance and the irrelevant pre-discovery distance. A Pearson product-moment correlation was run to test the relationship between the optimum distance and the log-transformed irrelevant distance. This revealed that there was no significant (2-tailed) correlation for the 0-ms (r = 0.3, n = 30, p = .11), 300-ms (r = 0.2, n = 30, p = .3) and 450-ms (r = -0.01, n = 30, p = .95) conditions, but that the 150-ms condition did display a positive correlation (r = 0.55, n = 30, p < .01). Longer optimal distances were therefore correlated with longer irrelevant distances in 1 of the 4 conditions. Because of the inconsistency of this effect it was decided that optimal distance should not be used as a covariate.

The effect of delay was investigated in terms of the differences in the mean irrelevant pre-discovery distance. Just as with experiment 3, the mean was taken from the 9 non-naïve trials in a batch of 10. Better performance by this metric would be indicated by shorter distances. A one-way repeated-measures ANOVA was conducted. Mauchly's test indicated that the assumption of sphericity had been violated ($\varkappa^2$ (5) = 12.402, p < .05; therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.802). The results show that there was a significant effect of reinforcement delay on the unnecessary distance travelled during the pre-discovery period, F(2.406, 69.781) = 5.339, p = .002. Figure 3.10 shows that the effect of

reinforcement delay was to increase the irrelevant distance travelled during the pre-discovery period. Bonferroni corrected post hoc *t*-tests revealed that the unnecessary distance travelled during the 450-ms condition differed significantly from the 0-ms ($t(29) = 2.85$, $p < .05$), 150-ms ($t(29) = 3.4$, $p < .05$) and 300-ms ($t(29) = 4.7$, $p < .05$) conditions.

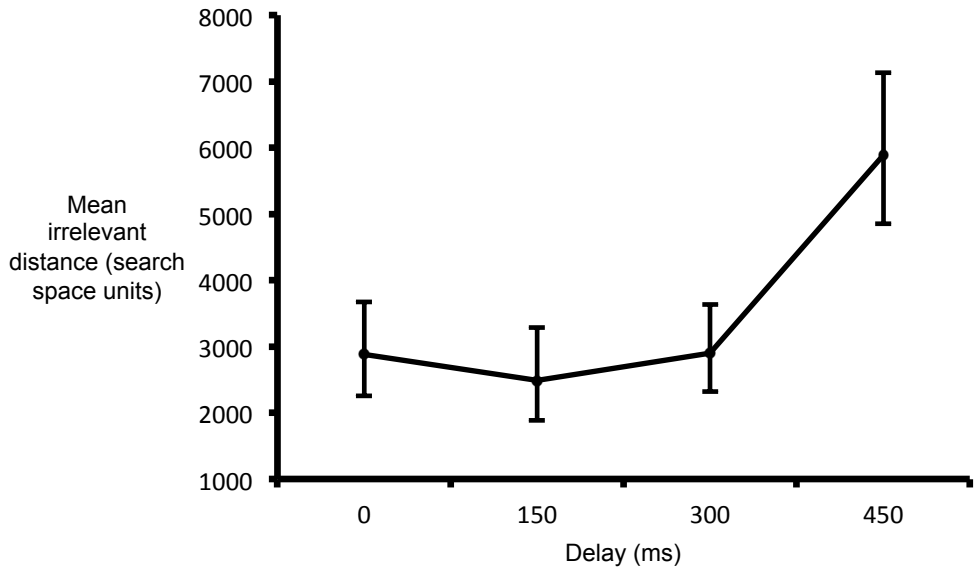

Figure 3.10 Mean irrelevant pre-discovery distance (and standard error) for the 4 levels of reinforcement delay. Values are back-transformed from the log transformation.

Learning

Experiments 1 and 2 included a threshold to determine the point at which a trial ended and this ensured that a trial couldn't be completed unless learning had taken place. Experiment 3 had no such threshold so it was important to check that at least some learning was occurring in each of the conditions; that is to say, whether the distance travelled in the later trials was shorter than that in the early trials. A learning ratio was calculated by dividing the irrelevant pre-discovery distance in trials 1 to 5 by that travelled in trials 6 to 10 of each batch. This metric not only provided a way to check for learning but also a means of testing whether there was an effect of delay on the extent of learning.

A one-way repeated-measures ANOVA was conducted on the log-transformed data. The results show that there was no significant effect of reinforcement delay on the

improvement in performance from early to late trials, $F(3, 87) = 0.43$, $p = .73$. In other words, whilst delay had a detrimental impact on overall performance, it doesn't appear to have had an effect on the relative improvement in performance over trials, which was large for all conditions (figure 3.11).

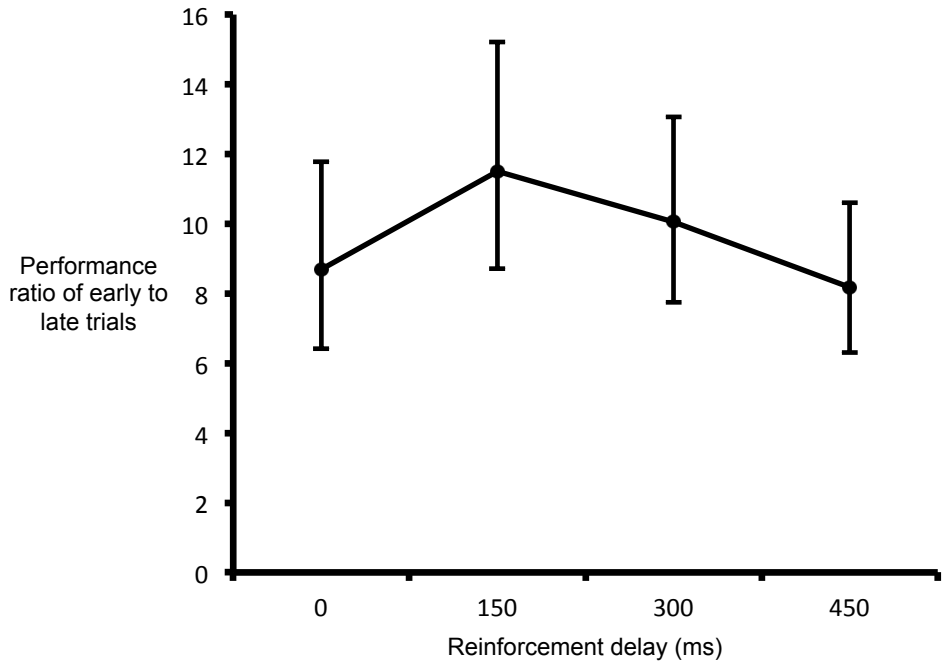

Figure 3.11 Mean performance ratio of early to late trials (and standard error) for the 4 levels of reinforcement delay. Values are back-transformed from the log transformation.

To test for differences in learning across trials, the distance for the early trials (trials 1 to 5) was compared to the distance for the late trials (trials 6 to 10) for all delay conditions. A two-way repeated-measures ANOVA was conducted on the log-transformed data. The results show that there was a significant main effect of reinforcement delay on the distance travelled, $F(3, 87) = 4.88$, $p = .003$; figure 3.12 shows that there was a general increase in the distance travelled with longer delay durations. There was also a significant main effect of phase (early or late trials), $F(1, 29) = 133.21$, $p < .001$; figure 3.12 clearly shows that the distance travelled in the late trials was shorter than that during the early trials. However, there was no significant delay-phase interaction, $F(3, 87) = 3.79$, $p = .13$; figure 3.12 shows that the improvement from early to late trials did not differ according to delay condition.

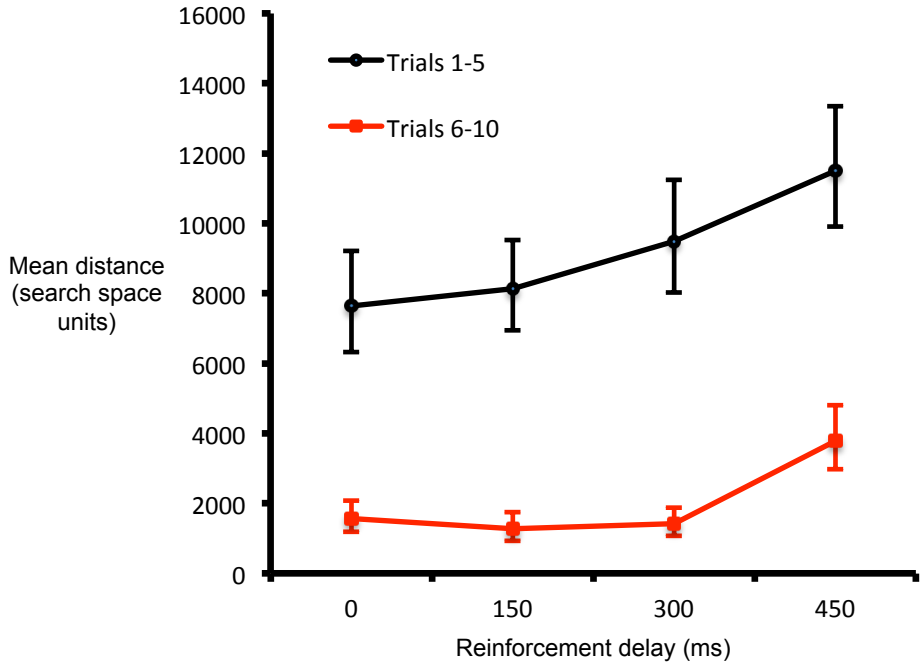

Figure 3.12 Mean distance for early and late trials (and standard error) for the 4 levels of reinforcement delay. Values are back-transformed from the log transformation.

Contamination

One of the predictions for a learning system with an eligibility period (whether it is indiscriminate or sensitive to the contents of the motor record) is that delayed reinforcement will hamper learning because it will contaminate the motor record with non-contingent output. If delayed reinforcement was hampering learning and performance due to contamination of the motor record then the effect of delay should be worse for some people than for others. Specifically, people who move further during the delay period will create more non-contingent motor output and therefore contaminate the motor record more than others. A test was therefore carried out in order to investigate the relationship between the pre-discovery irrelevant distance and the post-discovery speed for the 450-ms delay condition.

A Pearson product-moment correlation was carried out to test for a relationship between the log-transformed pre-discovery distance and the log-transformed post-discovery speed. It revealed that there was no correlation between irrelevant pre-discovery distance and the post-discovery speed (r = 0.22, n = 30, p < .13, one-tailed) (Figure 3.13).

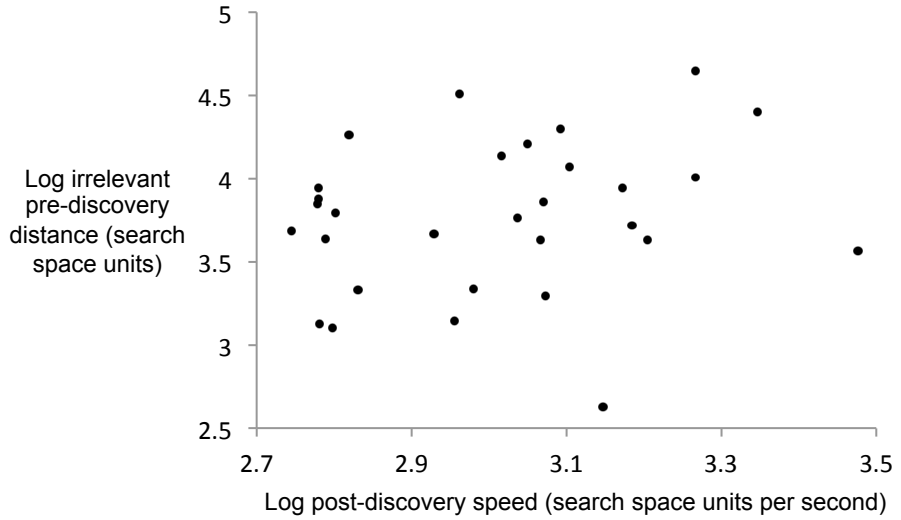

Figure 3.13 The relationship between the log of the post-discovery speed and the log of the irrelevant pre-discovery distance for the 450-ms delay condition.

Whilst speed gives an indication of the amount of contaminating (non-contingent) output that was occurring, it gives no indication as to the complexity of that movement. An estimate of the complexity of the movement during the post-discovery period was produced by calculating the change in the angle of movement over time (degrees per second). According to this metric, movement in a straight line would result in no change in angle and would be classed as a low complexity movement, whereas movement with lots of changes of direction would result in large changes in angle and would be classed as high complexity movement. If the type of movement during the delay period is an important part of the effect of delay because of how this impacts on contamination then we would expect large changes in angle during the post-discovery period to be correlated with longer irrelevant pre-discovery distances. A Pearson product-moment correlation was carried out to test for a relationship between the irrelevant pre-discovery distance and the post-discovery change in angle, it revealed no correlation between the two variables (r = -0.03, n = 30, p < .43, one-tailed) (Figure 3.14).

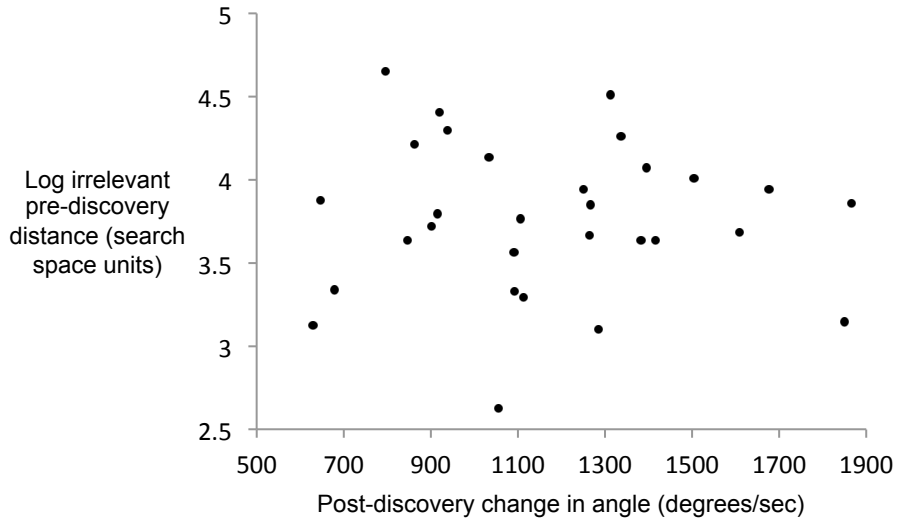

Figure 3.14 The relationship between the complexity of movement during the post-discovery period and the irrelevant pre-discovery distance for the 450-ms delay condition.

Discussion

Differing sensitivities to delay depending on type of motor task

The results showed that the effect of delay on performance was limited to the longest duration condition (450 ms) and therefore that the overall sensitivity to delay was considerably lower than for the task employed in experiments 1 and 2 where an effect was detected in the 75-ms condition. In terms of overall sensitivity, changing the repetition of reinforcement from within to between trials was akin to changing the task from a motor control (Foulkes & Miall, 2000; Miall & Jackson, 2006; Miall et al., 1985) or motor adaptation paradigm (Held et al., 1966; Kitazawa et al. 1995) to a reinforcement learning paradigm (Elsner & Hommel, 2004; Hollerman & Schultz, 1998; Shanks & Dickinson, 1991; Shanks et al., 1989). And yet both versions of the task were designed to be tests of reinforcement learning, with feedback characteristics that one would expect to find in such tasks (Jordan & Rumelhart, 1992). For instance, all reinforcement signals were qualitative, indicating whether or not the correct movement had been produced but providing no information on the relative level of performance. Reinforcement was also contingent on behaviour: if an individual had failed to move the joystick into the hotspot, they would have received no feedback at all.

One potential explanation for this result is that the important difference between tasks that target supervised learning mechanisms and those that target reinforcement learning mechanisms is not the informational content of the feedback but the frequency with which that feedback is delivered. The primary difference between the two versions of the joystick task in this sense is that feedback is delivered whilst movement is on-going in one case (within-trial repetition of reinforcement) whereas it signals the successful completion of a movement in the other (between-trial repetition of reinforcement). From the individual's perspective, the within-trial repetition experiments were effectively closed-loop tasks with the participants constantly adjusting their movements based on external feedback. In some ways this is similar to the situation in a standard motor control task (supervised learning) such as target pursuit (Foulkes & Miall, 2000; Miall & Jackson, 2006; Miall et al., 1985). By contrast, the between-trial repetition experiment is like an open loop task with the participants attempting to make the correct movement in the absence of feedback and on subsequent attempts adjusting the movement based on past experience.

An alternative explanation is that the differences between the two tasks call into question the distinctions that are sometimes drawn between different types of learning such as supervised and reinforcement learning (Jordan & Rumelhart, 1992; Wolpert et al., 2001). By employing a rich input device instead of a button- or lever-pressing response, the joystick task removed one of the most obvious distinctions between the experimental approaches to investigating supervised versus reinforcement learning. It could be that the difference in effect found here simply highlights the problems associated with attempting to investigate particular neural mechanisms at the behavioural level (Anderson, Fincham and Douglass, 1997). Definitions based on theories of learning or neuroanatomy might not accord well with the way behaviour pans out in reality. It might not make sense to refer to some types of learning independently of one another because they simply cannot operate independently in a normally functioning organism.

