EXPECTATION VIOLATIONS AND EMOTIONAL LEARNING

Cordelia Fine

Department of Psychology,
University College London.

Thesis submitted for the degree of Ph.D. in Psychology,
University of London.
October 2000.

Acknowledgements

Foremost, I'd like to thank James Blair for his indispensable guidance and encouragement. His patience and sense of humour have helped make my PhD an enjoyable experience. I am also indebted to Uta Frith for her constructive criticism, encouragement and support. Lisa Cipolotti, Caroline Edmonds, and John Morton also deserve thanks for giving their time to discuss many ideas in this thesis.

I am grateful to Peter Dayan for his supervision of my project during the year I spent in the Computer Science department. His expert advice, interest and patience were greatly appreciated. I am also extremely grateful to Essi Colledge, John Lumsden and Derek Mitchell for their help with the patient research.

Thanks also to all who have provided computer support and help and advice, in particular, Giulia Abbonizio, Sylvie Berthoz, Janice Brown, Essi Colledge, Caroline Edmonds, Sarah Paterson, Daniel Reisel, Caleb Smith, and especially Charmaine Bill. John Draper must also be thanked for his administrative omniscience and unfailing good humour.

I thank all my family for their great encouragement and support. And finally, thanks to Russell.

This PhD was supported by a Medical Research Council studentship and a University College London Graduate School Cross-disciplinary scholarship.

Abstract

Chapter 1 discusses the representation of reinforcement expectations. Chapter 2 tested the prediction that expectation violations will trigger arousal and rapid behavioural change. It was found that unexpected valence changes, but not magnitude changes, triggered arousal increases and rapid behavioural change. It was therefore suggested that an instrumental learning system represents both magnitude and valence information, but that a separable instrumental re-learning system only represents valence. These hypotheses were implemented in a computational model in Chapter 3. This model successfully simulated human behavioural data from three experiments in Chapter 2. The model also predicted that the instrumental learning and instrumental re-learning systems could be independently damaged. This was investigated in Chapters 4 and 5 with case studies of an amygdala patient and two orbitofrontal patients. The amygdala patient was severely impaired in instrumental learning. In contrast, the orbitofrontal patients were only impaired in instrumental re-learning. This dissociation supported the hypothesis that instrumental learning and re-learning are mediated by separable systems. Chapter 6 found support for the hypotheses that developmental psychopathy is associated with amygdala dysfunction and orbitofrontal cortex function by assessing instrumental learning and re-learning in a population of psychopathic individuals. Chapters 7 and 8 investigated further the effects of early amygdala damage on emotional and social cognition. A patient with early left amygdala damage was shown to be impaired in the recognition of fear and sadness, and showed a lack of empathy. These findings were predicted by the early amygdala dysfunction hypothesis of developmental psychopathy. Chapter 8 demonstrated a severe theory of mind impairment in the amygdala patient, in the absence of any executive dysfunction. This finding suggests that theory of mind is not simply a function of more general executive functions, and supports the hypothesis that the amygdala plays a role in the development of the circuitry mediating theory of mind. In the last chapter, future research directions are identified.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Response outcome expectations: representation and violation

1.1: Introduction

Few would disagree with the notion that much of our behaviour is guided by expectations of the positive or negative consequences of our actions (e.g., Tolman, 1932; Bolles, 1972; Balleine & Dickinson, 1998). This chapter is concerned with the question of how these expectations of future reward and punishment are represented cognitively.

1.2: Reinforcement expectations and expectation violations

The representation of the expected relationship between response and reinforcement is implicit or explicit in most influential models of emotional processing, for example, models of emotional experience (e.g., Mandler, 1984; Oatley & Johnson-Laird, 1987), anxiety (e.g., Gray, 1982), stimulus-reinforcement association re-learning (e.g., Grossberg, 1982; Rolls, 1990), frustration (Amsel, 1992), latent inhibition (Schmajuk, Lam, & Gray, 1996), and consciousness (Gray, 1995). These models also include mechanisms that detect violations of expectation. Moreover, the detection of these violations have important consequences, physically and/or psychologically. For example, in his model of the construction of emotional experience, Mandler proposes that discrepancies of perception, action, or thought are detected by a "difference detector" (Mandler, 1991) that triggers autonomic arousal (e.g., Mandler, 1964; Mandler, 1984; see figure 1.1). This arousal determines the intensity of the resulting emotional experience. In addition, the discrepancy results in an interruption of on-going behaviour (Mandler, 1964).

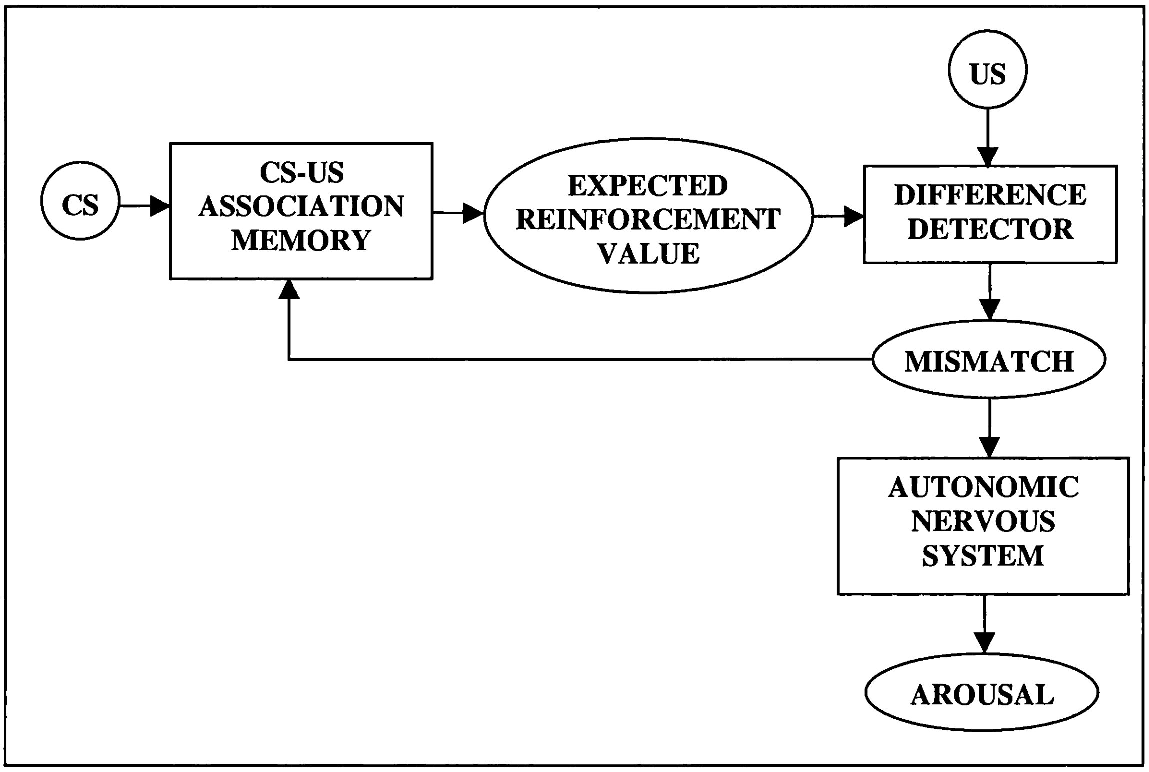

Figure 1.1: Possible representation of Mandler's discrepancy/arousal model (e.g., Mandler, 1984).

Gray's model of the behavioural inhibition system describes an information processing system that compares the current state of the world to predictions of the expected state of the world (e.g., Gray, 1982; Gray, 1987; Gray, 1995; see figure 1.2). For example, if reward is expected but does not occur, the behavioural inhibition system is activated and the animal will become aroused, increase attention, and inhibit current behaviour. It is proposed that activation of the behavioural inhibition system results in a state of anxiety.

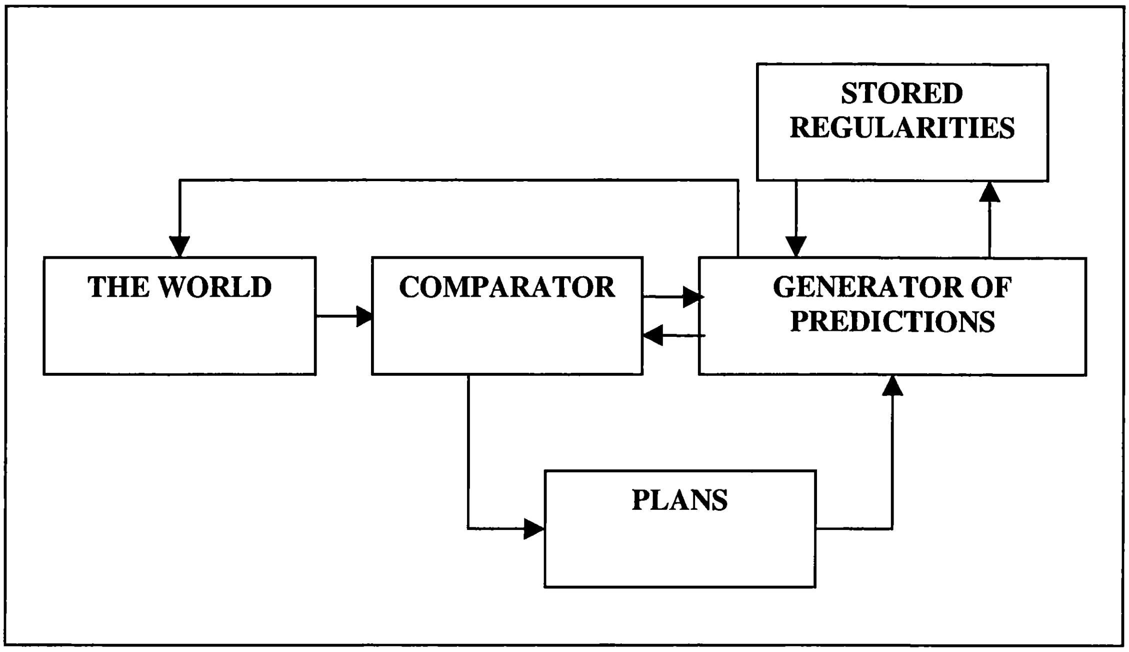

Figure 1.2: The comparator predicts the animals' expected state in the next instant of time, and compares this prediction to the actual state of the world. From Gray, 1982: p263.

Amsel's frustration theory proposes that a short-term increment in arousal follows the omission of an expected reward, or reward that is less than expected (Amsel, 1992). In Grossberg's model, expectation violations trigger the orienting response, which allows rapid re-learning of stimulus-reinforcement associations to occur (Grossberg, 1982).

Responses that result in reinforcement events that violate expectations are therefore of crucial importance in these important models of emotional processing. In each model, an expectation of future reward or punishment must be compared with actual reinforcement. However, depending on how reinforcement expectations are represented, it is possible that not all discrepancies of reinforcement will violate expectations. If a particular variable – for example, valence - is represented in a reinforcement expectation, then a change in the valence of reinforcement will violate expectations. In contrast, if a particular variable is not represented in a reinforcement expectation – for example, magnitude - then a change in the magnitude of reinforcement will not violate expectations. In some models, reinforcement expectations have been explicitly defined computationally. In Grossberg's (1982) model, expected and actual events are represented as vectors of activity. An expectation violation occurs when the difference between the expected and actual event is sufficiently large that the combination of the two vectors is suppressed as noise. A "vigilance parameter" determines the degree of mismatch that will be tolerated. However, Grossberg does not state what parameters of the stimulus are represented in the vectors. In Schmajuk et al.'s (1996) model of latent inhibition, differences in magnitude between expected and actual reinforcement are detected. Mandler (1984) and Amsel (1992) also assume that discrepancies in the magnitude of expected and actual reinforcement are detected and responded to behaviourally and psychophysiologically. However, despite the importance in these models of what constitutes an expectation violation, the question of what information is represented in expectations of future reinforcement has not been well researched. The next section introduces a possible methodology for exploring this question.

1.3: Inferring expectations from autonomic arousal

A methodology for investigating the representation of expectations is suggested by the research of Sokolov (1960) on the orienting response. The orienting response is a non-specific response that includes increases in autonomic arousal. Sokolov was interested in orienting responses to novel events. He argued that novelty is the noncoincidence between external stimuli and a neuronal (or internal) model of predicted stimuli. He suggested that:

> *"... the function of this neuronal model ...is to filter the signals in the nervous system. This means that a selective prevention of transmission of impulses from certain kinds of stimulation must take place. We can measure the property of this biological filter by changing stimuli and by measuring reaction."* Sokolov, 1960: p208.

In other words, by manipulating a stimulus parameter and measuring consequent arousal, it is possible to infer what stimulus parameters are represented in the internal predictive model. For example, Sokolov (1960) habituated subjects to a sound of a particular frequency. He then measured a number of indices of the orienting response, including skin conductance response and respiration rate, following sounds of different frequencies. The size of the orienting response was proportional to the magnitude of the difference in frequency between the novel stimulus and the habituated stimulus. From this it could be inferred that stimulus frequency was represented in the internal model.

A number of other studies have explored autonomic arousal responses following unexpected and/or novel stimuli. Lewis & Goldberg (1969) found that violations of colour, form, content and curvature of a stimulus resulted in an arousal response. Wilton & Boersma (1974) found that violations of principles of conservation produced arousal responses in children who had acquired the principle of conservation, but not in children who had not. In a study by Mandler (1964), subjects learnt a sentence or a sequence of digits. The sequences were either learnt to mastery, or were over-learned. Subjects were then asked to anticipate each word or digit as the experimenter said the sequence they had learnt. On critical trials, an unexpected word or digit was said after the subject had anticipated the correct one. There was a skin conductance increase following the discrepant events, and this was greater when the sequences had been over-learned than when they had only been learnt to mastery. This suggests that the greater the expectation for an event, the greater the arousal response when that expectation is violated. Similarly, Nakamura (cited in MacDowell & Mandler, 1989) produced unexpected endings for stories and measured skin conductance in subjects listening to the stories. Unexpected endings produced skin conductance responses, and these were greater following unexpected endings to well-known proverbs than to novel stories.

Thus these data support the position that violations of expectation produce arousal and that the greater the violation, the greater the arousal that follows. In a similar way, in can be predicted that violations of reinforcement expectations will also trigger autonomic arousal responses. Indeed, this prediction is made by the discrepancy/arousal hypothesis (e.g., Mandler, 1984), the behavioural inhibition system model (e.g., Gray, 1982), frustration theory (Amsel, 1992) and Grossberg's computational model (Grossberg, 1982). Following the approach of Sokolov (1960), arousal responses following reinforcement can be used to infer what information is and is not represented in reinforcement expectations that input into a violation detection mechanism. If a particular variable is represented in a reinforcement expectation, then discrepancy in that variable will produce arousal. In contrast, if a particular variable is not represented in a reinforcement expectation, then discrepancy in that variable will not produce arousal. Thus, by investigating what kinds of violations produce arousal, we can infer what information is represented in reinforcement expectations.

So far, only one study conducted by MacDowell & Mandler (1989) has tested the prediction that discrepant reinforcements produce arousal. In this experiment, subjects played an interactive adventure computer game in which both good and bad events could occur. The dependent variables assessed autonomic arousal responses, indexed by heart rate response, heart rate response acceleration, skin conductance response, and skin conductance response acceleration. There were three phases to the experiment. Session 1 was a one hour expectation acquisition session. In session 2, autonomic arousal responses to the expected events were recorded. In the third session, unexpected events occurred randomly, and autonomic arousal responses to the events were recorded (see table 1.1). Two unexpected events comprised outcomes better than the typical event, and two involved outcomes worse than the typical event.

EVENT	EXPECTED	UNEXPECTED
POSITIVE		
Gold	70 pieces	1000 pieces
Death	Kill weak monster	Powerful monster drops dead
NEGATIVE		
Gold	30 pieces	3 pieces
Death	Killed by powerful monster	Killed by weak monster

Table 1.1: Expected and unexpected events: from MacDowell & Mandler (1989).

It was found that, as predicted, discrepant events were associated with a greater increase in heart rate than expected events. However, this association was not significant for the other three measures of autonomic arousal. A further difficulty with interpreting the results was that the unexpected events differed to the expected events in ways other than whether they violated expectations. It is therefore possible that the unexpected events were intrinsically more arousing than the expected events. For example, in the positive gold condition, the unexpected event (1000 gold pieces) was of greater magnitude than the expected event (70 gold pieces). Thus the greater heart rate increase seen following the unexpected events cannot be unambiguously attributed to violations of expectation. An additional drawback is that MacDowell & Mandler (1989) did not establish that subjects had actually developed expectations during the first two sessions of the experiment.

The results of this study are therefore inconclusive. However, the use of a computer game provides a useful approach to investigating the effects of discrepant reinforcement on autonomic arousal. This approach was used in the experiments described in Chapter 2. The first aim of these experiments was to explore what discrepancies of reinforcement produce arousal, whilst controlling for absolute levels of reinforcement.

1.4: When a discrepancy is not a violation

The previous section discussed the possibility that not all discrepancies will result in arousal. If so, then an interesting question is why this might be the case. As Sokolov (1960) suggested, the answer may lie in the processing aim of the system detecting violations of expectations. In Mandler's model, violations result in interruption of ongoing behaviour, and the arousal readies the individual for a potentially important change in the environment (e.g., Mandler, 1964; Mandler, 1984; Mandler, 1991). In Gray's model, violations of expectation, together with punishment and non-reward, activate the behavioural inhibition system which prevents the individual from continuing with plans that are potentially maladaptive (e.g., Gray, 1982). Grossberg (e.g., Grossberg, 2000) and Rolls (e.g., Rolls, 1990; 1996) both stress the importance of detecting expectation violations in order to rapidly re-learn reinforcement contingencies, preventing perseverative responding. In other words, the models suggest that an important consequence of detecting an expectation violation is a rapid behavioural change or behavioural interruption.

This suggests a reason why not all discrepancies may be detected as violations; behavioural change may not always be appropriate, as Grossberg (1982) has argued. He notes that internal representations in cognitive systems must be able to:

> *" ... maintain themselves in a stable fashion against the erosive effects of behaviourally irrelevant environmental fluctuations yet ... nonetheless adapt rapidly in response to environmental fluctuations that are crucial to survival."* Grossberg, 1982: p536.