In terms of the theory of action acquisition offered by Redgrave and Gurney (2006) and Redgrave et al. (2008), the difference in sensitivity between the two different versions of the joystick task is difficult to interpret. It could indicate that the match

between the effective delay duration in experiments 1 and 2 and the onset latency of the phasic activity of dopamine neurons is purely incidental. In other words, the experiments were sensitive to delay for reasons other than the underlying learning mechanism. If this is true, then the theory needs to be able to account for the result from experiment 4 and also the range of findings from the reinforcement learning literature with which this result is in accordance. In other words, the current finding does not necessarily contradict the theory, but it does potentially call into question the explanation that the learning mechanism evolved to be short latency in order to avoid non-contingent contamination of the motor record. If the learning mechanism has evolved to work at such short latencies in a trade-off between the greater depth of processing that would come from a few hundred milliseconds of further processing (Redgrave et al., 2008) then we must assume that the detrimental impact of contamination is strong and presumably detectable at the experimental level.

One problem with comparing the susceptibility to delay in the two tasks covered here (and almost any two tasks featuring delay) is that they are not calibrated in terms of difficulty. Performance in difficult tasks is likely to suffer more when reinforcement is delayed so it could be the case that the task featuring within-trial repetition of reinforcement was simply more difficult than the task featuring between-trial repetition. This is not something that can be ruled out in the present case, but it could present a fruitful line of inquiry for future studies seeking to investigate differences at the methodological level between supervised learning and reinforcement learning.

Eligibility and contamination

As discussed in the introduction, the ability of animals to learn in the presence of delayed reinforcement implies the existence of some kind of short term record of movements and context, perhaps in the form of a gradually decaying eligibility trace (Singh & Sutton, 1996). It is assumed that any such trace must be time limited and previous research investigating response acquisition in the presence of delay indicates that the length of the trace could be in excess of 10 s (Lattal & Gleeson, 1990; Okouchi, 2009; Snycerski et al., 2004; Snycerski et al. 2005). However, it is also possible that the number and type of actions that occur within the eligibility period are also important,

producing more or less contamination depending on the type of behaviour involved. Consequently, it was suggested that we might be able to find evidence of movement specific contamination in the joystick task in the form of a correlation between the quantity or complexity of movement during the delay period and performance.

The results, however, revealed no correlation between performance and complexity or between performance and the speed of movement during the delay period and thus provide no support for the hypothesis that the amount of movement is a factor in learning with delayed reinforcement in addition to the duration of the delay period. In a sense, this result is not wholly unexpected. One of the major differences between the working memory system mentioned in the introduction (Baddeley, 2003) and the reinforcement learning process under investigation here is that the latter works at a very low level and, as we have noted, is essentially indiscriminate in nature; the latency at which phasic dopamine activity occurs precludes any rich information processing. However, the idea that any learning system would be unaffected by information load is, nonetheless, surprising.

Of course, it is possible that the experiment simply didn't provide the necessary conditions to enable the detection of an effect of contamination. For instance, longer delay durations would have provided more opportunity for the occurrence of non-contingent movements and it is conceivable that the amount of movement must reach some kind of threshold before it exerts a load over and above the basic effect of a decaying eligibility trace. Another explanation is that the effect of delay was the result of an entirely different brain process. For example, it is possible that the learning might reflect the effect of the declarative guidance of behaviour. Rather than having to rely primarily on the reinforcement of recent motor output, the participants might have been choosing successful shapes of movement or spatially inferring the movement they were required to make. If so, the relative complexity and distance of the movement might be unimportant as the overall trajectory would be object of learning. This issue is discussed further in the general discussion below.

General Discussion

The influence of declarative processes

The theme across all the experiments in this thesis is to develop and assess an experimental paradigm designed to record action acquisition in humans. Chapter 2 introduced the joystick task and detailed two studies featuring within-trial repetition of reinforcement and the present chapter has detailed a variant on the task featuring between-trial repetition of reinforcement. The advantages of between-trial repetition have already been discussed and certainly, the emergence of a shape of movement over trials, visible in figures 3.1 and 3.7, captures an intuitive sense of how it is that actions are learned. Indeed, it is possible that in such cases, the trajectory of movement became much more important than the need to achieve a particular end point. In this sense, at least some participants could have been solving the problem of 'what movement will bring an end to the trial' as opposed to 'where in the search space does the joystick need to be in order to bring an end to the trial'; or, more simply, learning 'what' instead of 'where'. This is presumably similar to the type of learning that is involved in producing a gesture and thus variations on the current task might present a useful way of investigating this type of learning for future research.

However, whilst many of the trials showed a stable progression of a particular shape of movement, a substantial number also displayed jumps in performance over the course of a single trial and, in a handful of cases, such as that shown in figure 3.15, seemingly one-shot learning. One reason for examples of learning such as this is that the basic structure of the task – location finding – meant that it was relatively easy to accidentally produce near optimal performance on a given trial. In other words, the path taken from the central starting position in an attempt to explore the search space could result in the joystick moving straight through the hotspot. Whilst it would be wrong to conclude that such cases definitely do not represent the kind of learning that the task was meant to record, the general presence of rapid learning in the experiment raises the possibility that performance in the task is driven to a large extent by learning at the declarative level.

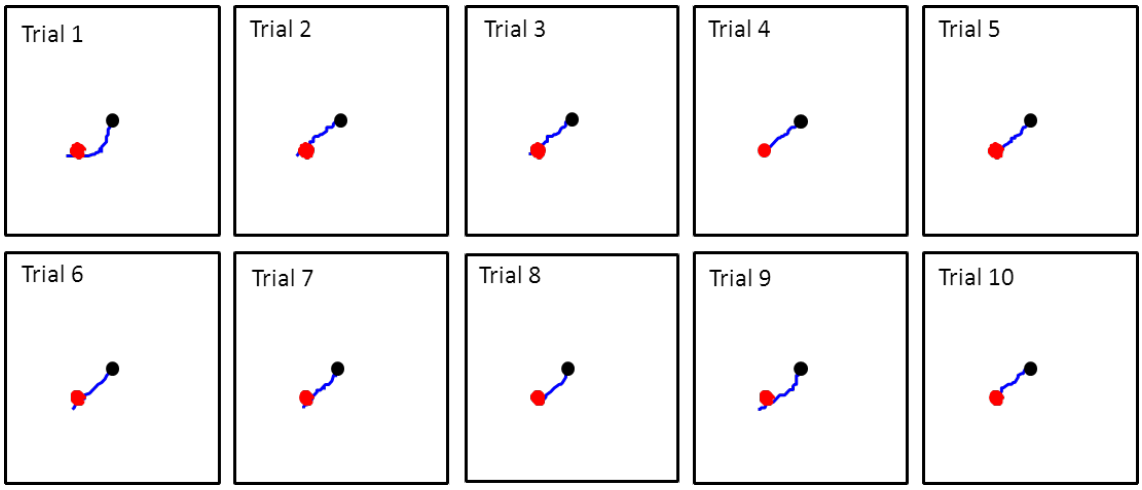

Figure 3.15 Data taken from experiment 4. The images show a batch of trials where a lucky first trial is followed by near perfect performance in all subsequent trials.

Avoiding the effects of declarative learning in humans is extremely difficult. In the experiments presented so far, attempts were made to reduce these effects. For instance, all participants were aware that the task was to move the joystick to the correct position. This intervention appears unusual at first as it removes such a large portion of what the participant must discover. However, whereas Thorndike (1911), for example, could rely on animals not to make a leap of insight; human participants, by contrast, can be relied on to do the opposite: the question is not so much if they will work something out but when. By giving some information about the task demands, the intention was to as far as possible avoid large shifts in performance caused by learning or insight at a declarative level. In other words, it was not possible for a participant to leap to the conclusion that the task was to search for a particular joystick position as this information was already provided.

It would be impossible to create a version of the joystick task that we could guarantee had excluded the contribution of spatial memory and general high-level cognitive guidance of behaviour (Anderson, Fincham and Douglass, 1997). However, whilst we might not be able to achieve process purity in the joystick task, it is, nonetheless possible to change the balance of the paradigm from one that might be more declarative in nature (i.e. involving the selection of a particular reaching movement) to one that is more nondeclarative (i.e. forcing the motor system to extract contingent motor output from multiple samples of self-generated behavioural variance). A potential way of achieving this is to make the required movement more complex and

therefore more difficult to infer spatially. This would have the benefit of making it harder to accidentally produce optimal performance and would give more scope for learning over and above the declarative guidance of behaviour. To this end, chapter 4 details the development of a gestural version of the joystick task and an investigation into the potential contribution of higher-level learning processes.

Chapter 4: declarative and nondeclarative components of action learning.

The versions of the joystick task described in chapters 2 and 3 provided simple and flexible approaches to the study of action acquisition. The task featuring within-trial repetition of reinforcement provided a particularly useful approach for presenting numerous new learning scenarios and a means of extracting a low variance learning metric in the form of the number of instances of reinforcement. It also offered a way of delivering large numbers of discrete stimulus presentations, ideal for investigating the effects of special stimulus properties on acquisition. The task featuring between-trial repetition of reinforcement provided a more traditional method of delivering reinforcement and as such represented a model of response acquisition that can be reasonably compared to response acquisition paradigms involving nonhuman animals. It also offered a means of recording the evolution of a movement trace over time, allowing learning to be quantified relative to optimal performance in a way that is finer grained than would be possible with a button-push paradigm. The present chapter describes a version of the task that requires the participant to generate more complex movements, the aim being to shift the task further towards the low-level learning processes that are the subject of the current research. It furthermore describes an attempt to investigate the contribution of high-level cognitive and spatial learning systems to the joystick task.

The declarative-nondeclarative split

There is a tradition in psychology and neuroscience of drawing a distinction between those mental processes that are described as flexible, insightful and conscious and those that are described as rigid, habitual and unconscious. In the field of long-term memory, for example, Squire (2004) summarises memory classifications with one main grouping criterion: a distinction between declarative and nondeclarative memory. He describes the difference as follows:

> Declarative memory is representational. It provides a way to model the external world, and as a model of the world it is either true or false. In contrast, nondeclarative memory is neither true nor false. It is

> dispositional and is expressed through performance rather than recollection. (p.173).

A similar distinction is made in the field of animal learning, with Balleine and Dickinson (1998) explaining that their "analysis of instrumental learning conforms to the popular distinction between declarative and procedural learning with contingency learning being declarative in nature and habit learning procedural (p.412)". This is further echoed in the field of skill learning and performance:

> Automatic processing is activation of a learned sequence of elements in long-term memory that is initiated by appropriate inputs and then proceeds automatically—without subject control, without stressing the capacity limitations of the system, and without necessarily demanding attention. Controlled processing is a temporary activation of a sequence of elements that can be set up quickly and easily but requires attention, is capacity-limited (usually serial in nature), and is controlled by the subject. (Schneider & Shiffrin, 1977, p.1).

In other words, many different approaches to the study of learning, memory and performance have converged on this compelling dichotomy. As we will see, the distinction has important implications for the study of action acquisition.

Perhaps the most convincing attempt to operationalise this difference has come from research into instrumental learning, employing techniques of overtraining and reinforcer devaluation. Adams (1982), provided the template for future research with a series of experiments investigating the effects of various degrees of exposure to reinforcers and response practice on the tendency of rats to behave as if they understand the consequences of their actions. The first experiment saw rats trained to press a lever in return for sucrose pellets. Half of the rats received a low amount of training and the other half received a relatively high amount of training. Following this, half of the low training group and half of the high training group received injections of lithium chloride paired with access to the sucrose pellets (thus devaluing the reinforcer) whilst the remaining two groups received unpaired injections (causing sickness without devaluation). All of the rats were then given an extinction session during which the number of lever responses was recorded. Adams found that the low training devaluation group had a lower rate of response during extinction than the low

training control group, indicating that the rats in the devaluation group had understood that pressing the lever would deliver food that was no longer palatable to them (i.e. similar to Adams & Dickinson, 1981). By contrast, no difference was found between the high training conditions, suggesting that the high training devaluation group had not integrated the devaluation with their understanding of the lever-food contingency. The implication of this finding is that behaviour initially under goal-directed control can be transformed into a habit through prolonged training. This experimental procedure provides an excellent example with which to compare other examples of declarative or nondeclarative learning.

Adding components to the behavioural repertoire and 'accidental' performance

Whilst Adams (1982) was able to demonstrate a convincing distinction between habitual and goal-directed elicitation of actions from the behavioural repertoire, an equivalent distinction regarding the different brain processes that might be involved in the acquisition of actions has been relatively neglected. This is presumably due in part to a general neglect of research into action acquisition in favour of the maintenance of actions (Lattal & Gleeson, 1990), but it also relates to methodological problems associated with demonstrating that nonhuman animals can produce not only the kind of uninsightful learning reported by Thorndike (1911), but also the kind of insightful learning that humans are capable of (Bayern, Heathcote, Rutz & Kacelnik, 2009).

The apparent tendency of animals to learn laboratory tasks uninsightfully has allowed researchers to gain some understanding as to how important it is for an essentially indiscriminate learning system to be exposed to the right kind of behavioural variance. The technique of shaping (see, for example, Petersen, 2004), for instance, is an exercise in constraining the behavioural variance of an animal subject. The experimenter effectively takes on the role of an external declarative system for the animal by identifying appropriate behavioural variance and ensuring that the animal executes the desired movements by reinforcing successive approximations of the desired behaviour. At the other end of the spectrum is the kind of gradual, error laden learning investigated by Thorndike (1911) in which no guidance of behaviour occurs

and, still more extreme, the phenomenon of superstitious learning in which the animal's learning systems are exposed to non-contingent reinforcement. Skinner (1948), for example, describes a study in which pigeons were placed into operant chambers and given access to a food hopper at regular intervals, irrespective of their behaviour. He found that in 6 out of 8 cases, operant conditioning took place: the pigeons developed stereotyped responses to the hopper, such as pronounced head-tossing, even though delivery of food was not contingent on those responses. In other words, Skinner's work suggests that if inefficiencies are sufficiently consistent, they can evolve into stable, and yet completely irrelevant, actions in the presence of non-contingent reinforcement. It should be noted, however, that the explanation of superstitious learning in terms of the reinforcement of recent motor output has been disputed (Aeschleman, Rosen & Williams, 2003; Timberlake & Lucas, 1985).

Finding evidence of such nondeclarative learning is notoriously difficult and the interventions and scenarios that are required to provide convincing demonstrations of such learning tend to be extreme. Examples include stochastic obscuring of associations (e.g. Knowlton, Mangels & Squire, 1996), dual task procedures (e.g. Nissen & Bullemer, 1987), implicit priming (e.g. Leiphart, Rosenfeld & Gabrieli, 1993) and lesion studies (e.g. Bayley, Frascino and Squire, 2005). Whilst such techniques are useful for the purpose of measuring nondeclarative learning in general, they are not readily compatible with the paradigm presented here for the general investigation of action acquisition in healthy subjects.

What characterises action acquisition in those nonhuman animals that are unable to learn through insight is that they are largely at the mercy of external constraints and current drives for the behavioural variance they produce and this in turn influences what they are able to learn and how fast they are able to learn it. The consistent yet irrelevant behaviour acquired during superstitious learning provides a caricature of the type of situation that the declarative guidance of behaviour helps humans to avoid. However, it also points to a possible hallmark of nondeclarative action acquisition that is of potential use in human research. Whilst pure examples of superstitious learning are, presumably, rare, perhaps more common are situations in which a successful task-relevant action is learnt which contains consistent inefficiencies; that is to say, an

action which contains movement components that are unnecessary but are nonetheless faithfully repeated each time the action is produced. The ability to identify such consistent inefficiencies in the joystick task under some circumstances but not others would provide some support for the idea that an indiscriminate learning mechanism is capable of driving action acquisition.

As discussed in chapter 3, we have thus far attempted to reduce the effects of the powerful declarative abilities that human participants bring to all tasks, by disclosing information about task demands before testing commences. However, whilst this intervention at the level of task instructions might have reduced the number of different approaches that were taken to the task, it also had two potential limitations. Firstly, by increasing a participant's knowledge we make the task easier to learn and therefore decrease the number of trials over which learning occurs, thus reducing the amount of effective data that can be collected. Secondly, it greatly reduces the likelihood of learners developing consistently inefficient behaviour: that is to say, extracting a suboptimal movement trace from their own behavioural variance. An alternative is to make the movement on which reinforcement is contingent more complex. In this way, the task instructions can remain explicit, but we provide more scope for participants to display inefficiencies of behaviour and increase the chances that learning will occur over a greater number of trials.

Testing the contribution to action acquisition of declarative versus nondeclarative processes

The learning mechanism proposed by Redgrave and Gurney (2006) does not just concern nondeclarative learning. Indeed, it is part of their position that this mechanism helps animals to determine whether they are agents of events as well as how to elicit those events by time stamping the information converging on the striatum at the moment of reinforcement and reselecting movements (and seeking out context) that immediately preceded the reinforcement. Under normal circumstances, it is assumed that this process would be intricately tied in with declarative processes. However, in the case of the joystick task, any such low-level mechanism for linking events with behavioural output would manifest itself in behaviour that has competing explanations in terms of high level executive systems and spatial learning. Specifically,

the ability to home in on a particular hotspot location could be based on an explicit memory of the hand position and simple inference of the path required to move to that point or it could be based on the stamping in of recent motor output.

In order to investigate the contribution of high-level spatially guided learning and low-level reinforcement learning, an experiment was carried out to measure acquisition under conditions featuring different amounts of spatial information and different types of practice. Spatial knowledge of the task was manipulated by either providing or concealing the precise location of hotspots during a learning phase. The type of motor output was also manipulated during the learning phase by varying whether or not participants had access to a visual representation of the current location of the joystick relative to the hotspots onscreen. The effect was measured in terms of performance in a test phase during which no visual information was provided onscreen about either the hotspot locations or joystick position relative to these locations. The overall aim was to determine how readily the joystick task – in many ways a spatial paradigm – relied on non-spatial processes and therefore, the potential value of this paradigm for widespread investigation of the process of interest. In other words, the aim was not to detect nondeclarative learning at all costs but to provide conditions in which its effects could show through in spite of competition from other learning processes.

The first question addressed concerned the issue of whether action acquisition relies primarily on the spatial guidance of movements when full information is available regarding the specific form that the action must take. If action acquisition in the joystick task relies primarily on spatial guidance rather than the learning of particular movements, then performance during a blind test phase should not differ depending on the type of behavioural variance produced during a learning phase in which participants have the same access to information about the locations of contingent areas of the search space. By contrast, if performance is influenced by the movements that have been practiced during a learning phase (i.e. the motor output that has been reinforced) then performance during a blind test phase should differ depending on the type of behavioural variance produced during the learning phase.

The second question concerned the issue of whether action acquisition relies primarily

on spatial information when the task is demonstrated in a spatial way, but no information regarding specific locations is provided. If people use their exploration of the search space to infer the location of contingent areas such that they can guide their movements using a spatial model of the environment, then performance during a blind test phase should not depend on whether particular spatial locations are provided visually or determined through exploration during a learning phase. By contrast, if learning depends on the movements that have been practiced, then performance during a blind test phase should differ depending on how the target locations were discovered, because discovery under blind conditions should result in more behavioural variance than when the positions are visually represented onscreen.

A third question concerned the variability of movements relative to the point of reinforcement. If, as Redgrave and Gurney (2006) suggest, the system controlling learning is a time stamp that reinforces motor output immediately preceding the moment of reinforcement, then we might expect that movements closest to the point of reinforcement should be less variable than those more distal; as they put it, "the maximal positive/negative reinforcing effect of [dopamine] would be directed to immediately contiguous motor efference copy" (p.973). Such a result would not be expected if the correct movement was simply inferred from exploration of the search space and performed based on a spatial model of relevant locations.

The final question concerns the issue of inefficiencies in the movements that are acquired. In a situation where participants understand the general structure of the movement they are required to make and are then given the opportunity to infer the specific nature of this movement through exploration of the search space, we would not expect them to develop particularly inefficient movements. However, if learning is strongly influenced by the reinforcement of motor output then the movements learnt will be more dependent on the variance of the exploratory behaviour and more likely to include inefficiencies.

Experiment 5: gesture discovery

Method

Participants

26 people (20 female) participated in all conditions of this study. Ages ranged from 18 to 20 years with a mean age of 19 (*SD*. = 0.6 years). Participants were all undergraduate students at the University of Sheffield who took part in the study in return for credits in the department's research participation scheme. All subjects were naive to the purpose of the experiment.

Defining the action

In the task featuring between-trial repetition of reinforcement, the task of locating a single location in space meant that it was relatively easy to accidentally produce the correct movement. The problem with this was that, on such occasions, the amount of behavioural variance from which the correct movement needed to be extracted was extremely small and this limited the value of the data that could be obtained. A potential solution to this problem is to make the hotspot smaller, thus making it harder to find the location by accident and return to it on subsequent attempts. However, experiment 3 indicated that, whilst smaller location sizes did indeed make the task more difficult to learn, there was a limit as to how small a hotspot could be made and still result in useful behavioural data. Very small hotspots introduce unwanted experimental error arising from the limitations of the apparatus. The limited precision of the analogue joystick mechanism meant that a small amount of noise was present in all movements. In essence this meant that the movement recorded by the apparatus was potentially not as smooth as the participant's input. The significance of this issue would necessarily increase with smaller targets. Small hotspots also place too much emphasis on fine motor control as opposed to action discovery. Consequently, a different type of action, a 'gesture', was defined for the current experiment which would increase the complexity of the required action without increasing the precision of movements required to perform it.

Whilst the type of action defined in this experiment will here be described as a gesture, it is important to explain what is meant by this term for the present purposes. In cognitive psychology and zoology a distinction is made between egocentric and

allocentric representations of items in space: items may be represented relative to the animal's own body position (egocentric) or relative to one another (allocentric). The former type of representation is the kind we use when reaching and manipulating objects and the latter is the kind of representation that allows us to navigate through an environment from novel starting positions and would enable us to draw a birds-eye view room plan (Galati et al. 2000; Wang & Spelke 2000). This issue is relevant to how gestures are defined in the current experiment. Gestures are typically thought of as highly practiced stereotypical responses that can be performed in a range of contexts. In order to demonstrate such gesture learning in the joystick task, a participant would need to learn not just a particular shape of movement, but that this shape can be produced anywhere within the search space and still result in reinforcement. This is a strong definition of a gesture and it is not how the term will be used in the current experiment. A gesture in the current experiment simply refers to a particular shape of joystick movement, always in the same position relative to the central starting point. The required movement can only be expressed through the joystick in this task and, to this extent, the required movement is always the same relative to the participant's body, irrespective of where their body is situated in space. From a purely spatial perspective, the type of learning that the joystick task aims to investigate, therefore, is egocentric rather than allocentric in nature.

The gesture was defined in terms of a movement through 3 hotspots. The hotspots were randomly placed on an annulus within the search space, as defined in chapter 2, with the additional rule that they could not overlap one another. The aim of the task was to move the joystick once into each of the 3 hotspots in the correct order, as illustrated in figure 4.1; once the joystick had moved into hotspots 1, 2 and 3 (in that order), the trial would end and the next trial would commence immediately.

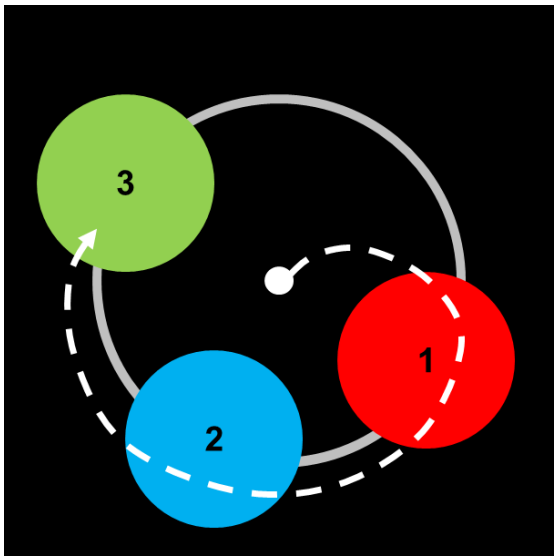

Figure 4.1 Arrangement of hotspots within the search space. The centre of each hotspot is placed randomly on the grey annulus such that the centre of the screen, the edge of the screen and other hotspots are never overlapped. The white dotted line indicates a path that could be taken through the hotspots in order to bring an end to the trial. Hotspots are drawn to scale.

An unavoidable consequence of varying the quantity of information provided to participants is that this would in turn affect the amount of exploration required in order to produce a successful movement. Whilst exploratory behaviour is informative, it is, nonetheless, desirable to be able to remove this behaviour from analysis in order to better focus on other aspects of performance. The primary reason for defining the gesture in terms of a movement through 3 discrete areas of the search space was to provide a way of splitting the trace into separate components at the methodological level rather than doing so post hoc. Due to the way the gesture was defined, the trial could be split into 2 main periods of movement (figure 4.2). The first period is an 'exploration' phase including all movements leading up to a successful entry into the first of the 3 hotspots. The second period, the 'gesture' phase, is the period that includes only the successful movement through the 3 hotspots. This distinction is important because even though the joystick task employs a stable central starting position, the successful movement as executed on a given trial will not always originate from this central point, especially when a participant is required to actively hunt for the correct movement. Thus, by defining a separate gesture phase it is possible to isolate a portion of the movement trace that always has a stable starting position irrespective of how much exploratory behaviour took place in a trial; this is therefore the stable portion of the movement trace that is reinforced trial on trial.

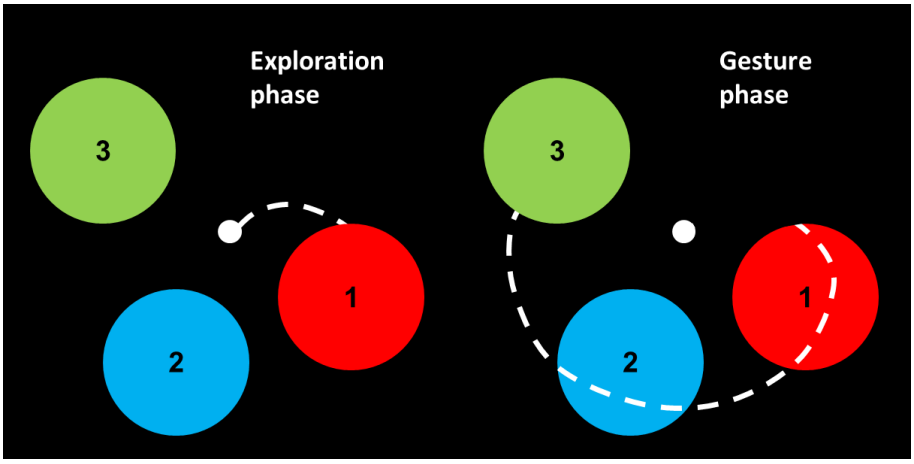

Figure 4.2 The 2 main periods of movement within a trial. The exploration phase includes all movements leading up to the successful entry into the first hotspot. The gesture phase includes only those movements following successful entry into the first hotspot.

A related benefit of defining the gesture in this way was to provide a means of defining portions of the movement trace that differed in terms of their proximity to the point of reinforcement. In other words, the gesture phase shown in figure 4.2 can be split into 2 further portions of movement (figure 4.3). Phase 1 includes all movements from the point of entry into the first hotspot up to the point of entry into the second hotspot. Phase 2 includes all those movements from the point of entry into hotspot 2 up to the point of entry into the third (and final) hotspot. Consequently, we can identify 2 portions of the movement on which reinforcement is contingent but that differ in terms of their temporal and spatial proximity to the point of reinforcement.

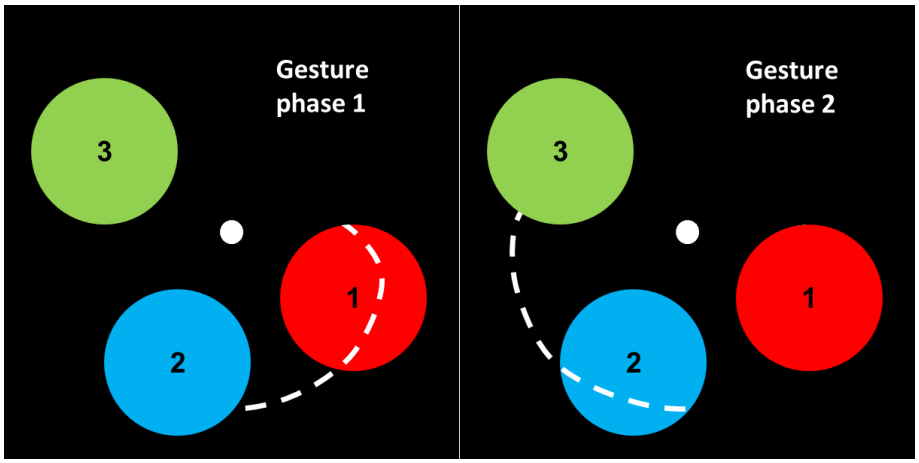

Figure 4.3 The gesture phase of the movement (shown in figure 4.2) can be split into 2 further periods of movement that vary in terms of their proximity to the point of reinforcement. In this diagram, gesture phase 2 is both temporally and spatially closer to the point of reinforcement (i.e. entry into hotspot 3).

The final reason for defining a gesture using separate, discrete regions of the search space was to reduce the constraints on the form that the final movement should take. A brain process that relies on the reinforcement of recent motor output could only differ from other brain processes if movements are allowed to vary from one trial to the next. In other words, the range of possible movements that can result in reinforcement must be large enough to ensure that trials differ not only in exploration distance but also in gesture distance. If a narrow channel of movement had been defined instead, then all participants would have been forced to move through this channel, thus restricting the behavioural variance that could be generated during the gesture phase and, consequently, reducing the opportunity for the characteristics of a particular learning system, be it spatial or reinforcement learning, to express itself in the data.

Hotspot size

The hotspot size chosen for this task was large relative to the size of the hotspots employed in the earlier experiments, each hotspot occupying 8.2% of the search space. The size of the hotspots relative to the search space is illustrated by all of the above figures (e.g. figure 4.1). The size was determined through pilot testing and chosen based on 3 main considerations. Firstly, the task could not be too difficult because there was a need to ensure that participants of different abilities could complete the requisite number of trials and demonstrate some improvement during the limited learning phase, regardless of which condition they were performing. Secondly, it was important that the emphasis should be on the extraction of a particular shape of movement and not on an individual's ability to demonstrate fine motor control. And thirdly, the task was designed so that it was possible to complete it spatially under blind conditions. It is conceivable that the use of extremely small hotspots might force people, under blind conditions, to adopt the kind of trial and error low-level learning that is the focus of this research; however, the issue here is not whether such learning could possibly occur, but whether it plays an important role in learning in the joystick because of the way the feedback (reinforcement) is presented. If the task was so difficult that the participants couldn't work out the

locations of the hotspots spatially from their behavioural variance then there would be no way of knowing whether this was indicative of a general tendency to perform the task non-spatially or whether it was caused by the challenging task conditions. By contrast, if the task is sufficiently easy that it can potentially be completed based on spatial inference, then this provides a better indication as to which learning system people are inclined to rely on and the contribution that the learning system of interest is likely to make to action learning in general.

Phases and trials

The experiment was split into 2 phases: a learning phase featuring 3 different stimulus conditions followed by a blind test phase which featured no stimuli irrespective of what was experienced during the learning phase.

The number of instances of reinforcement (trials) during the learning phase was restricted to 25. This was to ensure that all participants at the test phase had received the same amount practice on a particular arrangement of hotspots. At the test phase, it would have been possible to assign a threshold along with a requirement that a minimum number of trials be completed; however, the methodological advantages of using a threshold were not sufficient to outweigh the practical advantages of having a consistent number of trials for all attempts. Whilst a threshold was important in experiments 1 and 2, this was because the required action was defined in terms of the rate at which reinforcement was received, so choosing a threshold was necessary in order to be able to compare performance between conditions. When utilising between-trial repetition, by contrast, performance can be considered separately for each instance of reinforcement (e.g. distance travelled during a given trial) so there is no necessity to wait for a threshold to be achieved before performance can be measured.

The conditions

The independent variable was the amount of location specific information provided during the learning phase of the experiment. This was manipulated by the use of 3 stimulus conditions (figure 4.4): 1, no visual information provided (blind); 2, the

positions of the hotspots shown on screen (half); and 3, the positions of the hotspots and the current location of the joystick relative to them all shown onscreen (full). In the blind condition, participants would be forced to extract the relevant movements or infer the hotspot locations from their exploration of the search space. In the half condition, the relevant locations were provided directly to the participants, so all that was required was to translate these positions into correct movements of the joystick either by extracting this from behavioural variance or by more deliberate guidance of the joystick. The full condition required nothing more than to move a dot representing the current location of the joystick into the relevant hotspot positions, thus there was no need to discover the locations or how to move the joystick into them and thus behavioural variance was constrained by the participants themselves. The correct order of the hotspots was not revealed in any of the 3 conditions. In the full and the half conditions, the hotspots were all displayed in the same colour (red) and had no identifying features so there was no way of determining the correct order of the hotspots without exploring the search space and awaiting reinforcement.

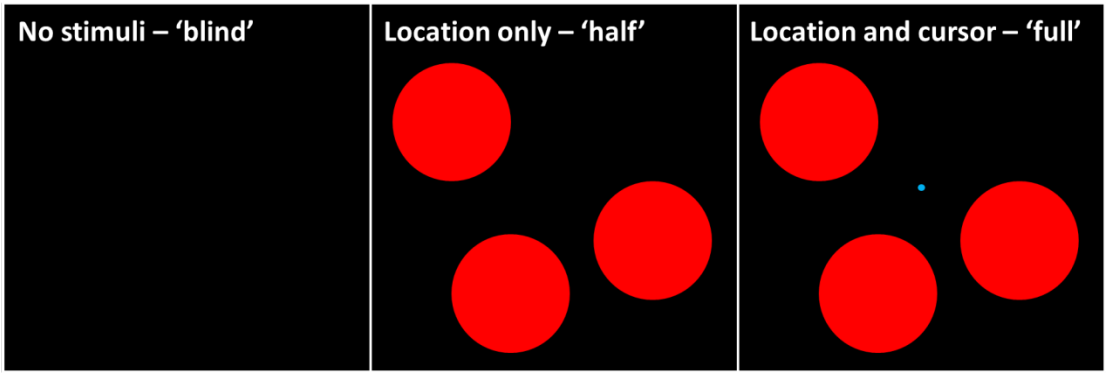

Figure 4.4 Visual information available to participants during the learning phase of the experiment for the 3 different conditions (hotspots shown in red and drawn to scale, cursor shown in blue).