Grossberg terms this the stability-plasticity dilemma. The dilemma arises from the system not "knowing" *a priori* what are and are not important violations of what is expected. The effects of reinforcement discrepancy on arousal and behaviour have been investigated previously in two experimental paradigms: extinction and reversal learning. In extinction, termination of an expected reinforcement abolishes both classical and instrumental responding to a conditioned stimulus (see LeDoux, 1998). In reversal learning, the reinforcement value of a stimulus is changed from positive to negative or vice versa. Control subjects rapidly adapt their instrumental responses in response to the changed reinforcement contingencies. For example, in the Intra-dimensional/Extra-dimensional Shift task – a visual discrimination and attentional set-shifting task - the mean number of errors on simple visual discrimination reversals was between 1 and 2 (Rahman, Sahakian, Hodges, Rogers, & Robbins, 1999).

Thus, the termination of reinforcement or reversal of valence of reinforcement results in rapid behavioural change. In contrast, the effects of magnitude changes on instrumental responding are not clear. A small number of studies have examined instrumental responses in rats following shifts in the magnitude of reward associated with a particular behavioural response (see Flaherty, 1982: p 422-423). However, the findings are inconclusive since in these studies, following changes in reinforcement magnitudes, the stimuli between which the animal had to choose were of equal value. Thus these experiments did not have the potential to demonstrate adaptive behavioural change following changes in reward magnitude. The second aim then of the experiments in Chapter 2 was to measure behavioural change following changes in the magnitude and valence of reinforcement, and to explore the relationship between arousal following expectation violations and behavioural change. If the models discussed here are correct in proposing that the function of detecting expectation violations is rapid behavioural change, then there should be an association between arousal responses and rapid behavioural change.

This section has so far omitted an influential neurocognitive model of the representation of reinforcement expectations and how they control behaviour. This is the somatic marker hypothesis (Damasio, Tranel, & Damasio, 1991; Damasio, 1994; Damasio, 1996; Bechara, Damasio, & Damasio, 2000). This model differs importantly to those discussed above by claiming that the arousal responses that precede behavioural decisions index expectations about the outcome of behaviour. In contrast, in the models discussed previously, arousal responses index violations of these expectations. The somatic marker hypothesis is the subject of the next section.

1.5: The somatic marker hypothesis

The somatic marker hypothesis proposes that bodily states, or "somatic markers", guide emotional decision-making:

> *"...when the somatosensory image which defines a certain emotional response is juxtaposed to the images which describe a related scenario of future outcome, and which triggered the emotional response via the ventromedial linkage, the somatosensory pattern marks the scenario as good or bad.*
>
> *"When this process is overt, the somatic state operates as an alarm or incentive signal. The somatic state is alerting you to the goodness or badness of a certain option-outcome pair."* Bechara et al., 2000: p297.

It can be seen from this description that the somatic marker hypothesis predicts the occurrence of conditioned autonomic responses to conditioned stimuli. These autonomic responses are labelled somatic markers. Furthermore, the model differs importantly from the classical conditioning account. Its major claim is that the function of these markers is to guide behaviour, by acting as an "alarm bell" or a "beacon of incentive" (Damasio, 1994). That is, the arousal responses that precede a behavioural choice are causal in determining behaviour (see figure 1.3). In contrast, in the models discussed previously, while arousal responses may accompany the presence of conditioned stimuli, they are not considered to play a causal role in emotional decision-making.

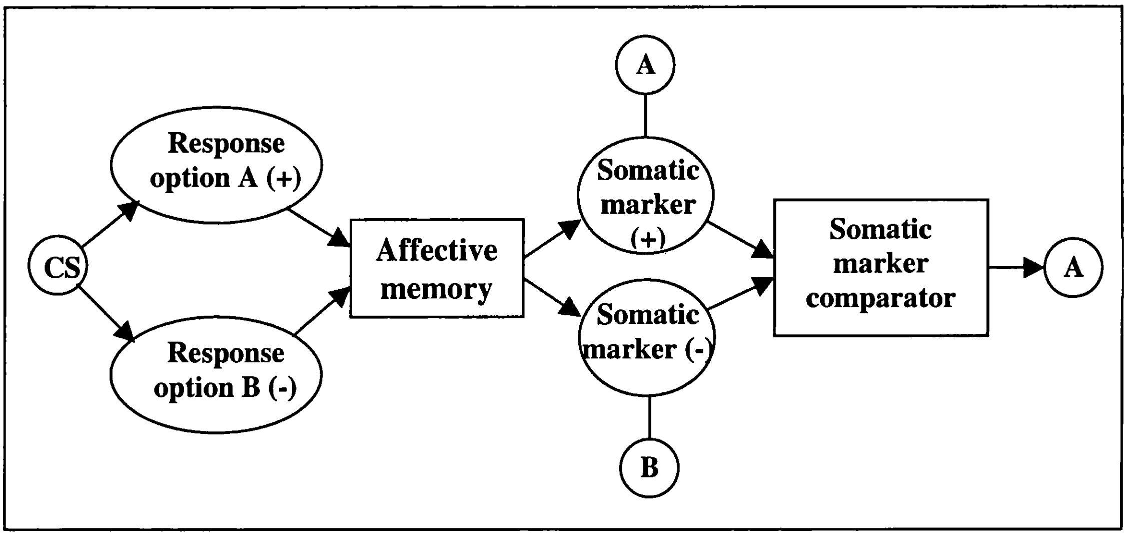

Figure 1.3: Possible representation of the somatic marker hypothesis (e.g., Damasio, 1994; Bechara et al., 2000). CS = conditioned stimulus.

1.6: Data in support of the somatic marker hypothesis

Findings from a gambling game, the Four Pack Gambling task, provide the main support for the somatic marker hypothesis (Damasio et al., 1991; Bechara, Damasio, Damasio, & Anderson, 1994; Bechara, Damasio, Tranel, & Damasio, 1997; Bechara, Damasio, Damasio, & Lee, 1999). This innovative task has provided one of the first consistent demonstrations of a cognitive deficit in patients with ventromedial prefrontal cortex damage. In addition, the task has provided an impetus for the development of other important decision-making tasks (e.g., Rogers, Everitt, Baldacchino, et al., 1999). In the Four Pack Gambling task, subjects are told to select cards at will from any of four decks. Two of the decks yield high rewards and high punishments (counterfeit money) that sum to a net loss (bad decks). The other two decks yield low rewards and low punishments that sum to a net gain (good decks). One good deck and one bad deck are associated with small but frequent punishment, and the other two decks are associated with larger but less frequent punishment. There are three dependent variables; one behavioural and two psychophysiological. The behavioural variable is the number of choices of good and bad decks. The two psychophysiological variables are skin conductance responses (SCRs) preceding the decision of which deck to choose, and SCRs following reinforcement.

Neurologically intact control subjects and patients with ventromedial frontal lobe (VMF) damage were given the task. VMF patients are characterised as showing poor social and emotional decision-making, including an insensitivity to the social repercussions of their behaviour (e.g., Eslinger & Damasio, 1985; Rolls, Hornak, Wade, & McGrath, 1994). This impairment cannot be explained in terms of defects in pertinent knowledge (Saver & Damasio, 1991), intellectual ability as assessed by IQ (e.g., Eslinger & Damasio, 1985; Rolls et al., 1994), or basic working memory (Bechara, Damasio, Tranel, & Anderson, 1998). The gambling task was designed to *"simulate real-life decisions in terms of uncertainty, reward, and punishment."* (Bechara et al., 1999: p5473). It was therefore of interest to investigate whether the VMF patient group would make poor decisions in the task, as they do in everyday life.

As predicted, unlike controls, the VMF patients did not learn to avoid the bad packs. In addition, their psychophysiological data differed to that of the controls. Both the VMF patients and the controls generated SCRs following rewards and punishments from their card selections. However, as the experiment progressed, controls began to generate large SCRs as they reached for cards from the bad decks, as if in anticipation of the punishment they might be about to receive. In contrast, the VMF patients did not produce anticipatory SCRs. Moreover, they failed to learn to avoid the high risk packs.

Bechara *et al.* (Bechara et al., 1994; 1997; 1999) interpreted the absence of anticipatory SCRs in the VMF patients as an absence of negative somatic markers to "warn" them away from the bad decks. They suggested that:

> *"the poor decision-making associated with VMF damage is related to an inability to integrate effectively all of the somatic state information triggered by the amygdala as well as other somatic effectors such as the hypothalamus and brainstem nuclei. ... When subjects decide to select cards from a specific deck, the neural activity pertaining to this information is signaled to VM cortices, which in turn activate the amygdala ... This latter activity would reconstitute a somatic state that integrates the numerous and conflicting instances of reward and punishment encountered with individual card draws from that deck."* Bechara et al., 1999: p5480.

A recent functional imaging study has investigated this claim. The study revealed activity in bilateral medial prefrontal areas both preceding and during SCRs in an emotional decision-making task (Critchley, Elliott, Mathias, & Dolan, 2000). The authors argue that this is consistent with a role for this region in generating and representing somatic markers. However, while these findings support the role of the medial prefrontal cortical areas in representing reinforcement and expectations of future outcomes, which are themselves associated with SCRs, this does not imply that these SCRs are causal in emotional decision-making.

1.7: Alternative accounts of the Four Pack Gambling task results

Indeed, in the same way it cannot be assumed that the larger anticipatory SCRs seen in controls preceding choices from the bad decks in the Four Pack Gambling task reflect the action of negative somatic markers. This is because the bad decks were associated with higher levels of reward as well as punishment. Thus the larger anticipatory SCRs for these decks may be explained by the higher absolute reinforcement values associated with the bad decks compared with the good decks. Damasio (1996) does consider a similar alternative account, but counters this with the observation that subjects develop their anticipatory SCRs before they can explicitly categorise the decks as good or bad. However, evidence of conditioning to a stimulus does not necessarily imply that the subject is explicitly aware of the stimulus-reinforcement contingencies (e.g., Edwards, 1990; Johnsrude, Owen, Zhao, & White, 1999).

With regard to the data from the VMF patients, their reduced anticipatory SCRs may be attributable to the passive nature of the decision-making stage of the experiment. Damasio, Tranel, & Damasio (1990) found that patients with bilateral lesions in orbital and lower mesial frontal lesions did not generate SCRs while passively looking at slides of conditioned stimuli. Possible alternative accounts of the VMF patients' poor behavioural choices include an instrumental re-learning deficit (Rolls, 2000) and an inability to resolve effectively between two competing response options (Rogers et al., 1999).

1.8: Summary and experimental aims of Chapter 2

Numerous important models of emotional processing include mechanisms that detect violations of response outcome expectations (e.g., Gray, 1982; Mandler, 1984; Schmajuk et al., 1996; Grossberg, 2000). Following Sokolov (1960), it is argued that the representation of expectations can be investigated by presenting the subject with outcomes that do not match their expectations. If mismatch on a particular stimulus parameter produces an arousal response, this indicates that that parameter is represented in the expectation. Conversely, the absence of an arousal response suggests that the parameter is not represented. The first aim of the next chapter was to explore the representation of response outcome expectations by measuring autonomic arousal responses following reinforcement discrepancies. The importance of the detection of response outcome expectations violations for rapid behavioural change and interruption of behaviour is stressed by a number of models (e.g., Mandler, 1964; Gray, 1982; Rolls, 1990). The second aim of Chapter 2 was to investigate whether rapid behavioural change is associated with arousal responses following violations of expectations.

Chapter 2

Autonomic arousal and behavioural change following expectation violations

2.1: Introduction of aims and predictions

The experiments reported in this chapter were designed to achieve the aims summarised in section 1.8 of the previous chapter. The first aim is to explore the representation of reinforcement expectations by measuring arousal responses following reinforcement expectation violations. In the models discussed in Chapter 1, it is either implicitly or explicitly proposed that both magnitude and valence violations are detected. It is therefore predicted that both magnitude and valence violations will trigger autonomic arousal. The second aim is to investigate whether arousal responses following expectation violations are associated with rapid behavioural change. This is because in many of the models discussed in Chapter 1, the detection of expectation violations results in behavioural inhibition or behavioural change (e.g., Gray, 1982; Grossberg, 1982; Mandler, 1984).

In the first three experiments, three variants of an instrumental learning and re-learning task were used. The task took the form of a computer game in which two stimuli were presented simultaneously, and subjects won or lost points according to the value of the stimulus they chose. In the first phase of the experiment (expectation acquisition), subjects learnt how many points each of the six different stimuli were worth. The points value of tokens, positive or negative, is referred to here as reinforcement. In the second phase (violation), the points value of some of the stimuli were changed, i.e., reinforcement violated expectations. Skin conductance responses (SCRs) following reinforcement and token choice behaviour were the dependent variables.

2.2: The use of SCR as an index of autonomic nervous system arousal

The autonomic nervous system (ANS) has two subdivisions: the sympathetic branch and the parasympathetic branch. The sympathetic branch serves what has come to be known as the 'emergency reaction' (Cannon, 1927), or autonomic nervous system arousal. This response comprises pupil dilation, inhibited salivary secretion, accelerated heart rate, increased respiration, increased electrodermal response, inhibition of digestion, secretion of adrenaline and noradrenaline, conversion of glycogen to bile, and inhibition of bladder contraction. ANS arousal is elicited by changes in both the physical and psychological environment. Psychological changes that elicit the ANS arousal response include emotionally significant stimuli, and novel or orienting stimuli. Physical changes include those that will potentially disrupt the balance of any organ. The parasympathetic division serves to reverse these responses since, on the whole, the two branches act antagonistically.

SCR was used as the measure of ANS arousal for three main reasons. First, SCR measurement is non-intrusive and causes no irritation to the subject. Second, the use of SCR enables comparison with other studies that have also chosen SCR as the dependent variable (e.g., Mandler, 1964; MacDowell & Mandler, 1989). Finally, and most importantly, SCR is determined only by the sympathetic branch of the autonomic nervous system, and increases monotonically with intensifying stimulation. SCR is also the best predictor of self-reported psychological arousal (Lang, Greenwald, Bradley, & Hamm, 1993). In contrast, heart rate –which has also been used as an index of autonomic arousal in psychological research - is innervated by both the sympathetic and parasympathetic branches of the ANS, and has a complex relationship with subjective reports of emotional arousal (Lang et al., 1993).

2.3: Experiment 1

It was hypothesised that magnitude of reinforcement would be represented in response outcome expectations. It was therefore predicted that violations of expected magnitude of reinforcement would result in significantly larger SCRs than those following expected reinforcements of the same value. It was also predicted that the detection of magnitude violations would be associated with rapid behavioural change. Specifically, it was predicted that when presented with stimulus combinations for which the correct stimulus to choose changed because of magnitude changes, subjects would show a rapid change in their stimulus choice behaviour.

2.4: Method

2.4.1: Subjects

Subjects were recruited locally and received payment and performance-related chocolate rewards. Since the hypotheses to be tested were concerned with psychophysiological and behavioural responses following expectation violations, only subjects who developed expectations were included in the analyses. Thus two subjects who made fewer than 60% correct token choices in phase 1 (expectation acquisition) of the experiment were excluded from the analyses. This exclusion criterion was also used in Experiments 2 and 3. The remaining 30 subjects comprised 12 male subjects and 18 female subjects. The mean age was 26 years (s.d. = 6).

2.4.2: Apparatus

An IBM-PC computer attached to a VGA colour monitor was used for game presentation, and for the storage of token presentation sequences and subjects' token choices. A MP100WSW Biopac physiological recording system was used with an IBM-PC computer for storage of skin conductance data. The two computers were interfaced.

Relative galvanic skin response was measured using a galvanic skin response amplifier module together with Ag-AgCl finger electrodes attached to the medial phalanges of the first and second fingers of the non-dominant hand. Standard laboratory electrode gel was used as a conductant.

2.4.3: Procedure

A computerised task was used. A brief verbal description of the computer game and the physiological measurement to be taken was given. The subject was then connected to the physiological recording equipment and requested to sit as still as possible. Only the dominant hand was used for computer play. The subject read the game instructions from the computer screen while his or her skin conductance response stabilised. The instructions were as follows:

> *Direct the snake around the field, using the cursor (arrow) keys. The aim of the game is to make your snake eat as many mice as possible.*

> *Tokens (coloured squares) will appear on the screen in pairs. These represent mice that you can win or lose.*
>
> *When the tokens appear, there will be a brief pause in the game. During this pause, you should decide which token you want your snake to eat.*
>
> *Move the snake to the token. When you hit the token, a message will appear telling you how many mice you have won or lost.*
>
> *A total will appear at the top of the screen, telling you how many mice you have won so far in the game. Your snake is very hungry. Try hard to win as many mice as you can!*
>
> *For every 1000 points you win in the game you will win a sweet. But for every 1000 points that you lose in the game, a sweet will be taken away.*

At the beginning of each trial, two small coloured tokens appeared on the screen simultaneously, equidistant from the snake's head (figure 2.1a). The screen then froze for four seconds. The subject was instructed to decide which token they were going to eat during this period. When the screen unfroze, the subject moved the snake to the token of choice using the keyboard cursor keys. A message then appeared on the screen telling the subject how many mice had been won or lost (figure 2.1b).

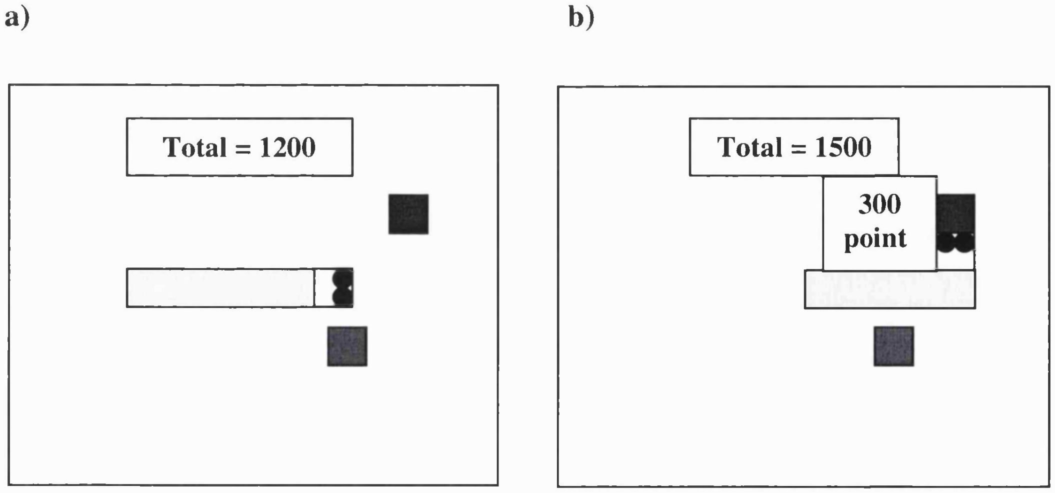

Figure 2.1: Computer screen display (a) before subject token choice and (b) after reinforcement.