The importance of the full condition was to help determine the extent to which the half condition forced participants to learn from their behavioural variance, if at all. The full condition provides a non-intrusive manipulation of behavioural variance. The main way in which the half and full conditions differ is in terms of the type of movements they cause to occur. During the learning phase, the full condition makes it easier to complete trials in a very direct, highly efficient manner, whereas the half condition requires a certain amount of approximation of movements, consequently, the

conditions should force participants to practice a different set of movements whilst having the same access to information about the location of the hotspots. If behavioural variance is not an important aspect in learning to complete the trials, then performance in these conditions should not differ at test phase when the stimuli are no longer available to guide performance. Specifically, the transition from the learning phase to the blind test phase should have minimal effect on the half condition but should negatively impact on performance in the full condition. The comparison between these stimulus conditions also allows us to infer whether or not there was a strong unintended negative impact during the transition from learning to test phase, caused simply by adaptation to the unfamiliar conditions. If such an effect was present, this would show up in decreased performance for both the full and the half conditions.

The blind condition provides a situation in which the correct movement or an understanding of the location of the hotspots must be extracted from the participant's own behavioural variance. Consequently it differs from the half condition in terms of how much information must be inferred or reinforced by practicing different movements through space. Performance in the test phase will give an indication as to how learning is occurring in the blind condition. If the blind condition results in similar performance at test phase to the half condition then this will be an indication that learning relies less on what was practiced and more on the ability to infer the correct movement and generate this movement in a spatially guided fashion.

Reinforcement

The reinforcing stimulus was the same as that used in experiments 3 and 4.

Instructions

As discussed in previous chapters, there are good reasons why we might wish to keep participants naïve as to what it is that they must learn during the task, the primary reason being that there is simply more for the participant to discover. However, once again, a decision was made to provide all participants with full task knowledge. The issue in the current experiment is that the independent variable (visual information)

also influences a participant's understanding of the task requirements. This means that if people starting in the blind condition were also naïve regarding the task structure, then the blind and the half conditions would differ not only in terms of knowledge about the particular locations of hotspots but also in terms of the understanding of the task requirements. Consequently, full task knowledge was provided prior to the experiment so as to reduce the differences between the blind and the half conditions. Prior to starting the experiment, all participants received three practice trials with full visuals (cursor and hotspot positions) as a visual explanation of what would be involved in the task. This happened just once for each participant and the arrangement of the hotspots in these three practice trials was always different to any they would subsequently encounter in the experiment. In other words the blind condition differed in that it required the specific arrangement of hotspot locations to be learned through exploration of the search space. Once again, the correct order of the hotspots was not revealed to the participants, irrespective of condition, so even in the full visuals condition, this needed to be learned from exploration of the search space.

Procedure

The participants were told that the experiment was designed to test their ability to learn gestures. The experimental program was started and they were asked to complete the 3 practice trials, which featured full visual information: hotspots and the relative position of the joystick were displayed onscreen. It was explained that the aim of the task was to move the joystick into the 3 hotspots in the correct order and that doing so would result in the end of the trial. It was further explained that they would have to work out the correct order by exploring the search space. Once the 3 practice trials were completed, the participants were told that the rest of the experiment would involve exactly the same task but that the visual information provided and the positions of the hotspots would change depending on the condition. They were further informed that it would be made clear to them that the hotspot positions had been changed. It was explained that for each new set of hotspot positions, they would have to complete 50 trials, the last 25 of which would always be blind. Following these instructions, the experimental trials were started. The participants were told at the start of each new condition that the arrangement of hotspots had been changed; they

were also told during a 2 minute interval between all learning and test phases that the locations of the hotspots would remain the same once the trials recommenced. No onscreen instructions were provided other than text to indicate that the current trial had been successfully completed: “You produced the correct movement!”. A summary of the protocol is presented in figure 4.5; it should be noted that the order of blocks 1, 2 and 3 was counterbalanced to ensure that participants experienced these blocks in different orders.

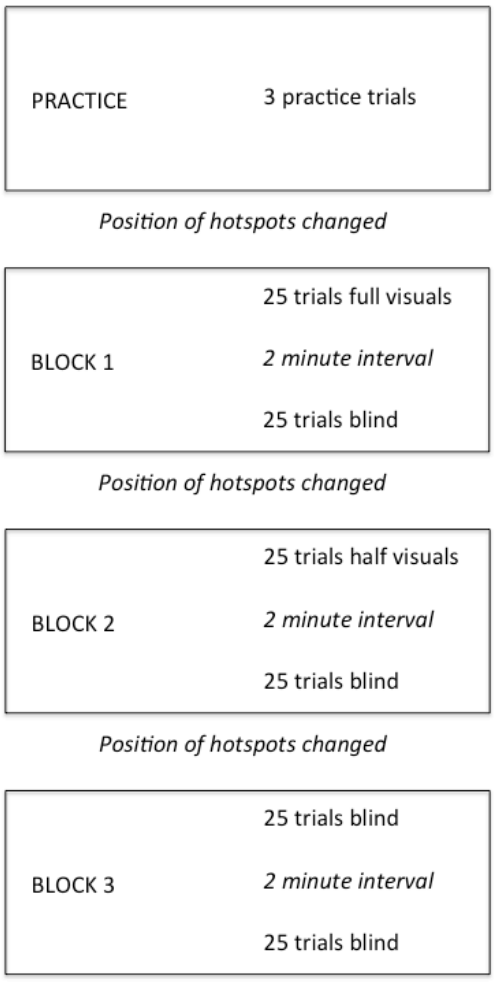

Figure 4.5 Summary of the experimental procedure. Blocks 1, 2 and 3 were counterbalanced so that different participants experienced the blocks in different orders. Each block comprised 25 learning trials followed by a 2-minute break and then another 25 test trials.

Design

All participants completed 3 practice trials featuring full information regarding the hotspot locations and current position of the joystick onscreen. For the experimental trials, each learning phase consisted of 25 trials, during which the positions of the 3

hotspots remained the same. Following the learning phase there was a 2 minute break and the test phase commenced, also comprising 25 trials with the hotspot positions remaining in the same place as they had been for the immediately preceding learning phase. The experiment had a repeated measures design, with all participants completing 3 learning and 3 test phases and therefore experiencing each of the 3 conditions. The position of the hotspots was changed at the start of each new learning phase. The order of the conditions was counterbalanced.

Results

Irrelevant distance

The first analysis was to summarise the basic performance measure – the irrelevant distance travelled during a trial – for each of the conditions at both learning and test phase. A two-way (2 x 3) repeated-measures ANOVA (phase by stimuli) of the log-transformed data revealed that there was a significant main effect of the phase of the experiment ($F(1, 25) = 4.47$, $p = .045$) and of stimuli ($F(2, 50) = 42.38$, $p < .001$). There was also a significant interaction between phase and stimuli ($F(2, 50 = 31.4$, $p < .001$). Figure 4.6 shows that the stimulus manipulation greatly affected the distance travelled. Bonferroni corrected post hoc *t*-tests confirmed this, revealing that, during the learning phase, both the blind ($t(25) = 3.7$, $p < .05$) and the full ($t(25) = 7.84$, $p < .05$) stimuli conditions differed significantly from the half condition, indicating that under these conditions, differing amounts of behavioural variance were being produced. The graph also shows that the means for the full and the half conditions converged between the learning and the test phases of the experiment. The half condition showed no change between the phases and did not differ from the full condition at test phase, whilst performance in the full condition significantly declined ($t(25) = 6.72$, $p < .05$) across the phases. This suggests that the shared feature of these conditions – knowledge of the hotspot locations – was the most important determinant of performance under blind conditions, notwithstanding the relatively large difference in the amount of behavioural variance produced during the learning phase. One possible explanation for why the performance in the full condition declined from the learning to the test phase is that this condition required participants to get used to unfamiliar (blind) stimuli, something that was clearly not the case in the blind

condition where conditions at test phase were identical to those during the learning phase. However, the half condition also resulted in a large change in the stimuli provided and they did not show a similar decline in performance, suggesting that the performance drop in the full condition was not due to the unfamiliar stimuli at test phase. Figure 4.6 also shows a convergence between the blind and half conditions from the learning phase to the test phase and an analysis confirmed that the distance travelled in the blind condition significantly decreased ($t(25) = 3.65$, $p < .05$) across phases. Somewhat surprisingly, however, the blind-half comparison at test phase only approached significance ($t(25) = 2.23$, $p > .05$), indicating just how high the level of performance was in the blind condition, in spite of the fact that the specific hotspot locations were never presented at learning or test phase.

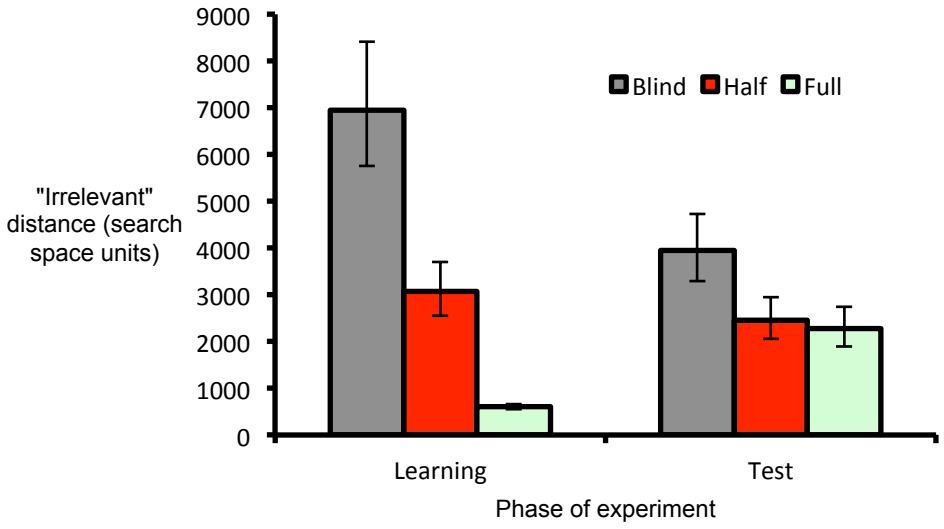

Figure 4.6 Mean irrelevant distance travelled (and standard error) for the 3 stimulus levels at both learning and test phases. Values are back-transformed from the log transformation.

Consistent inefficiencies

Whilst there was no attempt in the present study to investigate superstitious learning in the strict Skinnerian sense (that is to say, by providing non-contingent outcomes), it is possible to investigate this kind of learning in the weaker sense of participants learning consistent but inefficient movements. In these terms, all movement traces, apart from the optimum trace, lie on a 'superstition spectrum', containing a lesser or greater degree of unnecessary movement. In order to investigate this between conditions, an analysis was undertaken into performance in terms of the consistent behaviour achieved during the test phase of the experiment.

Over the course of the 25 trials during the test phase, participants produce a range of movements and often make missed attempts at the movement they intended, perhaps because of the blind conditions they are performing under. Consequently, the record of all movement traces creates a noisy representation of the core movement that the participant was settling on. In order to identify occasions where the sequence of movements remained stable over multiple trials, similar to that illustrated in figure 4.7, a consistency criterion was applied to the data. By identifying runs of trials during which the participant was able to maintain a consistent movement, it is possible to pick out traces that better represent the movement they had learnt or, as the case may be, the accuracy of their inference of the hotspot locations. The important thing about these traces is that they are occasions in which the participants have ceased to explore the search space and are instead performing a movement based on what they have learnt over previous trials – it is assumed that if the participants were still exploring the search space then they wouldn't be able to achieve this consistency.

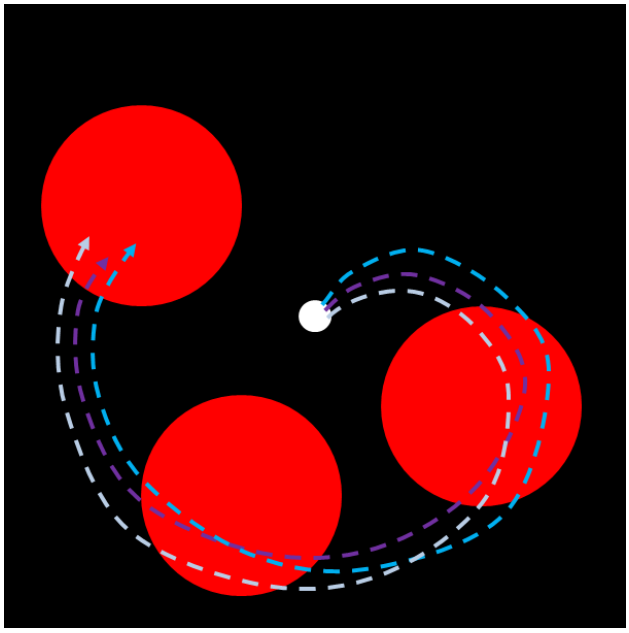

Figure 4.7 Example of the kind of attempts that the consistency criterion was designed to capture. The 3 attempts depicted in the diagram differ from one another but feature a consistent shape of movement.

Being able to identify situations in which performance is less representative of exploratory behaviour is important when making comparisons between the blind and half conditions. The blind condition necessarily requires more in the way of exploration

than the half condition because information about the particular arrangement of hotspots can only come from the behavioural variance that is produced. Consequently, for any given trial, it is not clear how much of the performance in the blind condition is due, on the one hand, to their, as yet, incomplete exploration of the search space and, on the other, to genuine limitations in the efficiency that can be achieved through the learning processes in operation. By identifying consistent trials, participants in both conditions can be compared in terms of an action they have settled on and the influence of exploratory behaviour is therefore reduced. If participants under both conditions are learning the movement spatially by simply moving the joystick to inferred (blind condition) or recalled (half condition) hotspot locations then the length of the consistent movement traces should not depend on how they learnt those hotspot positions. If, on the other hand, the blind condition had resulted in the participants learning a particular movement rather than inferring the positions of the hotspots, then we would expect to see less efficient – more 'superstitious' – movement traces.

Trials for all participants during the test phase were assessed for consistency, with 'consistency' being defined in terms of the difference in the distance travelled between 2 given trials. For each trial (excluding trials 1 and 2), this difference was calculated by comparing the current trial (T^0) with each of the 2 preceding trials (T^{-1} and T^{-2}), and also by comparing those preceding trials with one another. The sum of these differences was then calculated in order to generate a total difference score. A threshold was set at 0.5 times the mean trace length for the 3 trials; if the total difference score exceeded this threshold, then the trials were not classed as being consistent. A further constraint was applied which dictated that in order to be treated as consistent, trials could not be in excess of 4 times the optimal distance; this was to exclude occasions where consistency was achieved by virtue of extremely long trace lengths. For some individuals, the threshold was reached more than once within a testing session; in such situations, only the final instance was submitted to analysis. The data for 5 participants was removed from analysis as these people failed to reach the threshold in at least one of the 3 conditions. Figures 4.8, 4.9 and 4.10 show the movement traces for the consistent trials overlaid on top of one another.

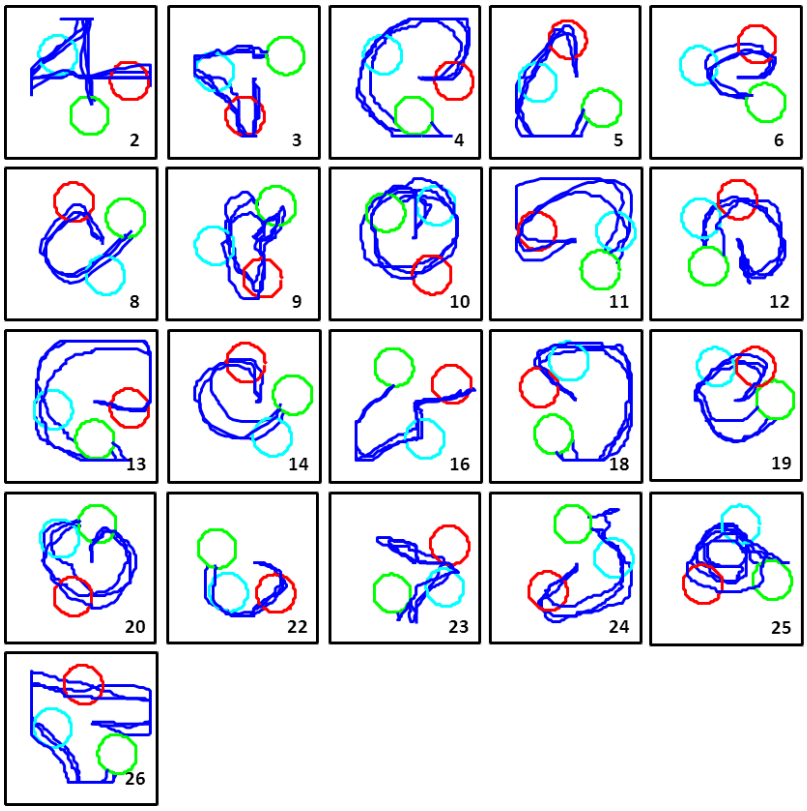

Figure 4.8 Combined movement traces for the 3 consistent trials at test phase in the blind condition.