The screen was frozen for four seconds, with the reinforcement message, to allow measurement of SCRs following reinforcement. The playing field then cleared for the next trial, and the total score message at the top of the screen was updated. Subjects received a chocolate reward for every 1000 points they won. The chocolate rewards were placed in front of the subject, or removed, as they were won or lost. The game lasted approximately 30 minutes.

Six different colours of token were used. The tokens were presented in pairs in five blocks. Each block contained 21 trials, comprising the 21 possible token combinations (i.e., six same-colour token combinations and 15 different-colour token combinations). There were three phases to the experiment: a familiarisation phase (block 1), an expectation acquisition phase (blocks 2-3), and a violation phase (blocks 4-5). There were no breaks between any of the phases, nor was the subject informed that there were different phases in the experiment. Token pair combinations were selected randomly within each phase of the experiment. In phase 1, each token was associated with a particular value, as shown in the Phase 1 column of Table 2.1. Discrepancy was produced in the second phase by changing the points associated with Tokens 1-4, as shown in the Phase 2 column of table 2.1. For example, Token 1 was worth 100 points in phase 1, and 300 points in phase 2. The control tokens, Tokens 5 and 6, kept the same points values throughout the experiment.

Token	**Phase 1 (expectation acquisition)**	**Phase 2 (violation)**
1 (magnitude change)	100	*300*
2 (magnitude change)	-100	*-300*
3 (magnitude change)	300	*100*
4 (magnitude change)	-300	*-100*
5 (control)	300	300
6 (control)	-300	-300

Table 2.1: Points values of tokens in phase 1 and phase 2 of Experiment 1. Italic text indicates a magnitude change in phase 2.

2.4.4: Data Treatment

Skin conductance amplitude change was calculated for the 1-4 second window after reinforcement onset. Amplitude increases that occurred before or after this time period, and decreases in amplitude, were scored as zero. Following a linear transformation of the addition of 1 to subjects' SCRs, a log to base 10 transformation was performed on this data. All SCR data are presented in microSiemens. Transformed SCRs greater than 3 microSiemens were presumed to be too large to have arisen from the experimental stimuli and were excluded from the analyses. This exclusion criterion was used in all four experiments.

Scoring of token choices in phase 1 was as follows. A trial was scored as "correct" if the subject chose the token with the largest value. A trial was scored as "incorrect" if the subject chose the token with the least value. If the tokens were of equal value, the trial was not scored.

2.5: Results & Discussion

2.5.1: Data Analyses

SCRs before token choices and after reinforcement were measured using a custom-written Matlab program, blind to experimental condition. All analyses were performed using SPSS software.

2.5.2: Behavioural Data

The behavioural exclusion criterion of at least 60% correct token choices in phase 1 guaranteed that all subjects had developed reinforcement expectations, as indexed by token choices. In phase 1 (expectation acquisition), the mean number of correct token choices was 23/26 (s.d. = 3). Figure 2.2 shows subjects' token choice performance categorised in terms of percentage correct choices when a positive and a negative token were presented together (mean = 93%, s.d. = 12), two positive tokens of different values (mean = 79%, s.d. = 28), and two negative tokens of different values (mean = 71%, s.d. = 30). One-sample t-tests showed that performance in all categories of token combinations was greater than chance, $t(29) \geq 3.8$, $p < .005$. Thus, subjects were sensitive to differences in both magnitude and valence between the tokens.

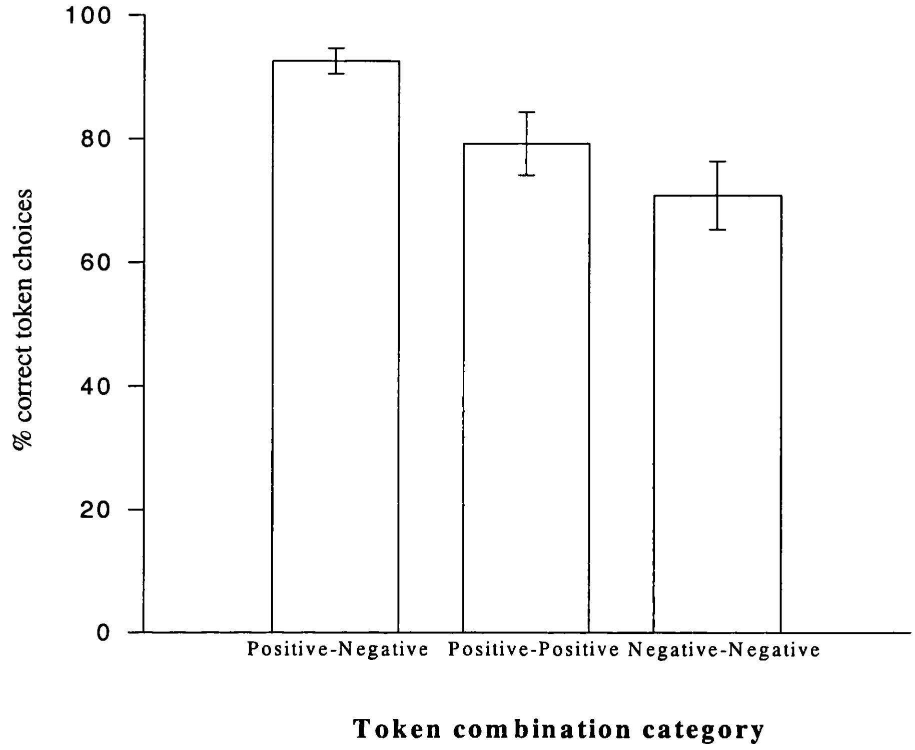

Figure 2.2: Token choice performance (percentage correct) for token combinations consisting of: a positive and a negative token; two positive tokens of different values; two negative tokens of different values.

A one-way repeated measures ANOVA with three-level factor Token Combination (Positive-Negative, Positive-Positive, Negative-Negative) revealed a significant main effect of Token combination, F(1, 29) = 14.9. p = .001. Paired samples t-tests revealed that subjects performed significantly better on Positive-Negative token combinations than on Positive-Positive token combinations, or Negative-Negative token combinations, t(29) ≥ 3.0, p < .01. There was no significant difference between the subjects' performance on Positive-Positive and Negative-Negative token combinations, t(29) = 1.1, p = ns. These findings suggest that subjects learnt whether tokens were positive or negative more reliably or more quickly than they learnt the tokens' specific magnitudes.

2.5.3: SCRs following magnitude changes

It was predicted that changes in the magnitude of reinforcement would result in significantly larger SCRs than those following expected reinforcements of the same value. That is, it was hypothesised that an unexpected reinforcement of 300 points would produce a larger SCR than an expected reinforcement of 300 points. To test this hypothesis, mean SCR following unexpected reinforcements in phase 2 (violation) were compared with mean SCRs following expected reinforcements of the same value in phase 1 (expectation acquisition). To calculate mean SCRs following magnitude changes, SCRs following the first choice of each of the four magnitude change tokens (Tokens 1-4) in phase 2 were used. To calculate mean SCRs following expected reinforcements, SCRs following the last choice of each of the four magnitude change tokens in phase 1 were used. The difference between these two mean SCR values was then calculated. This is referred to as the Magnitude Change difference score. Thus, the Magnitude Change difference score represents the increase in SCR due to violation of expectations. To control for global changes in arousal throughout the course of the experiment, a Control difference score was calculated in the same way as the Magnitude Change difference score. This was done using SCRs following choices of the two control tokens.

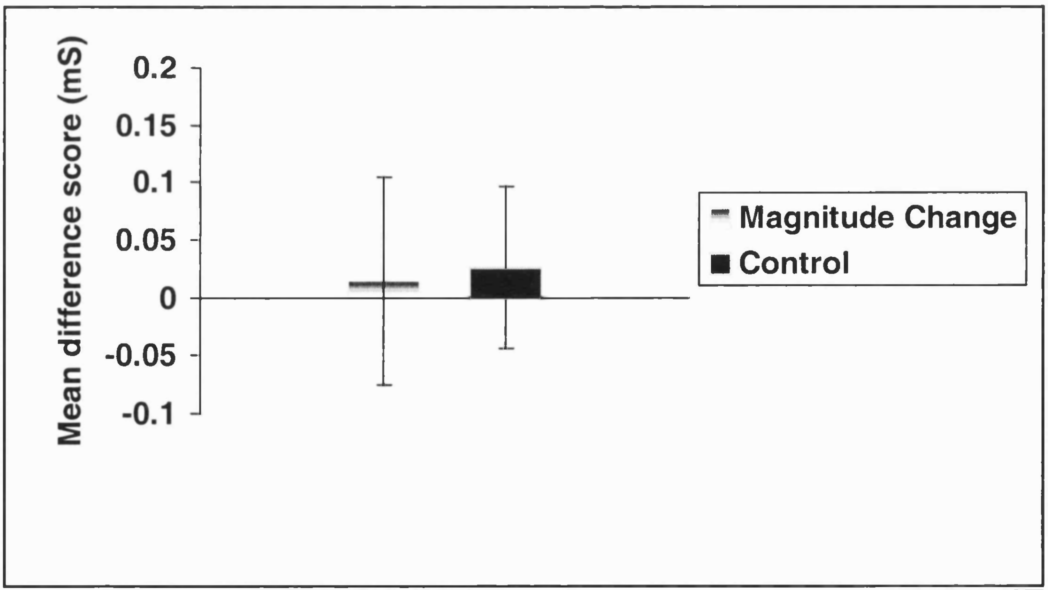

Figure 2.3: Mean Magnitude Change and Control difference scores, Experiment 1.