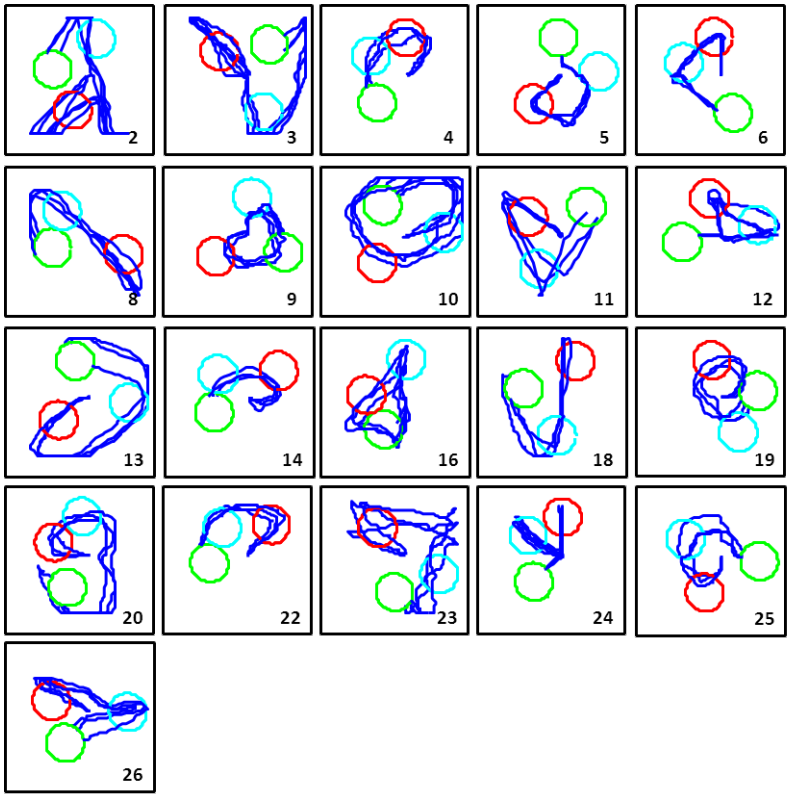

Figure 4.9 Combined movement traces for the 3 consistent trials at test phase in the half condition.

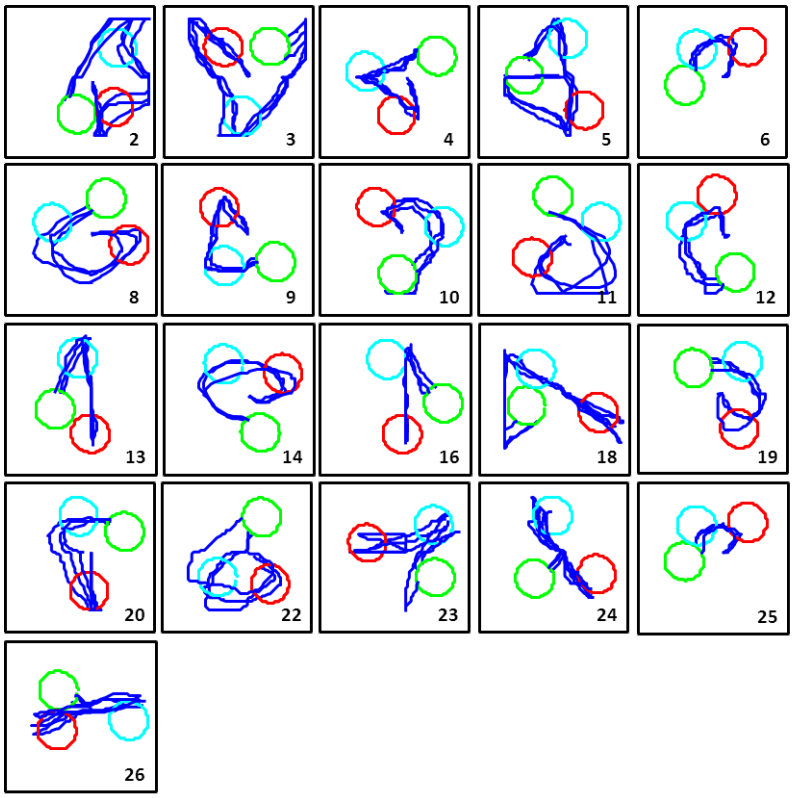

Figure 4.10 Combined movement traces for the 3 consistent trials at test phase in the full condition.

A one-way repeated-measures ANOVA was conducted to investigate the effect of stimulus condition on the mean irrelevant distance for the consistent trials during the test phase. The ANOVA of the log-transformed data revealed that there was a significant effect of stimulus condition on the mean irrelevant distance travelled during the consistent trials, $F(2, 40) = 5.5$, $p = .008$. As figure 4.11 illustrates, there was a substantial difference between the blind and the full conditions and this was confirmed by a Bonferroni corrected post hoc *t*-test ($t(20) = 3.53$, $p < .05$). However, the more relevant comparison between the blind and half conditions did not reach significance and would not have done so even without the conservative post hoc correction ($t(20) = 1.86$, $p > .05$). This result indicates that the blind condition did not force participants to settle on solutions that were significantly less optimal than the half condition and consequently gives no indication that this condition caused a different type of learning to occur to the half condition.

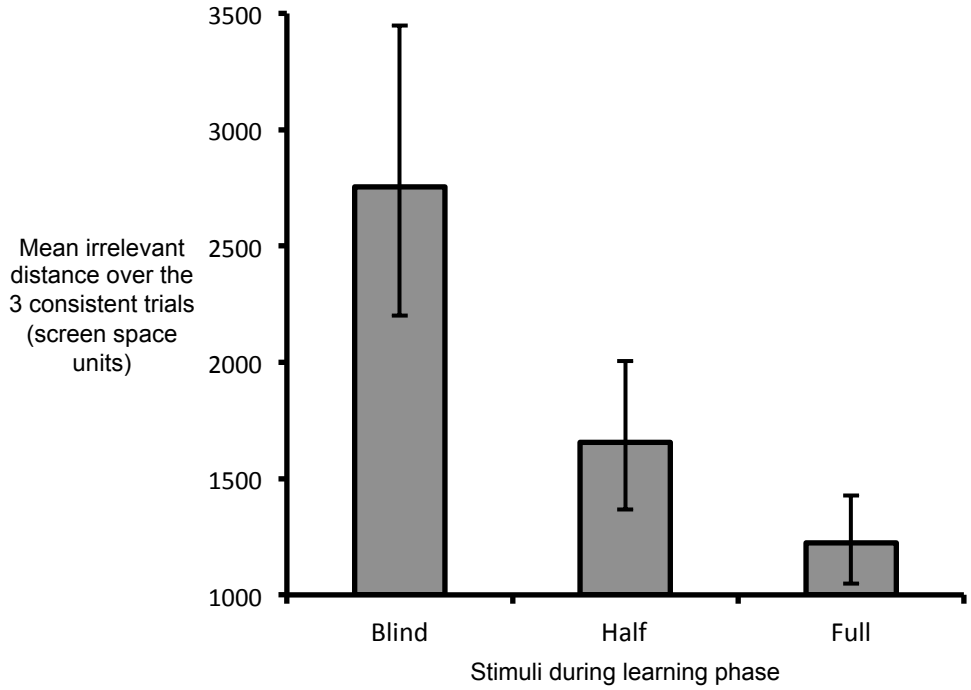

Figure 4.11 Mean irrelevant distance (and standard error) for the consistent trials at test phase. Values are back-transformed from the log-transformation.

Gesture phase of movement trace

The analyses described so far show that, at test phase, the blind and half conditions result in similar performance. However, the mean values appear to suggest that the blind group were generally less efficient at producing the required movement. One possibility is that performance was more similar than these data imply and that the results described so far largely reflect differences in the persistent need to explore the search space for the blind group, even after the 25 learning trials. In order to investigate performance for all trials with less emphasis on exploratory behaviour, an analysis was performed on the irrelevant distance travelled during the gesture phase of the movement trace (as defined in the methods section, also see figure 4.2). A two-way (2x3) repeated measures ANOVA (experiment phase by stimulus condition) was carried out on the log-transformed data. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of stimulus ($\varkappa^2$ (2) = 7.55, p < .05; therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.79). The analysis revealed that there was no main effect of phase (F(1, 25) = 3.52, p > .05), but that the main effect of stimuli (F(1.58, 39.37) = 23.2, p < .001) and the phase by stimuli interaction (F(2, 50) = 27.87, p < .001) were significant.

Bonferroni corrected post hoc comparisons revealed that the irrelevant gesture length for the blind and the half conditions differed significantly during the learning phase ($t(25) = 2.39$, $p < .05$), but not at test phase $t(25) = 1.35$, $p > .05$. Figure 4.12 shows the general convergence in gesture length between the learning and test phase. The greater similarity in the results at test phase for this metric appears to support the idea that much of the apparent difference in the blind condition for previous analyses was probably due to differences in exploratory behaviour as opposed to the action that had been acquired.

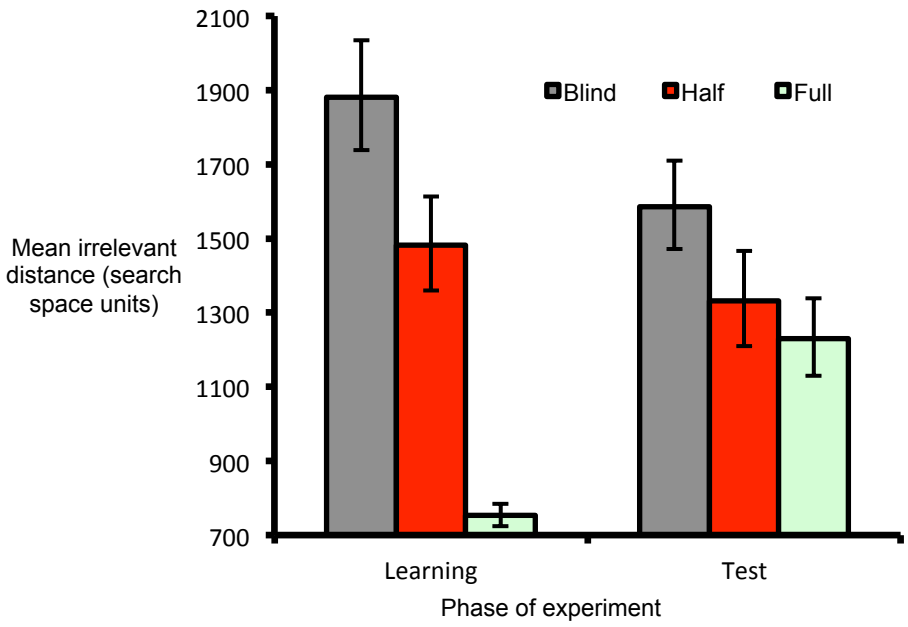

Figure 4.12 Mean irrelevant distance (and standard error) for the gesture phase of the movement trace only. Values are back-transformed from the log-transformation.

Early versus late phases of the gesture period

As indicated in the introduction, a possible symptom of learning through the reinforcement of motor output is that the variability in movements might differ depending on their proximity relative to the point of reinforcement. If the gesture is learnt based on the reinforcement of recent motor output, then it is possible that movements closest to the point of reinforcement will be less variable than those situated further from the point of reinforcement. If the gesture is inferred spatially, by contrast, there is no reason to expect one portion of the movement to vary more or less than any other. An important aspect of this prediction is that it doesn't depend on

absolute distances travelled and, consequently, the results are less readily explained in terms of, for example, the differences in task difficulty between the conditions.

The gesture portion of the movement trace for the test phase of the experiment was split into 2 further portions (phase 1 being furthest from the point of reinforcement and phase 2 being closest) as detailed in the methods section (see also figure 4.3). Following this, the standard deviation of the distances for both portions was calculated for each participant in both the blind and the half conditions (i.e. the standard deviation encompassed all 25 distances – one for each trial – in the test phase for each participant). A two-way (2 x 2) repeated measures ANOVA (gesture phase by stimuli) was performed on the log-transformed data. The analysis revealed no main effects of phase ($F(1, 25) = 2.2$, $p > .05$) or of stimuli ($F(1, 25) = 3.68$, $p > .05$) and no phase-stimuli interaction ($F(1, 25) = .000$, $p > .05$). Figure 4.13 shows that neither condition differed across the phases nor did the conditions differ from one another. In other words, contrary to the prediction, the variability in the length of the movement trace did not depend on its proximity to the point of reinforcement for either condition.

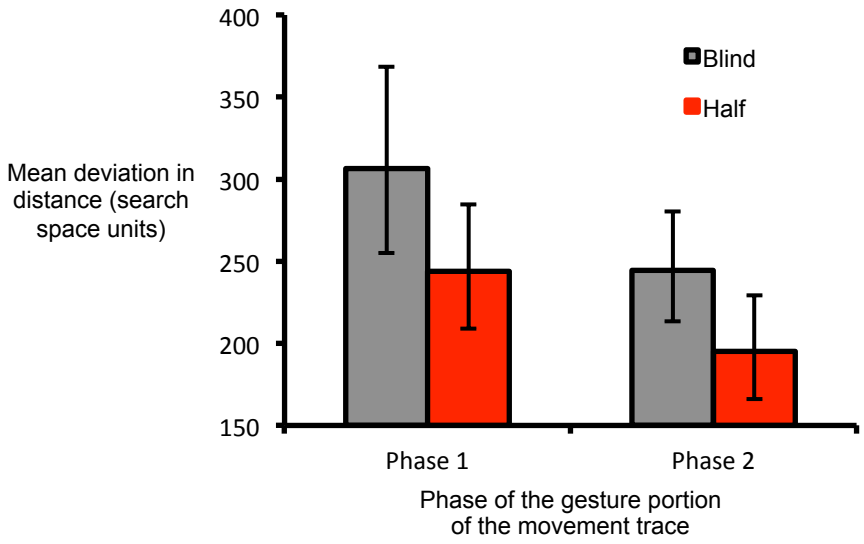

Figure 4.13 Mean trace length deviation (and standard error) for the two phases of the gesture portion of the movement trace. Phase 1 is the portion of the trace furthest from the point of reinforcement and phase 2 is the portion that is closest. Values are back-transformed from the log-transformation.

Whilst distance provides a good overall estimate of performance, it takes no account of the relative position in space of the portion of movement trace under investigation.

It is always possible that small differences in distance represent relatively substantial differences in the shape of movement. Consequently, the deviations in distance shown in figure 4.13 might not fairly represent the variations in the gesture that the participants were attempting to perform. In order to take into account something of the shape of the movement, the point of entry into the first and the final hotspots was extracted from the data and the deviation in the point of entry was then calculated. The prediction is the same as for the 2 portions of the gesture period: the deviation for the hotspot closest to the point of reinforcement will be lower than that for the hotspot further away. There is also the further possibility that the overall variation in the point of entry for the blind condition might be smaller than that for the half condition irrespective of the proximity to the point of reinforcement. This is because, gestures learnt as a set of movements that are stable relative to the hotspots cannot vary greatly because each point of entry is dependent on the last. By contrast, if the position of the hotspots is spatially inferred, the point of entry into a given hotspot is less important because the position of the next hotspot will not be determined based on previous movements.

In order to calculate the deviation scores, an arbitrarily determined stable point was first set on the perimeter of each hotspot. Following this, a deviation distance was calculated by determining the distance of the point of entry from the stable point around the perimeter of the hotspot. The circumference of each hotspot was 1037 screen units so the maximum deviation distance was half this, at 518.5 units. The standard deviation of all deviation distances was then calculated for each participant in both the blind and the half conditions at test phase (i.e., once again, the standard deviation described the variation of all 25 deviation distances for each participant in both conditions). These variance scores for the point of entry were then submitted to analysis.

A two-way (2 x 2) repeated measure ANOVA (hotspot number by stimuli) was conducted to investigate the effect of the position of the hotspot relative to the point of reinforcement and the effect of the stimulus condition on the deviation in point of entry. The analysis revealed that there was no main effect of hotspot ($F(1, 25) = 0.04$, $p > .05$) or of stimuli ($F(1, 25) = 0.67$, $p > .05$) and no hotspot-stimuli interaction ($F(1, 25)$

= 0.96, $p > .05$). Figure 4.14 shows that the consistency of the point of entry into the hotspots was not affected by either the proximity of that hotspot to the point of reinforcement or the stimulus conditions experienced during the learning phase.

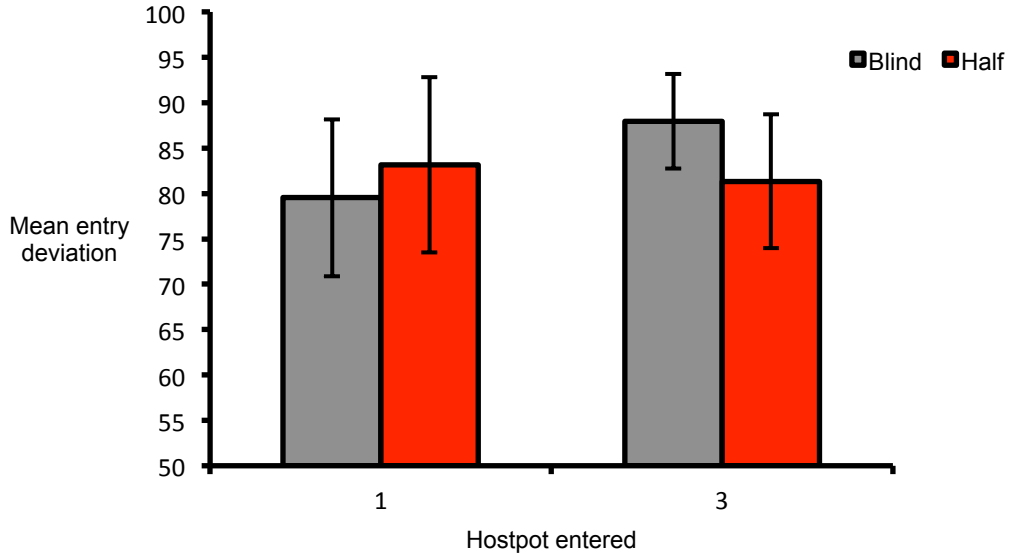

Figure 4.14 Mean point of entry deviation (and standard error) for the first and the final hotspot in the gesture. The maximum possible deviation on a given trial is 518.5 units.

Procrustes analysis

In order to consider the shape of movement on its own, a Generalised Procrustes Analysis (GPA; Gower, 1975) was conducted on the data for the blind condition. With this technique it is possible to minimise the influence of the size, orientation and the position in space of the movement trace because the analysis first scales, rotates and then transforms the coordinates of each configuration such that they are superimposed before they are compared. All data were first resampled to ensure that each movement trace to be analysed consisted of 1000 landmarks (coordinates). A mean movement configuration was generated from Procrustes adjusted movement traces for all trials of each participant's data. Following this, each of the adjusted traces was compared to the mean trace for that participant with the root mean square residual providing a score for the variability of the trace from the mean trace (Miall, Leschziner, Miall & Stein, 1997).

A paired-sample *t*-test was used to compare the mean variability for all participants

during the learning phase with the mean variability at test phase. This revealed that there was a significant difference in the variability of the movement traces produced in the learning and test phases, $t(25) = 3.55$, $p = .002$, two-tailed. Figure 4.15 shows that there was a substantial reduction in variability from the learning to the test phase indicating that movement conformed more to an underlying shape of movement with practice at the task.

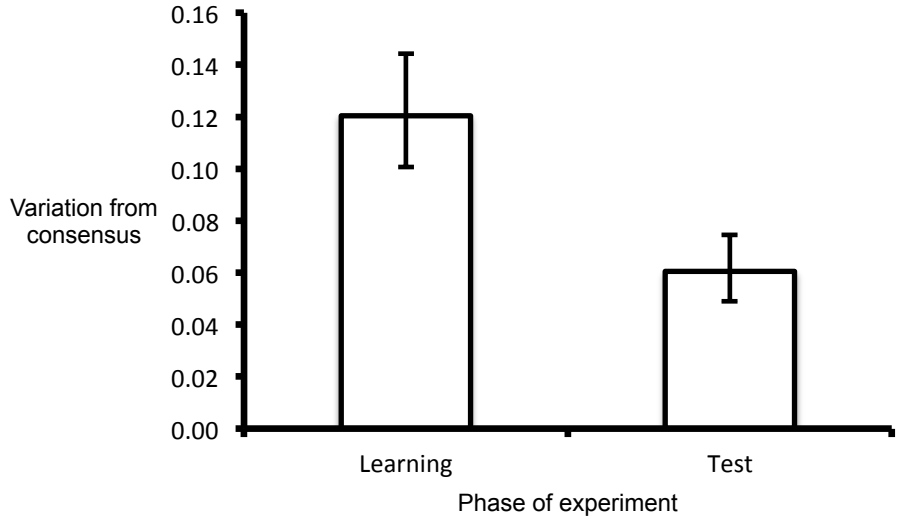

Figure 4.15 Mean GPA determined variability (and standard error) at learning and test phase for the blind condition.

Discussion

The current task involved the learning of patterns of movement in human participants under different stimulus conditions. Using this paradigm it was possible to find and document evidence of pattern learning and show that the ability to generate these movements with no visual feedback was similar irrespective of the visual conditions under which the pattern was originally learned.

As indicated in the introduction, the experiment was designed to look at 4 main questions, all related to the general issue of whether action acquisition is influenced by the reinforcement of recent motor output when an alternative means of acquisition is available in the former of high-level cognitive/spatial guidance of movements. Comparisons between the full and the half conditions revealed that when the shape of the gesture is known, the type of practice has no effect on the ability of people to

perform that gesture when those locations are subsequently hidden. Both conditions provided full visual information about the positions of the hotspots and both required participants to discover the correct order by exploring the search space. However, the conditions differed markedly in the amount of behavioural variance that was produced when navigating between these hotspots. The full condition allowed participants to move between hotspots in a near optimal fashion during the learning phase, whereas the half condition resulted in a greater quantity of irrelevant movements. And yet, despite this, performance at test phase did not differ between the 2 conditions, indicating that the different movements practiced had little impact on learning.

The comparisons between the half and blind conditions revealed that when information is provided about the general form of the gesture, exploratory behaviour results in no differences in the ability of people to perform that gesture under blind conditions. Both conditions provided full information about the general form that the gesture would take. However, once again, the conditions differed in terms of the amount of behavioural variance that was produced in navigating between these hotspots. The half condition provided information about the specific locations of the hotspots so no exploratory behaviour was required to determine their locations during the learning phase, whereas the blind condition required participants to learn about the gesture by exploring the search space resulting in a greater quantity of irrelevant movements during learning. Once again, performance at test phase did not reflect the differences in the movements practiced during the learning phase.

The general finding that movements produced during the learning phase resulted in no differences in performance during the blind test phase was reflected in further, more specific analyses. The blind and the half groups did not differ in terms of the variance in movements relative to the point of reinforcement. Nor did these conditions differ in terms of the efficiency of the consistent movements acquired. In other words, these results also provided no reason to believe that learning was influenced to any large degree by the movements practiced.

It was, however, possible to learn whether or not there was a basic improvement in the consistency of movement from the learning to the test phase. A Procrustes

analysis of the data for the blind condition indicated that there was a significant decrease in the variability of movement traces around an underlying mean movement trace derived from all traces. In other words, the movement traces were more similar in shape to one another during the test phase than the learning phase. This confirms a valuable feature of the paradigm to measure to ability of people to learn and refine self-generated patterns of movement without the ability to visually monitor movements relative to a target.

The Procrustes analysis also proved to be a potentially better way of detecting improvements across phases than the other performance metric described in the results section. Whilst the distance during the gesture period for the blind condition was found to decrease (indicating better performance) across phases of the experiment ($t(25) = 2.81$, $p = .01$), the same comparison based on the Procrustes derived metric of variability proved marginally more sensitive to the change ($t(25) = 3.55$, $p = .002$). In general the Procrustes analysis probably offers a better description of performance because performance is defined in terms of consistency rather than how close the distance travelled is to the optimum distance. This is important because the optimum route between hotspots is unlikely to align well with the natural dynamics of movements that follow a more rounded path.

The theory of action acquisition proposed by Redgrave and Gurney (2006) and Redgrave et al. (2008) differs from some traditional accounts of skill and action acquisition. Their proposal is that circuitry in the midbrain and basal ganglia are actively involved in the early stages of action acquisition from the detection of agency to the reselection of temporally contiguous motor output and the filtering of task relevant features from behavioural variance. This view is in line with a large amount of evidence that the basal ganglia are contribute heavily during the early stages of action acquisition (Brasted & Wise, 2004; Costa, 2007; Costa, Cohen, Nicolelis & Carolina, 2004; Ungerleider, 2002). The perspective differs from some existing accounts in that it is generally assumed that these areas of the brain, and indiscriminate processes of reinforcement in general, are associated with more gradual and habitual learning (Bayley et al., 2005; Buffalo, Stefanacci, Squire & Zola, 1998; Squire, 2004). In traditional accounts of skill acquisition this amounts to the gradual automatisation

phase, characteristic of the final stages of skill learning in humans (Anderson, 1982; Fitts & Posner, 1967; Newell & Rosenbloom, 1981).

In the context of the current experiment these different perspectives on action acquisition call for different interpretations of the findings. The traditional account describes the learning process as going through a series of stages that start with the initially rapid improvement in performance typical of a declarative phase of learning (Anderson, 1982). From this point of view, the learning phase in the current experiment (i.e. the first 25 trials) would be part of the rapid declarative phase of learning and there would be no reason to expect much contribution from the reinforcement of motor output; such low-level effects would be expected further down the line, once the action had been repeated sufficiently often that it starts to become automatic. Consequently, the finding that, for example, the 'full' condition in the current experiment showed a decline in performance between the learning and test phase would support the idea that performance was being guided at a declarative level. According to Redgrave and Gurney's account, by contrast, we would expect the learning mechanism to influence mostly the early phase of action acquisition, when the link between the movements, context and reinforcing stimulus remains novel. According to this account, it is somewhat surprising that the highly efficient movements produced during the learning phase in the full condition did not transfer to the test phase.

In fact, Redgrave and Gurney (2006) allude to a transition between learning mediated by the phasic activity of dopamine neurons, that results in the discovery of novel actions, and learning that arises from the repetition of the newly discovered action, through "traditional reinforcement learning mechanisms" (p.972). It is possible, therefore, that there is a major distinction between these two types of reinforcement learning, with the former causing a tendency simply to reselect and repeat important recent motor output and the latter playing more of a role in the reinforcement and storage of that motor output. Such a distinction would have important implications for how we view the joystick task in general and the results of the experiments in chapters 2 and 3.

The findings from the current experiment suggest that we cannot assume that reinforcement learning is responsible for the effects uncovered in the previous experiments based solely on the fact that feedback was presented in a way that is in accordance with this type of learning. Indeed, there were indications from the experiment featuring between-trial repetition of reinforcement (experiment 4) that learning may have been under the influence of higher-level declarative systems. The effect of delay, therefore, could have been due to the declarative system being fooled by the different positions in space that resulted from delayed reinforcement as opposed to a low level learning mechanism stamping in non-contingent, yet contiguous, motor output. The effect of delay in experiment 4 was similar to that found in other investigations of reinforcement learning (e.g. Shanks et al., 1989; Shanks & Dickinson, 1991), though very different to the effect found in experiments 1 and 2 and this difference was interpreted as possibly being caused by the rate at which reinforcement was delivered. However, an explanation in terms of different learning systems might make more sense. Buehner and May (2004) were able to demonstrate, for example, that the results concerning delay reported by Shanks et al. (1989) could be abolished by a small change in the task instructions. This finding therefore implies that the results concerning delay in the Shanks study may have little to do with the operation of simple associative learning mechanisms and much more to do with participants being fooled at a declarative level.

The sensitivity to delay found in experiments 1 and 2 is not so readily explained in terms of learning at the declarative level. As discussed in chapter 2, the sensitivity was in line with what we would expect based on Redgrave and Gurney's (2006) theory of the function of the phasic activity of dopamine neurons. This interpretation is further supported by the finding that the basal ganglia are much faster to respond to reinforcement contingencies than the prefrontal cortex (Pasupathy & Miller, 2005) and furthermore that prefrontal systems and model based learning in general require more time to operate due to a greater level of computational complexity (Bogacz, McClure, Li, Cohen & Montague, 2007; Daw, Niv & Dayan, 2005). It is therefore possible that participants in experiments 1 and 2 were forced to rely on lower level learning mechanisms because the time between instances of reinforcement and adjustments of hand position didn't afford them the opportunity to process their responses using

more computationally sophisticated brain processes. By contrast, experiment 4 allowed participants a period of approximately 5 s between one attempt and the next and thus gave them time to make the subsequent response based on a deeper level of information processing. The emphasis that the task in experiments 1 and 2 placed on the reselection and repetition of movements rather than on the more gradual refinement of behavioural variance, might therefore make it a more effective means of investigating the mechanism of action acquisition described by Redgrave and Gurney.

Results are in line with previous evidence showing that humans find it hard to use their nondeclarative systems

The opportunity for participants to demonstrate declarative learning was deliberately left open in the experiment 5 in order to test the general tendency of people to reinforce the movements they make rather than simply infer correct movements. However, it is possible that even if the crucial aspects of the task had been made opaque to higher-level processes, the task may have uncovered little in the way of nondeclarative learning (i.e. just an overall decline in learning).

The best evidence on this issue, either way, is to be found in situations where people are forced to rely on nondeclarative systems as is sometimes the case with instances of brain damage and, here, the evidence is mixed. Buffalo et al. (1998), for instance, have found that medial temporal lobe lesions (lesions designed to impair declarative memory) have no effect on the ability of macaques to learn a concurrent discrimination task: the monkeys learn this task gradually and apparently habitually whether they are lesioned or not. Similar brain damage in humans has a devastating impact on learning, but some capacity for learning is retained and, importantly, the rate of acquisition is comparable to the gradual habitual acquisition that is the norm in monkeys (Bayley et al., 2005). Furthermore the brain damaged participants are able to demonstrate none of the flexibility that one would expect had they learned the task declaratively, indicating that they had instead learnt the task habitually and in a way very similar to the monkeys. In other words, Bayley et al. (2005) were able find evidence of nondeclarative memory in humans and a similar propensity to learn nondeclaratively as nonhuman animals.

A somewhat different pattern of results is found in blindsight research. Following damage to the primary visual cortex, both humans and monkeys can be prompted to use their preserved visual abilities in laboratory tests, and the performance of both improves over time. However, despite the fact that both humans and monkeys display evidence of unconscious perception, monkeys appear better able to make use of their preserved abilities outside of heavily cued testing sessions, where responses are not prompted and behaviour is more naturalistic (Allen-Hermanson, 2010; Humphrey, 1995; Humphrey, 2000; Stoerig & Cowey, 1997). In other words, when it comes to having the confidence to rely on nondeclarative processes to navigate and respond to the environment, monkeys, in spite of their very similar visual systems, are apparently better able to rely on nondeclarative processing. Though, it is clearly true that such differences are marginal and only likely to appear under special circumstances. Both humans and nonhuman animals are adept at utilising nondeclarative learning and control across all areas of control.

Not only do humans have a highly developed ability to process information declaratively and, arguably, a greater tendency to rely on this type of information processing, but the range of behaviour that is monitored at the declarative level is surprising large. The Chevreul Pendulum Illusion (Easton and Shor, 1975), for example, is a situation where human participants are fooled by an apparent inability to monitor movements consciously. The illusion is created by asking a participant to use their hand to suspend an object, such as a wedding ring, on the end of a piece of thread. When participants are then asked to focus on the object, the object will often begin to oscillate despite the participant having no sensation that they are causing the movements. The muscle movements powering the pendulum are apparently so small that they are simply not registered and thus the participant is left with the sensation that the pendulum is powered through some external means. What is surprising about this illusion is that it suggests that the judgements of causality necessary to ascribe agency to oneself are heavily reliant on conscious awareness and, to some extent, external monitoring of muscle movements. Split brain animals and patients also demonstrate how important the external monitoring of behaviour can be in the form of cross-cueing integration of information between the hemispheres (Jakobson,

Servos, Goodale & Lassonde, 1994; Levy, Trevarthen & Sperry, 1972; Savaki, Kennedy, Sokoloff & Mishkin, 1993). Put simply, one side of the brain finds out what the other side is doing or thinking by witnessing the body movements it produces or by listening to the verbalisation of behaviour. This process of monitoring is so efficient that in spite of the severing of the corpus callosum, it is possible for humans and monkeys to behave in a way that, outside of the laboratory, is almost indistinguishable from normal behaviour. In other words, there are signs that the declarative monitoring and guidance of behaviour is more than just a computationally expensive veneer overlaying a largely nondeclarative and computationally cheap set of learning and control processes. Rather, we might be incorrigibly reliant on declarative level processing, in spite of our limited attentional resources even for relatively simple guidance of motor behaviour.

The declarative threshold

Pursuing the issue of reselection and repetition of responses in isolation to reinforcement is one possible future avenue for research, especially as regards Redgrave and Gurney's (2006) theory. However, the nondeclarative acquisition of actions through the reinforcement of motor output is, nonetheless, still an important issue. If action acquisition is dependent on the neural learning mechanism detailed by Redgrave and Gurney (2006), then we must assume that the type of learning displayed by the animals in Thorndike's (1911) original studies was largely a product of this mechanism. Indeed, the data and descriptions of behaviour provided by Thorndike are in line with what one would expect from an indiscriminate learning mechanism with no additional declarative guidance. On the first trial inside a puzzle box, a cat's performance of the required action is an incidental consequence of the general escape behaviour it produces on being place in the enclosure. In other words, it performs the behaviour by accident. But, unlike the participants in the current study, so too does it complete the second trial by accident: perhaps slower, perhaps quicker, but the cat remains apparently clueless as to what needs to be done. The process continues and the trend is towards more efficient performance of the necessary action over subsequent trials and yet the cat remains in the dark as to how it is achieving the desired outcome. Thorndike persuades us of this by observing that trained cats

displayed irrelevant behaviour such as trying to perform the required action even when the door to the puzzle box was open or performing the action when the operandum had been removed. Attempts to teach the cats or allowing them to see other animals completing the action had no effect on performance. Furthermore, in most of the puzzle boxes, learning did not shift abruptly from poor performance to efficient performance; rather there was a gradual trend towards an increase in efficiency. As Thorndike notes, this was never a smooth curve and there was no guarantee that a given trial would be completed more quickly than the trial that preceded it. In other words, the cats displayed behaviour with none of the properties of insightful learning (Bayern et al., 2009). In a sense then, it was as if the cats were escaping from the puzzle box by accident on the first attempt and escaping by accident on the final attempt, the only thing that had changed was the speed with which the accident was performed.

The question Thorndike's (1911) study raises for the future of research into action acquisition in humans is whether there might be any conditions under which people acquire actions in a similar way to Thorndike's nonhuman animal subjects. Dekeyser (2007) raises the interesting concept of a declarative threshold that must be crossed during the process of acquiring a new skill. Most skills, Dekeyser observes, are learnt with a distinctive rate of acquisition, described by a power law learning curve: initially performance improves rapidly, but the rate of improvement then declines and follows a shallow slope of improvement for all subsequent practice. Dekeyser notes that the two distinct portions of the learning curve, as defined by their different slopes, are generally thought to represent the different processes that are contributing to learning at different points during skill acquisition. Early on, the skill is approached declaratively, with the learner developing a structural understanding of the task requirements. Once this phase is over, the learner has, in Dekeyser's words, crossed a declarative threshold and learning continues through a 'automatisation' phase during which the rate of improvement is lower and the contribution of declarative processes is diminished.