Magnitude Change and Control difference scores are shown in figure 2.3. The Magnitude Change score (mean = 0.015, s.d. = 0.47) was not significantly different from the Control difference score (mean = 0.026, s.d. = 0.38), t(29) = 0.11, p = ns. In other words, there was no SCR increase due to expectations of the magnitude of reinforcement being violated. This does not support the hypothesis that magnitude is represented in response outcome expectations. However, one possible explanation of the absence of support for this hypothesis is that the change in magnitude was not large enough to produce an arousal increase.

2.5.4: Token choice behaviour following magnitude changes

The next analysis investigated whether subjects adapted their token choices following magnitude changes. In phase 2, there were two token combinations for which the correct token to choose changed. Tokens 1 and 3 changed from 100 to 300 and 300 to 100 respectively. Thus Token 3 was the correct token to choose in the first phase, and Token 1 in the second phase, for that combination. Tokens 2 and 4 changed from -100 to –300 and –300 to –100 respectively. Thus Token 2 was the correct token to choose in the first phase, and Token 4 in the second phase, for that combination.

In order to investigate behavioural change, token choice behaviour for these two token combinations in the first and second blocks of phase 2 were calculated. If magnitude changes trigger rapid and adaptive behavioural change, then token choice performance when Tokens 1 and 3 and Tokens 2 and 4 are presented together for the first time in phase 2 should be better than chance. The dependent variable was frequency of correct token choices for these two token combinations in block 1. The mean frequency of correct token choices was 48% (s.d. = 33). A one-sample t-test revealed that this performance was not significantly better than chance, t(29) = .27, p = ns. In contrast, performance in the second block of phase 2 (mean = 68%, s.d. = 68) was significantly better than chance, t(29) = 2.2, p < .05. It was reasoned that the failure to see behavioural change in block 1 may have been due to subjects not having had the opportunity to learn the new values of the magnitude change tokens before the trials of interest. For example, a subject may have had to make a choice between Tokens 1 and 3 without having observed the new value of Token 1 and/or Token 3. Therefore the analysis for block 1 was repeated, but data points were excluded if a subject had not yet had the opportunity to learn the new reinforcement values of both the tokens. This meant that for 11 subjects, one trial was excluded. For three subjects, both trials were excluded. The mean frequency of correct token choices was 52% (s.d. = 43). A one-sample t-test revealed that this was not better than chance, t(26) = .23, p = ns.

In summary, changes in the expected magnitude of reinforcement had no effect on autonomic arousal response or instrumental behaviour. Magnitude changes did not produce an increase in mean SCR over-and-above the mean SCR to an expected reinforcement of the same value. Furthermore, although behavioural change following magnitude changes would have resulted in winning more points, no rapid behavioural adaptation to the new token values was seen in the first block of trials following the magnitude changes. However, behavioural change did occur in the second block.

2.6: Experiment 2

In Experiment 1 it was found that subjects performed significantly better than chance when deciding between tokens of the same valence but different magnitude. For example, when presented with tokens worth 100 and 300 points, subjects chose the 300 point token nearly 80% of the time. This indicates that subjects were representing both valence and magnitude in the response outcome expectations that guided behaviour. However, magnitude changes did not produce an increase in arousal. It was argued in Chapter 1 that arousal increases following discrepancy in a parameter indicate that that parameter is represented. The failure to find an arousal response following magnitude changes is therefore rather unexpected. It suggests the possibility that the system that represents the response outcome expectations that guide behaviour is not also involved in detecting violations of those expectations. In other words, a second system may detect expectation violations. The results of Experiment 1 suggest the hypothesis that magnitude is not represented in the system that compares expected with actual reinforcement. Experiment 2 tested the prediction that violations of expected valence, but not magnitude, of reinforcement would produce arousal increases and trigger rapid behavioural change. Stimuli that changed in magnitude of reinforcement with the same absolute discrepancy as the valence change stimuli were also included, to investigate whether the null result of Experiment 1 was due to insufficient level of discrepancy.

2.7: Method

2.7.1: Subjects

Subjects received payment for their participation, in addition to performance-related chocolate rewards. Four subjects who made fewer than 60% correct token choices in phase 1 (expectation acquisition) of the experiment were excluded from the analyses. There were 13 male subjects and 17 female subjects with a mean age of 23 years (s.d. = 4).

2.7.2: Procedure

The experimental procedure was similar to Experiment 1. Each colour of token was associated with a certain number of points in both phases (see table 2.2). For example, in phase 1 (expectation acquisition), Token 1 and Token 3 were both associated with 100 points reward, as shown in the Phase 1 column of table 2.2. Discrepancy was produced in the second phase when the points associated with Tokens 1-4 were changed, as shown in the Phase 2 column of table 2.2. Tokens 1 and 2 changed valence and Tokens 3 and 4 changed magnitude. For example, Token 1 was associated with 100 points loss in phase 2 (valence change) and Token 3 was associated with a 300 point reward in phase 2 (magnitude change). The Control tokens 5 and 6 kept the same points values throughout the experiment.

Token	**Phase 1** **(expectation acquisition)**	**Phase 2** **(violation)**
1 (valence change)	100	***-100***
2 (valence change)	-100	***100***
3 (magnitude change)	100	**300**
4 (magnitude change)	-100	**-300**
5 (control)	300	300
6 (control)	-300	-300

Table 2.2: Points values of tokens in phase 1 and phase 2 of Experiment 2. Bold italic text indicates a valence change in phase 2. Bold text indicates a magnitude change in phase 2.

2.8: Results & Discussion

2.8.1: Behavioural Data

The mean number of correct token choices in phase 1 was 23/26 (s.d. = 3). Figure 2.4 shows subjects' token choice performance categorised in terms of percentage correct choices when a positive and a negative token were presented together (mean = 92%, s.d. = 25), two positive tokens of different values (mean = 82%, s.d. = 25), and two negative tokens of different values (mean = 66%, s.d. = 29). One-sample t-tests showed that performance in all categories of token combinations was greater than chance, $t(29) \geq 3$, $p < .01$. Thus, subjects were sensitive to differences in both magnitude and valence between the tokens. A one-way repeated measures ANOVA with three level factor Token Combination (Positive-Negative, Positive-Positive, Negative-Negative) revealed a significant main effect of Token Combination, $F(1, 29) = 30.5$, $p < .001$. Paired samples t-tests revealed that subjects performed significantly better on Positive-Negative token combinations than on both Positive-Positive and Negative-Negative token combinations, $t(29) \geq 2.3$, $p < .05$. In addition, subjects performed significantly better on Positive-Positive token combinations than on Negative-Negative token combinations, $t(29) = 3.3$, $p < .01$. As in Experiment 1, these findings suggest that subjects learnt the valence of tokens more reliably or faster than they learnt the magnitude of reinforcement. In addition, it was found that performance on Positive-Positive token combinations was better than performance on Negative-Negative token combinations. This is probably attributable to the fact that subjects selected positive tokens more frequently than negative tokens, and therefore had more trials to learn the magnitude of positive tokens than they did negative tokens.

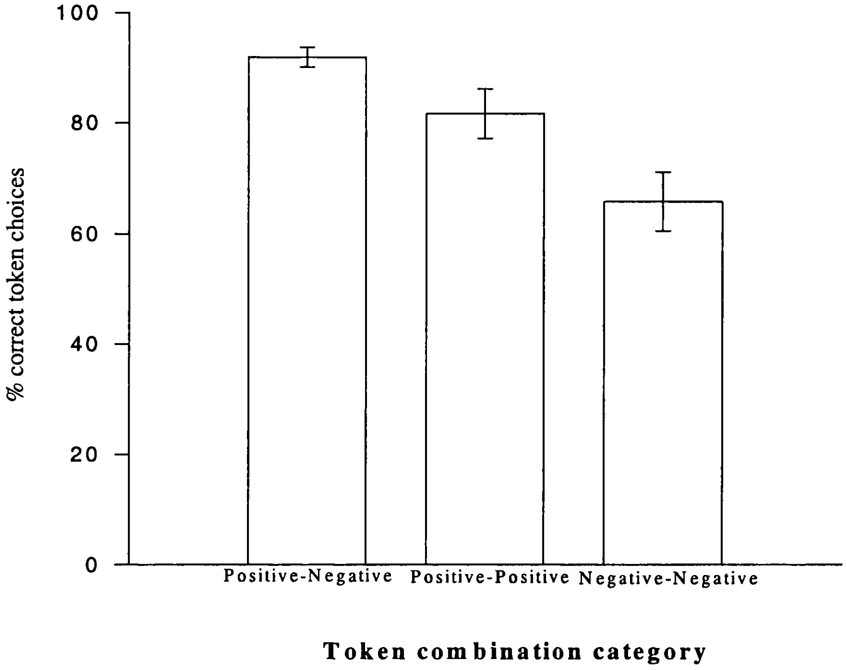

Figure 2.4: Token choice performance (percentage correct) for token combinations consisting of: a positive and a negative token; two positive tokens of different values; two negative tokens of different values. In this and all subsequent graphs, error bars correspond to standard errors.