Dekeyser's use of the concept of a declarative threshold is potentially useful for identifying situations in which learning occurs nondeclaratively. However, his use of

the term is not how we would normally understand a threshold to work: in Dekeyser's sense, learning ceases to be declarative when the threshold is exceeded. If we instead consider this concept from the opposite point of view, it becomes much more useful to the present purposes. Exceeding the declarative threshold in the context of the current research might be thought of as any instance in which an aspect of learning becomes available to awareness and therefore to the powerful higher level cognitive processing abilities of the human participant. The question is: can we identify, or contrive to produce, any instances of action acquisition that do not exceed this revised definition of the declarative threshold? Doing so would provide a potential means of investigating the low level learning processes central to the current research.

The role that a tutor can play in helping people to learn skills such as juggling, gives us a clue as to how far declarative learning processes can take us in learning to perform such actions. The aspects that a tutor cannot convey – the components of an action that only come from practice – are, by contrast, good candidates for nondeclarative learning. Juggling is an excellent example of an action with the frustrating property that when one follows the guidance, it still seems like there is an instruction missing. The basic combination of actions required is simple and yet, in spite of the apparent simplicity of the required sequence of movements, there is a component of the action that one cannot execute through an understanding of the task structure alone. Some aspects of such skills are presumably, too fast, too fine or simply too far outside of our frame of reference to be perceived and/or executed declaratively. Instead, one must simply practice the skill and generate the behavioural variance from which the important aspects of that variance can be gradually extracted. This frustrating component of skill learning is popularly referred to as the 'knack' of the skill and such knacks might provide a way of investigating the contribution of low-level, nondeclarative learning mechanisms to the process of action acquisition.

Chapter 5: summary and going forward

There is currently no methodological standard for the investigation of action acquisition in humans in the way as there is, for example, for the investigation of reward learning in nonhuman animals. For many reasons, this situation is unlikely to change any time soon. The scope of the subject of action acquisition is extremely broad, encompassing multiple learning processes and types of feedback. The current work has described the development of a novel behavioural paradigm for the investigation of action acquisition in humans. From the outset, the aim has been to constrain this broad research question by focusing the theoretical emphasis on a low-level learning mechanism detailed by Redgrave and Gurney (2006) and Redgrave et al. (2008). However, even thus constrained, the development was far from straightforward.

The prospects for the joystick task as a means of investigating action acquisition are mixed. It provides many methodological advantages over button-press reinforcement learning paradigms; however, it suffers, as do so many other methodological techniques, in that it is not straightforward to separate the effects of declarative level processing from low-level reinforcement learning mechanisms. What is perhaps most important is that it proved extremely difficult to demonstrate the specific influence of behavioural variance on learning – a key element of the theory of action acquisition under investigation.

However, whilst the reinforcement of behavioural variance remains a difficult phenomenon to investigate with the joystick task, as discussed in chapter 4, it seems likely that this means of measuring the influence of Redgrave and Gurney's action acquisition mechanism may not ultimately prove to be the most efficient approach. By focussing our intention instead on the repetition and reselection as opposed to the reinforcement of motor output, we may be able to gain more insight into how such a mechanism might contribute to the earliest stages of action acquisition. In this regard, the version of the task described in chapter 2 is likely to provide a useful starting point from which to approach future research.

Going forward, there are countless areas of research open to investigation. Here, two areas of potential inquiry are identified. The first concerns another test of Redgrave and Gurney's (2006) theory in a patient population and potentially utilising the joystick task paradigm. The second concerns behavioural variance and would likely require an alternative methodological approach.

Parkinson's disease

As already discussed, Redgrave and Gurney's (2006) theory of action acquisition concerns the activity of neurons in the ventral midbrain, including the substantia nigra. Parkinson's disease is characterised by a loss of cells in the substantia nigra (Pavese & Brooks, 2009), which is a major site of dopamine neuron cell bodies (Schultz, 1998). An idea put forward by Redgrave and Gurney within this author's research group, is that this implies that people suffering from Parkinson's disease should be less able to learn associations between novel stimuli and their own motor output. Consequently, they should find it particularly difficult to add new actions to their behavioural repertoire (Redgrave et al., 2008; Redgrave, et al., 2010). Interestingly, another prediction arising from the theory is that for patients who are being treated with dopamine agonists, the general therapeutic effect of the medication should not include an improvement in the ability to learn novel actions. This is because the phasic release of dopamine (as opposed to the tonic levels) would be relatively unaffected by the treatment and it is the phasic release of dopamine that is thought to have the time-stamp effect that is crucial to learning.

In order to test the potential of the joystick task as a means of investigating a deficit in action acquisition in this population, a pilot study was run on one male (66 year-old) participant suffering from Parkinson's disease (fully medicated at the time of testing). The task employed was a modified version of experiment 3 and it was carried out mainly to assess the ability Parkinson's disease sufferers to perform the joystick task. Somewhat surprisingly, given the prediction, performance on the task was excellent for hotspot sizes up to the smallest tested – 0.48% of the search space – and certainly in the general range found for undergraduate participants in experiment 3 (M = 1838.61, SD = 1659.60, values in untransformed screen units). Going forward, there is

value in running further experiments with people suffering from Parkinson's disease, using different versions of the joystick task. Such investigations would not only enable the testing of the general hypothesis outlined above but may also prove informative of the paradigm itself.

Imitation learning and the generation of behavioural variance

The emphasis of Thorndike's (1911) research was very much on the nondeclarative end of the learning spectrum: on the accidental nature of learning. In his view, the animals in the puzzle-boxes were utterly at the mercy of the environment they were placed into: "If all cats, when hungry and in a small box, will accidentally push the button that holds the door, an occasional cat in a large room may very well do the same" (p.73). Given a similar puzzle-box task to Thorndike's cats, humans could bring a powerful declarative learning system to bear on the problem and, provided they could locate an operandum, would not be unduly affected by parameters such as the size of the room they were placed in. However, this is not to say that there can be no room for indiscriminate behaviour, even in the sophisticated learning of humans. As has already been argued, in the case of knacks, exposing our learning system to samples of essentially accidental behavioural variance might be the only way that we can learn to reliably elicit a desired outcome because we are forced by our very ignorance of the paradigm to explore portions of movement space we have no good reason to think will be of use.

We might accept that there are advantages to be gained from indiscriminate learning; however, this raises the question of how a learning agent can generate behavioural variance from which commonalities can be drawn in the first place. Given that we are not prone to motor babbling outside of the first few months of life and that knack learning makes up just a small proportion of most of the learning we produce, there doesn't seem to be much scope for producing behaviour which doesn't have a direct purpose. However, the question of the origins of behavioural variance is an important one and as Staddon and Niv (2008) have noted:

> It is something of a historical curiosity that almost all operant-conditioning research has been focused on the strengthening effect of reinforcement and almost none on the question of origins, where the behavior comes from in the first place, the problem of behavioral variation, to pursue the Darwinian analogy.

One possible source of behavioural variation in humans, which is more naturalistic than the kind of forced behavioural variance produced during the exploratory period of the joystick task, is imitation. In spite of the obvious constraints that imitation implies, it can result in the generation of behaviour that is at odds with anything we might produce with a strict goal-directed approach to the situation. There is considerable evidence that both children and adult humans imitate causally irrelevant aspects of behaviour (Lyons, Young, & Keil, 2007; McGuigan, Makinson & Whiten, 2011; McGuigan, Whiten, Flynn, & Horner, 2007), even behaviour that chimpanzees will not imitate (Horner & Whiten, 2005). A potential advantage of imitating aspects of behaviour that are apparently causally irrelevant (and potentially inefficient) is the possibility that these components of the behaviour have some hidden value. For example, by copying the style of a particular tennis player, a novice is likely to gain substantial advantages from principles of body-shape, follow-through and small-step movements without necessarily understanding the relevance of these behaviours. Imitation learning is therefore a potentially fruitful avenue of research for investigating how learning is derived from behavioural variance under normal circumstances.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B, 34b*(2), 77-98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology, 33B*(2), 109–121.

Aeschleman, S. R., Rosen, C. C., & Williams, M. R. (2003). The effect of non-contingent negative and positive reinforcement operations on the acquisition of superstitious behaviors. *Behavioural Processes, 61*(1-2)*,* 37-45.

Allen-hermanson, S. (2010). Blindsight in monkeys, lost and (perhaps) found. *Journal of Consciousness Studies, 17*(1-2)*,* 47–71.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4)*,* 369.

Anderson, K. G., & Elcoro, M. (2007). Response acquisition with delayed reinforcement in Lewis and Fischer 344 rats. *Behavioural Processes, 74*(3)*,* 311-8.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 932.

Apicella, P., Legallet, E., & Trouche, E. (1997). Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Experimental Brain Research, 116*(3)*,* 456-66.

Asmus, F., Huber, H., Gasser, T., & Schöls, L. (2008). Kick and rush: paradoxical kinesia in Parkinson disease. *Neurology, 71*(9), 695.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10)*,* 829-39.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37*(4-5), 407-19.

Bard, C., Turrell, Y., Fleury, M., Teasdale, N., Lamarre, Y., & Martin, O. (1999). Deafferentation and pointing with visual double-step perturbations. Experimental brain research. *Experimentelle Hirnforschung. Expérimentation Cérébrale, 125*(4)*,* 410-6.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics, 13*(5), 834-846.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47*(1)*,* 129-41.

Bayern, A. M. P. von, Heathcote, R. J. P., Rutz, C., & Kacelnik, A. (2009). The role of experience in problem solving and innovative tool use in crows. *Current Biology:, 19*(22)*,* 1965-8.

Bayley, P. J., Frascino, J. C., & Squire, L. R. (2005). Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature, 436*(7050)*,* 550-3.

Bernstein, N. (1967). *The co-ordination and regulation of movements.* Oxford: Pergamon.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28*(3)*,* 309-69.

Berthouze, L., & Lungarella, M. (2004). Motor skill acquisition under environmental perturbations: on the necessity of alternate freezing and freeing of degrees of freedom*. Adaptive Behavior, 12*(1)*,* 47-64.

Black, J., Belluzzi, J. D., & Stein, L. (1985). Reinforcement delay of one second severely impairs acquisition of brain self-stimulation. *Brain Research, 359*(1-2)*,* 113-9.

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning*. Brain Research, 1153*, 111-21.

Braun, D., Aertsen, A., Wolpert, D. M., & Mehring, C. (2009). Motor task variation induces structural learning. *Current Biology, 19*(4)*,* 352-7.

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *The Quarterly Journal of Experimental Psychology, 57B*(2), 179-191.

Brasted, P. J., & Wise, S. P. (2004). Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. *European Journal of Neuroscience, 19*, 721-740.

Buffalo, E. A., Stefanacci, L., Squire, L. R., & Zola, S. M. (1998). A re-examination of the

concurrent discrimination learning task: the importance of anterior inferotemporal cortex, area TE. *Behavioral Neuroscience, 112*(1)*,* 3-14.

Cardinal, R. N. (2006). Neural systems implicated in delayed and probabilistic reinforcement*. Neural Networks, 19,* 1277-1301.

Collins, S. H., & Ruina, A. (2001). A three-dimensional walking robot with two legs and knees. *The International Journal of Robotics Research, 20*(7)*,* 607-615.

Colwill, R. M., & Rescorla, R. A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes, 11*(1), 120-132.

Colwill, R. M., & Rescorla, R. A. (1988). The role of response-reinforcer associations increases throughout extended instrumental training. *Learning & Behavior, 16*(1)*,* 105–111.

Comoli, E., Coizet, V., Boyes, J., Bolam, J. P., Canteras, N. S., Quirk, R. H., . . . Redgrave, P. (2003). A direct projection from superior colliculus to substantia nigra for detecting salient visual events*. Nature Neuroscience, 6*(9)*,* 974-80.

Costa, R. M. (2007). Plastic corticostriatal circuits for action learning: what's dopamine got to do with it? *Annals of the New York Academy of Sciences, 1104*, 172-91.

Costa, R. M., Cohen, D., Nicolelis, M. A. L., & Carolina, N. (2004). Differential corticostriatal plasticity during fast and slow motor skill learning in mice. *Current Biology, 14*, 1124-1134.

Daroff, R. B. (2008). Paradoxical kinesia. *Movement Disorders, 23*(8), 1193.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12)*,* 1704-11.

DeKeyser, R. (2007). Skill Acquisition Theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97-113)*.* New Jersey: Lawrence Erlbaum Associates.

Desmurget, M., & Grafton, S. (2000). Forward modelling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences, 4*(11)*,* 423-431.

de Wit, S., Barker, R. A., Dickinson, A. D., & Cools, R. (2011). Habitual versus goal-directed action control in Parkinson disease. *Journal of Cognitive Neuroscience, 23*(5)*,* 1218-29.

de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., & Fletcher, P. C. (2009).

Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *The Journal of Neuroscience, 29*(36), 11330-8.

de Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(1), 1-11.

D'Hooge, R., & De Deyn, P. P. (2001). Applications of the Morris water maze in the study of learning and memory. *Brain Research Reviews, 36*(1), 60-90.

Dickinson, A. (2001). The 28th Bartlett Memorial Lecture Causal learning: An associative analysis. *The Quarterly Journal of Experimental Psychology B, 54*(1), 3-25.

Dickinson, A., & Balleine, B. W. (2000). Causal cognition and goal-directed action. In C. M. Heyes & L. Huber (Eds.), *The evolution of cognition* (pp. 185-204). Cambridge: The MIT Press.

Dickinson, A., Balleine, B., & Watt, A. (1995). Motivational control after extended. *Animal Learning & Behavior, 23*(2), 197-206.

Dickinson, A., Watt, A., & Griffiths, W. (1992). Free-operant acquisition with delayed reinforcement. *The Quarterly Journal of Experimental Psychology Section B, 45*(3), 241–258.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., . . . Redgrave, P. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science, 307*(5714), 1476-9.

Dowd, E., Monville, C., Torres, E. M., Wong, L.-F., Azzouz, M., Mazarakis, N. D., & Dunnett, S. B. (2005). Lentivector-mediated delivery of GDNF protects complex motor functions relevant to human Parkinsonism in a rat lesion model. *The European Journal of Neuroscience, 22*(10), 2587-95.

Easton, R. D., & Shor, R. E. (1975). Information processing analysis of the Chevreul pendulum illusion. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 231-6.

Elsner, B., & Hommel, B. (2004). Contiguity and contingency in action-effect learning. *Psychological Research, 68*(2-3), 138-54.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York: Appleton-

Century-Crofts.

Fitts, P. M., & Posner, M. I. (1967). *Human Performance*. Oxford: Brooks/Cole.

Forssberg, H. (1999). Neural control of human motor development. *Current Opinion in Neurobiology, 9*(6)*,* 676-82.

Foulkes, A. J. & Miall, R. C. (2000). Adaptation to visual feedback delays in a human manual tracking task. *Experimental Brain Research, 131*, 101-110.

Galati, G., Lobel, E., Vallar, G., Berthoz, a, Pizzamiglio, L., & Le Bihan, D. (2000). The neural basis of egocentric and allocentric coding of space in humans: a functional magnetic resonance study. *Experimental Brain Research, 133*(2)*,* 156-64.

Georgiou, N., Iansek, R., Bradshaw, J. L., Phillips, J. G., Mattingley, J. B., & Bradshaw, J. A. (1993). An evaluation of the role of internal cues in the pathogenesis of parkinsonian hypokinesia. *Brain, 116*(6)*,* 1575-87.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika, 40*(1), 33-51.

Graziano, M. (2006). The organization of behavioral repertoire in motor cortex. *Annual Review of Neuroscience, 29,* 105-34.

Graziano, M. S. A., Taylor, C. S. R., & Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron, 34* (5), 841-851.

Guarraci, F. A., & Kapp, B. S. (1999). An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential pavlovian fear conditioning in the awake rabbit. *Behavioural Brain Research, 99*(2), 169-79.

Hagemann, N., & Memmert, D. (2006). Coaching anticipatory skill in badminton: Laboratory versus field-based perceptual training. *Journal of Human Movement Studies, 50*(6)*,* 381-398.

Haggard, P., Leschziner, G., Miall, R. C., & Stein, J. F. (1997). Local learning of inverse kinematics in human reaching movement. *Human Movement Science, 16*(1), 133-147.

Hallett, M. (2008). The intrinsic and extrinsic aspects of freezing of gait. *Movement Disorders, 23*(Suppl 2), S439-43.

Held, R., Efstathiou, A., & Greene, M. (1966). Adaptation to displaced and delayed visual feedback from the hand. *Journal of Experimental Psychology, 72*(6)*,* 887-891.

Hellon, R. F. (1986). Are single-unit recordings useful in understanding

thermoregulation? *The Yale Journal of Biology and Medicine, 59*(2), 197-203.

Hikosaka, O., & Wurtz, R. H. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. I. Relation of visual and auditory responses to saccades. *Journal of Neurophysiology, 49*(5), 1230-53.

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

Histed, M. H., Pasupathy, A. & Miller, E. K. (2009). Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. *Neuron, 63,* 244-253.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1*(4)*,* 304-9.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition, 8,* 164– 181.

Horvitz, J. C., Stewart, T., & Jacobs, B. L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research, 759*(2), 251-8.

Humphrey, N. (1995). Blocking out the distinction between sensation and perception: Superblindsight and the case of Helen. Behavioral and *Brain Sciences, 18*(2)*,* 257-258.

Humphrey, N. (2000). In Reply [Reply to commentaries on "How to solve the mind-body problem". *Journal of Consciousness Studies, 7*(4)*,* 98–112.

Jakobson, L. S., Servos, P., Goodale, M. A., & Lassonde, M. (1994). Control of proximal and distal components of prehension in callosal agenesis. *Brain, 117*(5)*,* 1107-13.

Jaśkowski, P., Jaroszyk, F., & Hojan-Jezierska, D. (1990). Temporal-order judgments and reaction time for stimuli of different modalities. *Psychological Research, 52*(1)*,* 35-8.

Jay, M. F., & Sparks, D. L. (1987). Sensorimotor integration in the primate superior colliculus. I. Motor convergence. *Journal of neurophysiology, 57*(1), 22-34.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: supervised learning with a distal teacher. *Cognitive Science, 16*(3)*,* 307-354.

Jordan, M. I., & Wolpert, D. M. (1999). Computational motor control. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences (Vol. 2, pp. 601-618).* Cambridge: MIT Press.

Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine, 14*(8)*,* 811-819.

Kish, G. B. (1955). Learning when the onset of illumination is used as reinforcing stimulus. *Journal of Comparative and Physiological Psychology, 48*(4), 261-4.

Kitazawa, S., Kohno, T., & Uka, T. (1995). Effects of delayed visual information on the rate and amount of prism adaptation in the human. *The Journal of Neuroscience, 15*(11)*,* 7644.

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

Kohfeld, D. L. (1971). Simple reaction time as a function of stimulus intensity in decibels of light and sound. *Journal of Experimental Psychology, 88*(2)*,* 251-7.

Lattal, K., & Gleeson, S. (1990). Response acquisition with delayed reinforcement. Journal of experimental psychology. *Animal Behavior Processes, 16*(1)*,* 27-39.

Leiphart, J., Rosenfeld, J. P., & Gabrieli, J. D. (1993). Event-related potential correlates of implicit priming and explicit memory tasks. *International Journal of Psychophysiology, 15*(3), 197-206.

Levy, J., Trevarthen, C., & Sperry, R. W. (1972). Perception of bilateral chimeric figures following hemispheric deconnexion. *Brain, 95*(1)*,* 61-78.

Lewis, G. N., Byblow, W. D., & Walt, S. E. (2000). Stride length regulation in Parkinson's disease: the use of extrinsic, visual cues. *Brain, 123*(10), 2077-90.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology, 67*(1)*,* 145-63.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*, 19751–19756.

Mars, R. B., Coles, M. G. H., Hulstijn, W., & Toni, I. (2008). Delay-related cerebral activity and motor preparation. *Cortex, 44*(5)*,* 507-20.

Matsumoto, N., Minamimoto, T., Graybiel, A. M., & Kimura, M. (2001). Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. *Journal of Neurophysiology, 85*(2), 960-

76.

May, P. J., McHaffie, J. G., Stanford, T. R., Jiang, H., Costello, M. G., Coizet, V., . . . Redgrave, P. (2009). Tectonigral projections in the primate: a pathway for pre-attentive sensory input to midbrain dopaminergic neurons. *The European Journal of Neuroscience, 29*(3)*,* 575-87.

McGuigan, N., Makinson, J., & Whiten, A. (2011). From over-imitation to super-copying: adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *British Journal of Psychology, 102*(1), 1-18.

McGuigan, N., Whiten, A., Flynn, E. F., & Horner, V. (2007). Imitation of causally necessary versus unnecessary tool use by 3- and 5-year-old children. *Cognitive Development, 22,* 356–364.

Miall, R. C. & Jackson, J. K. (2006). Adaptation to visual feedback delays in manual tracking: evidence against the Smith Predictor model of human visually guided action. *Experimental Brain Research, 172,* 77-84.

Miall, R. C., Weir, D. J., & Stein, J. F. (1985). Visuomotor tracking with delayed visual feedback. *Neuroscience, 16*(3)*,* 511-20.

Miall, R. C., Weir, D. J., Wolpert, D. M., & Stein, J. F. (1993). Is the cerebellum a Smith predictor? *Journal of Motor Behavior, 25*(3), 203-16.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks, 9*(8)*,* 1265-1279.

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE, 49*(1), 8-30.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience, 16*(5), 1936-47.

Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*(2)*,* 239–260.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale: Erlbaum.

Nieuwboer, A., Rochester, L., Müncks, L., & Swinnen, S. P. (2009). Motor learning in Parkinson's disease: limitations and potential for rehabilitation. *Parkinsonism & Related Disorders, 15*(Suppl 3)*, S53*-8.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology, 19*(1)*,* 1–32.

Okouchi, H. (2009). Response acquisition by humans with delayed reinforcement. *Journal of the Experimental Analysis of Behavior, 91*(3)*,* 377-90.

Ostlund, S. B., Winterbauer, N. E., & Balleine, B. W. (2009). Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *The Journal of Neuroscience, 29*(25)*,* 8280-7.

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

Pavese, N., & Brooks, D. J. (2009). Imaging neurodegeneration in Parkinson's disease. *Biochimica et Biophysica Acta, 1792*(7)*,* 722-9.

Peterson, G. B. (2004). A day of great illumination: B. F. Skinner's discovery of shaping. *Journal of the Experimental Analysis of Behavior, 82*(3), 317-28.

Ravel, S., & Richmond, B. J. (2006). Dopamine neuronal responses in monkeys performing visually cued reward schedules. *The European Journal of Neuroscience, 24*(1), 277-90.

Redgrave, P. & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience, 7,* 967-975.

Redgrave, P., Gurney, K., & Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews, 58*(2)*,* 322-39.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., . . . Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews Neuroscience, 11*(11)*,* 760-72.

Reiner, A., Jiao, Y., Del Mar, N., Laverghetta, A. V., & Lei, W. L. (2003). Differential morphology of pyramidal tract-type and intratelencephalically projecting-type corticostriatal neurons and their intrastriatal terminals in rats. *The Journal of Comparative Neurology, 457*(4)*,* 420-40.

Renner, K. E. (1964). Delay of reinforcement: A historical review. *Psychological Bulletin, 61*(5), 341-361.

Rousselet, G. A, Thorpe, S. J., & Fabre-Thorpe, M. (2004). How parallel is visual processing in the ventral pathway? *Trends in Cognitive Sciences, 8*(8), 363-70.

Rubinstein, T. C., Giladi, N., & Hausdorff, J. M. (2002). The power of cueing to circumvent dopamine deficits: a review of physical therapy treatment of gait disturbances in Parkinson's disease. *Movement Disorders, 17*(6)*,* 1148-60.

Sacks, O. (1982) *Awakenings*. London: Pan Books.

Salamone, J. D., Correa, M., Farrar, A. M., Nunes, E. J., & Pardo, M. (2009). Dopamine, behavioral economics, and effort. *Frontiers in Behavioral Neuroscience, 3*(13).

Savaki, H., Kennedy, C., Sokoloff, L., & Mishkin, M. (1993). Visually guided reaching with the forelimb contralateral to a "blind" hemisphere: a metabolic mapping study in monkeys. The Journal of Neuroscience*, 13*(7)*,* 2772-2789.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences, 3*(6)*,* 233-242.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*(1)*,* 1-66.

Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *Journal of Neurophysiology, 56*(5)*,* 1439-61.

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

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience, 1*(3), 199-207.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology, 57,* 87-115.

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. *The Journal of Neuroscience, 13*(3), 900-13.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306)*,* 1593-9.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition, 19*(4)*,* 353-60.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgment of causality by human subjects. *Quarterly Journal of Experimental Psychology, 41B,* 139–159.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*(3)*,* 367-395.

Sheridan, T. B. (1993). Space teleoperation through time delay: review and prognosis. *IEEE Transactions on Robotics and Automation, 9*(5)*,* 592-606.

Sheridan, T. B., & Ferrell, W. R. (1963). Remote manipulative control with transmission delay. *IEEE Transactions on Human Factors in Electronics, HFE-4*(1), 25-29.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Transactions on Autonomous Mental Development, 2*(2)*,* 70-82.

Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning, 22*(1-3), 123-158.

Sizemore, O. J., & Lattal, K. A. (1977). Dependency, temporal contiguity, and response-independent reinforcement. *Journal of the Experimental Analysis of Behavior, 27*(1)*,* 119-125.

Skinner, B. F. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology, 38*(2), 168.

Skinner, B. (1962). Operandum. *Journal of the Experimental Analysis of Behavior, 5*(2), 224-224.

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis.* New York: Appleton-Century-Crofts.

Snycerski, S., Laraway, S., Huitema, B. E., & Poling, A. (2004). The effects of behavioral history on response acquisition with immediate and delayed reinforcement. *Journal of the Experimental Analysis of Behavior, 81*(1)*,* 51-64.

Snycerski, S., Laraway, S., & Poling, A. (2005). Response acquisition with immediate and delayed conditioned reinforcement. *Behavioural Processes, 68*(1)*,* 1-11.

Soltani, A., & Wang, X.-J. (2008). From biophysics to cognition: reward-dependent adaptive choice behavior. *Current Opinion in Neurobiology, 18*(2)*,* 209-16.

Sparks, D. L. (1986). Translation of sensory signals into commands for control of saccadic eye movements: role of primate superior colliculus. *Physiological Reviews, 66*(1)*,* 118-71.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171-7.

Staddon, J. E. R. & Niv, Y. (2008). Operant conditioning. *Scholarpedia, 3*(9)*:*2318.

Staubli, U., & Lynch, G. (1987). Stable hippocampal long-term potentiation elicited by "theta" pattern stimulation. *Brain Research, 435*(1-2)*,* 227-34.

Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron, 51*(5)*,* 651-9.

Stoerig, P., & Cowey, A. (1997). Blindsight in man and monkey. *Brain, 120*(3)*,* 535-59.

Stubbs, A. (1969). Contiguity of briefly presented stimuli with food reinforcement. *Journal of the Experimental Analysis of Behavior, 12*(2), 271-8.

Sumner, P., Adamjee, T., & Mollon, J. D. (2002). Signals invisible to the collicular and magnocellular pathways can capture visual attention. *Current Biology, 12*(15)*,* 1312-6.

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*(4), 2520-9.

Thompson, A. (1997). An evolved circuit, intrinsic in silicon, entwined with physics. *Evolvable Systems: From Biology to Hardware, 1259,* 390-405.

Thorndike, E. L. (1911). *Animal intelligence: experimental studies.* New York: The Macmillan Company.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology, 39*(1)*,* 212-222.

Thorpe, S. J., & Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science, 291*(5502), 260-263.

Timberlake, W., & Lucas, G. A. (1985). The basis of superstitious behavior: chance contingency, stimulus substitution, or appetitive behavior? *Journal of the Experimental Analysis of Behavior, 44*(3)*,* 279-99.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., Lecea, L. de & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science, 324*(5930), 1080-4.

Ungerleider, L. (2002). Imaging Brain Plasticity during Motor Skill Learning. *Neurobiology of Learning and Memory, 78*(3), 553-564.

Versnel, H., Zwiers, M. P., & van Opstal, A. J. (2009). Spectrotemporal response properties of inferior colliculus neurons in alert monkey. *The Journal of Neuroscience, 29*(31)*, 9725*-39.

Vorhees, C. V., & Williams, M. T. (2006). Morris water maze: procedures for assessing spatial and related forms of learning and memory. *Nature Protocols, 1*(2)*,* 848-58.

Wang, R. F., & Spelke, E. S. (2000). Updating egocentric representations in human navigation. *Cognition, 77*(3)*,* 215-50.

Willingham, D. B., Salidis, J., & Gabrieli, J. D. E. (2002). Direct comparison of neural systems mediating conscious and unconscious skill learning. *Journal of Neurophysiology, 88*(3)*,* 1451-1460.

Winograd, T. (1975). Frame representations and the declarative/procedural controversy. In D. D. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 185–210). New York: Academic Press.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences, 5*(11)*,* 487-494.

Wurtz, R. H., & Albano, J. E. (1980). Visual-motor function of the primate superior colliculus. *Annual Review of Neuroscience, 3*(1)*,* 189-226.

Appendices

Appendix 1

Onscreen instructions for experiment 1 (within-trial repetition of audio reinforcement)

1. Welcome to the Joystick Task. Please read the instructions using the space bar to move from page to page.
2. You will start with a practice session composed of 3 trials to get you limbered up.
3. The task is to find the correct position to place the joystick in. Finding it will result in a "pip" sound.
4. Keep the joystick in the position where you found the pip until you hear another sound indicating success.
5. Trials will end automatically and move on to the next one after a short delay.
6. The first trial will start when you press "Space bar"; thereafter trials start automatically. Trials are counted in by 3 beeps.
7. Press the "Space bar" to start the practice session.
8. You found the correct position! Get ready for trial (...) of 3.
9. You found the correct position!
10. Practice session over; time for the real thing. There are 18 trials in total.
11. Please complete trials as quickly as possible; some will be more difficult than others.
12. Press "Space bar" to start.
13. You found the correct position! Get ready for trial (...) of 18
14. You found the correct position and that was the last one. Phew!
15. The experiment is now over. Thanks for participating!

Appendix 2

Onscreen instructions for experiment 2 (within-trial repetition of visual reinforcement)

1. If you suffer from epilepsy DO NOT continue with the task as screen flashes could potentially induce a seizure.
2. Welcome to the Joystick Task. Please read the instructions using the space bar to move from page to page.
3. You will start with a practice session composed of 3 trials to get you limbered up.
4. The task is to find the correct position to hold the joystick in. Finding it will result in a screen flash.
5. Keep the joystick in the position where you found the screen flash until you hear a sound indicating success.
6. Trials will end automatically and move on to the next one after a short delay.
7. The first trial will start when you press "Space bar"; thereafter trials start automatically. Trials are counted in by 3 beeps.
8. If you have any problems or queries please tell Tom and he will be happy to assist you.
9. Press the "Space bar" to start the practice session.
10. You found the correct position! Get ready for trial of (...) of 3.
11. You found the correct position!
12. Practice session over; time for the real thing. There are 18 trials in total.
13. Please complete trials as quickly as possible; some will be more difficult than others.
14. Press "Space bar" to start.
15. You found the correct position! Get ready for trial (...) of 18
16. You found the correct position and that was the last one. Phew!
17. The experiment is now over. Thanks for participating!

Appendix 3

Onscreen instructions for experiment 3 (between-trial repetition of audio reinforcement)

1. Welcome to the Joystick Task. Please read the instructions using "Space bar" to move from page to page.
2. You will start with a practice session composed of 3 trials to get you limbered up.
3. The task is to find the correct area to put the joystick in as quickly as possible. As soon as you find the correct area the trial will end.
4. The next trial will begin automatically after a short delay. When it begins, you need to search for the same area again.
5. The first trial will start when you press "Space bar" thereafter trials start automatically. All trials are counted in by 3 beeps.
6. If you have any problems or queries please tell Tom and he will be happy to assist you.
7. Press the "Space bar" to start the practice session.
8. You found the correct position! Get ready for the next trial...
9. You found the correct position!
10. Practice session over time for the real thing. Press "space bar" to move through these instructions.
11. There are 4 sets of 10 trials: i.e. 40 trials in total.
12. Within a set of 6 trials the correct area to move the joystick to remains the same.
13. When you complete a set of 10 trials, the position of the area changes to another, randomly determined, location and you have to start searching afresh.
14. Please complete trials as quickly as possible; some will be more difficult than others.
15. Press "Space bar" to start.
16. You found the correct position! Get ready for the next trial...
17. You found the correct position and have finished set 1 of 4!

18. Remember, the correct area has now been moved, at random, to a new location. Press space bar when ready to start the second set of trials.
19. You found the correct position and have finished set 2 of 4!
20. Remember, the correct area has now been moved, at random, to a new location. Press space bar when ready to start the third set of trials.
21. You found the correct position and have finished set 3 of 4!
22. Remember, the correct area has now been moved, at random, to a new location. Press space bar when ready to start the final set of trials.
23. You found the correct position and that was the last one. Phew!
24. The experiment is now over. Thanks for participating!

Appendix 4

Onscreen instructions for experiment 4 (between-trial repetition with delayed reinforcement)

1. Welcome to the Joystick Task. Please read the instructions using "Space bar" to move from page to page.
2. You will start with a practice session composed of (_)trials to get you limbered up.
3. The task is to find the correct area to put the joystick in as quickly as possible. As soon as you find the correct area the trial will end.
4. The next trial will begin automatically after a short delay. When it begins, you need to search for the same area again.
5. The first trial will start when you press "Space bar"; thereafter trials start automatically. All trials are counted in by 3 beeps.
6. If you have any problems or queries please tell Tom and he will be happy to assist you.
7. Press the "Space bar" to start the practice session.
8. Practice session over; time for the real thing. Press "space bar" to move through these instructions.
9. There are 8 sets of 10 trials: i.e. 80 trials in total.
10. Within a set of 10 trials the correct area to move the joystick to remains the same.
11. When you complete a set of 10 trials, the position of the area changes to another, randomly determined, location and you have to start searching afresh.
12. Please complete trials as quickly as possible; some will be more difficult than others.
13. Press "Space bar" to start.
14. Found it! Get ready for the next trial...
15. You found the correct position and that was the last one. Phew!
16. The experiment is now over. Thanks for participating!