2.8.2: SCRs following valence changes and magnitude changes

It was predicted that violations of expected valence, but not magnitude, of reinforcement would result in significantly larger SCRs than those following expected reinforcements of the same value. To test this hypothesis, mean SCRs following valence and magnitude changes in phase 2 were compared with mean SCRs following expected reinforcements of the same value in phase 1. The two magnitude change stimuli were worth +/- 100 points in phase 1, but +/-300 points in phase 2. It was therefore not possible to compare SCRs following the magnitude change tokens between the two phases of the experiment because any increase in SCR might be attributable to the larger reinforcement in phase 2. For this reason, SCRs following reinforcement from the magnitude change tokens in phase 2 (+/- 300 points) were compared with SCRs following expected reinforcements of the equivalent points value in phase 2, namely, the control tokens (also +/- 300 points). Mean Valence Change, Magnitude Change, and Control difference scores were calculated in the same way as in Experiment 1 (see section 2.4.3).

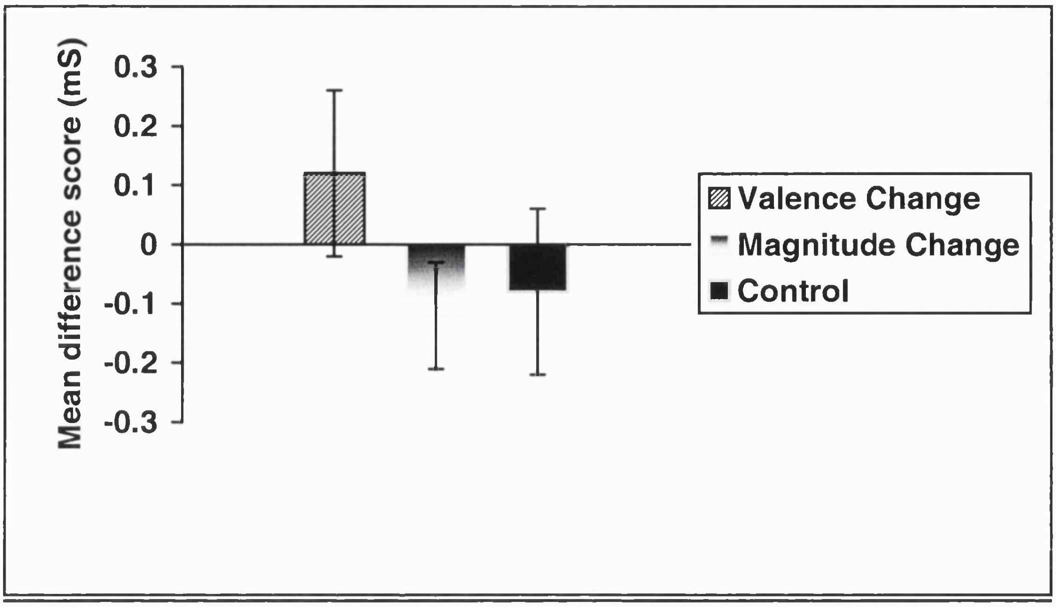

Figure 2.5: Valence Change, Magnitude Change and Control difference scores, Experiment 2.

Figure 2.5 and table 2.3 show the mean Valence Violation, Magnitude Violation, and Control difference scores. It can be seen that only valence changes produced an increase in SCR. For both magnitude changes and control tokens, mean SCRs were smaller in phase 2 than in phase 1. A paired-samples t-test revealed that, as predicted, the mean Valence Change difference score was significantly greater than the mean Magnitude Change difference score, t(29) = 1.65, p = .05; one-tailed. That is, the increase in SCR observed following valence changes was significantly greater than the change following magnitude changes. However, contrary to prediction, the mean Valence Change difference was not significantly greater than the mean Control difference score, t(29) = 1.23, p = ns. The Magnitude Change difference score did not differ significantly to the Control difference score, t(29) = 0.36, p = ns.

Difference score	**Mean in mS (standard deviation)**
Valence Change	0.12 (0.41)
Magnitude Change	-0.12 (0.74)
Control	-0.08 (0.14)

Table 2.3: Mean (standard deviation) Valence Change, Magnitude Change and Control difference scores, Experiment 2.

These findings therefore suggest the prediction that unexpected changes in the valence of reinforcement produce arousal. In addition, as in Experiment 1 magnitude changes did not produce SCRs greater than those seen following expected reinforcements of the same value. Moreover, it is important to note that for both valence and magnitude changes, discrepancy was of 200 points. Thus the absence of SCR increases following magnitude changes was probably not due to insufficient discrepancy of reinforcement. The findings therefore support the hypothesis that only valence is represented in the system that compares actual with expected reinforcement. The results of Experiments 1 and 2 together contrast with those of MacDowell & Mandler (1989), who found that changes in the magnitude of expected reinforcement produced heart rate increases. However, MacDowell & Mandler's findings may have been due to the fact that unexpected reinforcements were of a greater magnitude than the expected reinforcements.

2.8.3: Token choice behaviour following valence violations

The next analysis investigated whether the observed SCR increase following valence violations was associated with an adaptive reversal of behaviour. Tokens 1 and 2 both changed valence in phase 2 (100 to –100 points, and vice versa). Thus for the combination of Tokens 1 and 2, Token 1 was the correct token to choose in phase 1, but Token 2 was the correct token to choose in phase 2.

In order to investigate behavioural change, token choice behaviour for this token combination in the first and second blocks of phase 2 was calculated. If valence changes trigger rapid and adaptive behavioural change, then token choice performance when Tokens 1 and 2 are presented together for the first time in phase 2 should be better than chance. The dependent variable was frequency of correct token choices for these two token combinations in block 1. The mean frequency of correct token choices was 43% (s.d. = 50). A one-sample t-test revealed that this performance was not significantly better than chance, t(29) = .72, p = ns. In contrast, performance in the second block of phase 2 (mean = 66%, s.d. = 49) was significantly better than chance, t(29) = 1.9, $p < .05$; one-tailed. It was reasoned that the failure to see behavioural change in the first block may have been due to subjects not having had the opportunity to learn the new values of the valence change tokens before the trials of interest. For example, a subject may have had to make a choice between Tokens 1 and 2 without having observed the new value of Token 1 and/or Token 2. Therefore the analysis for the first block was repeated, but data points were excluded if a subject had not yet had the opportunity to learn the new reinforcement values of both the tokens. This meant that for 16 subjects, there were no data available. The mean frequency of correct token choices was 57% (s.d. = 51). Contrary to prediction, a one-sample t-test revealed that this performance was not significantly better than chance, t(13) = 0.52, p = ns. Thus there was no evidence of rapid behavioural change following valence violations. However, over half of the subjects' data had to be excluded, thus this null result may have been due to a lack of power.

2.9: Experiment 3

The results of Experiments 1 and 2 suggest that unexpected changes in the valence of reinforcement produce arousal whereas magnitude changes do not. The simplest account of these findings is that the system that detects expectation violations only represents information about the expected valence of reinforcement following a response, with no representation of magnitude. However, a second possible explanation is that reinforcement expectation representations are coded both in terms of valence and magnitude, but that a valence change is necessary to trigger ANS arousal. These two hypotheses differ in the predictions that they make about the relative size of the arousal increase that should be observed following large and small valence violations. The first hypothesis predicts that the size of the valence change will have no effect on the size of the arousal increase. In contrast, the second hypothesis predicts that the size of the valence change will have an effect. Specifically, the greater the size of the change from positive to negative, or vice versa, the greater should be the increase in arousal following the expectation violation (Sokolov, 1960; Mandler, 1964; Nakamura, 1984 (cited in MacDowell & Mandler, 1989)). One aim of Experiment 3 therefore was to measure SCRs following both large and small valence changes to ascertain whether the results support the hypothesis that only valence is represented, or the hypothesis that a change in valence is necessary to trigger arousal but that magnitude is nonetheless represented.

2.10: Method

2.10.1: Subjects

Subjects received payment for their participation, in addition to performance-related chocolate rewards. As in Experiments 1 and 2, four subjects who made fewer than 60% correct token choices in the Expectation Acquisition phase of the experiment were excluded from the analyses. There were 17 male subjects and 13 female subjects with a mean age of 27 years (s.d. = 9).

2.10.2: Procedure

The experimental procedure was similar to Experiments 1 and 2. Each colour of token was associated with a certain number of points in the two phases of the experiment (see table 2.4). Discrepancy was produced in phase 2 by changing the points associated with Tokens 1-4, as shown in the "Phase 2" column of table 2.4. For the small valence change tokens, Tokens 1 and 2, there was a change in reinforcement of +/- 200 points in phase 2. For example, Token 1 was associated with 100 points reward in phase 1 and 100 points loss in phase 2. For the large valence change tokens, Tokens 3 and 4, there was a change in reinforcement of +/- 600 points. For example, Token 4 was associated with 300 points loss in phase 1 and 300 points reward in phase 2. The two control tokens, Tokens 5 and 6, kept the same value throughout the experiment.

Token	**Phase 1** **(expectation acquisition)**	**Phase 2** **(violation)**
1 (small valence change)	100	**-100**
2 (small valence change)	-100	**100**
3 (large valence change)	300	***-300***
4 (large valence change)	-300	***300***
5 (control)	300	300
6 (control)	-300	-300

Table 2.4: Points values of tokens in phase 1 and phase 2 of Experiment 3. Bold text indicates a small valence change in phase 2. Bold italic text indicates a large valence change in phase 2.

2.10.3: Data Treatment

Treatment of skin conductance data and scoring of token choices in phase 1 was as for Experiments 1 and 2.

2.11: Results & Discussion

2.11.1 Behavioural Data

In phase 1 (expectation acquisition), the mean number of correct responses was 20/26 (s.d. = 6). Figure 2.6 shows subjects' token choice performance categorised in terms of percentage correct choices when a positive and a negative token were presented together (mean = 90%, s.d. = 12), two positive tokens of different values (mean = 81%, s.d. = 26), and two negative tokens of different values (mean = 69%, s.d. = 28). One-sample t-tests showed that performance in all categories of token combinations was greater than chance, t(29) ≥ 3.8, $p < .005$. Thus, subjects were sensitive to differences in both magnitude and valence between the tokens. A one-way repeated measures ANOVA with three level factor Token Combination (Positive-Negative, Positive-Positive, Negative-Negative) revealed a significant main effect of Token Combination, F(1, 29) = 18.5, $p < .001$. Paired samples t-tests revealed that subjects performed significantly better on Positive-Negative token combinations than on Negative-Negative token combinations, t(29) = 4.3, $p < .001$. The difference between performance on Positive-Negative and Positive-Positive token combinations was close to significance, t(29) = 2.0, $p < .1$, as was the difference in performance between Positive-Positive and Negative-Negative token combinations, t(29) = 1.9, $p < .1$. As in Experiments 1 and 2, these findings suggest that subjects learnt the valence of tokens more reliably or faster than they did the magnitude of reinforcement.

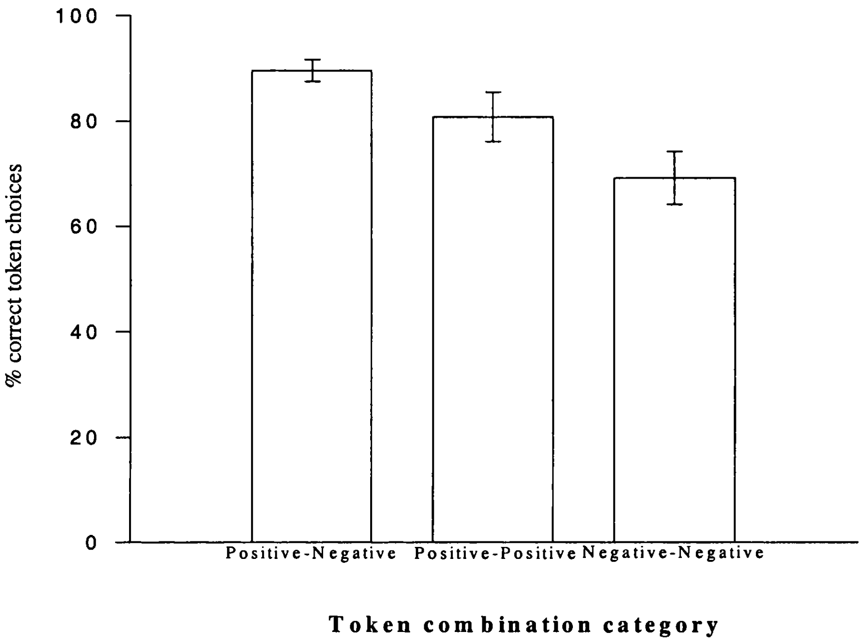

Figure 2.6: Token choice performance (percentage correct) for token combinations consisting of: a positive and a negative token; two positive tokens of different values; two negative tokens of different values.

2.11.2: SCRs following valence changes

The first analysis tested the prediction that SCRs following valence changes would be greater than those following expected reinforcements of the same value. To test this hypothesis, mean SCRs following valence changes in phase 2 were compared with mean SCRs following expected reinforcements of the same value in phase 1. Mean Valence Change, and Control difference scores were calculated in the same way as in Experiments 1 and 2 (see section 2.4.3).

Figure 2.7 and table 2.5 shows the mean Valence Change and Control difference values. As predicted, the Mean Valence Change difference score was significantly greater than the Control difference score, t(29) = 1.65, p = 0.05; one-tailed.

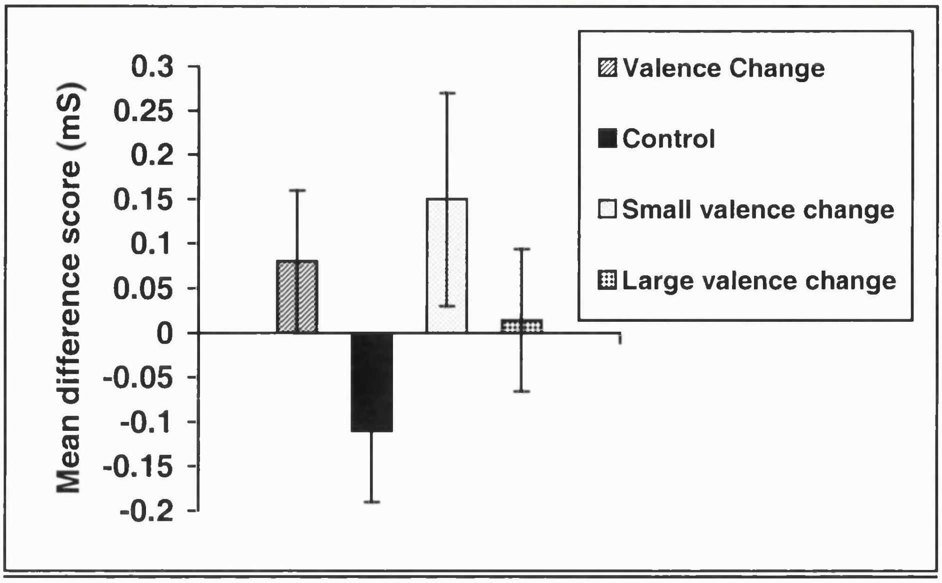

Figure 2.7: Valence Change and Control difference scores, Experiment 3.

The second prediction was that SCR increases following large valence changes would be larger than those following small valence changes. Mean Small Valence Change and Large Valence Change difference scores were calculated in the same way as in Experiments 1 and 2. Means and standard deviations of the Large Valence Change and Small Valence Change difference scores are shown in table 2.5. Mean SCR increase following large valence changes was not significantly greater than mean SCR increase following small valence changes, t(29) = 1.06, p = ns. Indeed, as can been seen in table 2.5 and figure 2.7, mean SCR increase following large valence changes were smaller than the mean SCR following small valence changes.

Difference score	Mean in mS (standard deviation)
Valence Change	0.08 (0.46)
Control	-0.11 (0.46)
Small Valence Change	0.15 (0.66)
Large Valence Change	0.014 (0.46)

Table 2.5: Mean (standard deviation) Valence Change, Control, Small Valence Change and Large Valence Change difference scores, Experiment 3.

Thus, Experiment 3 replicated the finding that valence changes produce SCR increases. The finding that the size of this SCR increase is not influenced by the magnitude of the valence change suggests that magnitude of reinforcement is not represented in the expectations that are compared with actual reinforcement. Of course, it is still possible that an even greater discrepancy of magnitude would trigger an increase in SCR. However, this appears unlikely since in the current experiment, small valence changes resulted in a (non-significantly) greater SCR increase than large valence changes.

2.11.3: Token choice behaviour following valence violations

The next analysis investigated whether the valence changes were associated with an adaptive change in token choice behaviour. Tokens 1 and 2 both changed valence in phase 2 (100 to –100 points, and vice versa), as did Tokens 3 and 4 (300 to –300 points, and vice versa). Thus in phase 1, for the combinations of Tokens 1 and 2, 3 and 4, 1 and 4, and 2 and 3, Tokens 1 and 3 were the correct tokens to choose. However in phase 2, Tokens 2 and 4 were the correct tokens to choose. The dependent variable was the frequency of correct token choices.

In order to investigate behavioural change, token choice behaviour for these token combinations in the first and second blocks of phase 2 were calculated. If valence changes trigger rapid and adaptive behavioural change, then token choice performance when Tokens 1 and 2 are presented together for the first time in phase 2 should be better than chance. The dependent variable was frequency of correct token choices for these two token combinations in block 1. The mean frequency of correct token choices was 62% (s.d. = 24). A one-sample t-test revealed that this performance was significantly better than chance, t(29) = 2.6, p < .05. Performance in the second block of phase 2 (mean = 73%, s.d. = 27) was also significantly better than chance, t(29) = 4.5, p < .001. This finding is therefore consistent with the idea that the processing aim of the mechanism that detects violations of reinforcement expectations is rapid instrumental re-learning. It therefore seems possible that the absence of evidence for instrumental re-learning following valences changes in Experiment 2 may have been due to the small number of trials available for analysis. This arose because in Experiment 2 there was only one token combination for which the correct token to choose changed following valence changes.

It was next investigated whether correct token choices in the first block of phase 2 were more frequent following small valence changes than following large valence changes. It was found that correct token choices were not made significantly more often following presentation of the two large valence tokens (mean = 60%, s.d. = 50) than following presentation of the two small valence tokens (mean = 70%, s.d. = 47), t(29) = .83, p = ns. Thus just as large valence changes did not result in a greater increase in SCR than small valence changes, nor did large valence changes trigger more rapid behavioural change compared with small valence changes.

2.12: Experiment 4

The results of experiments 1-3 suggested that both magnitude and valence are represented in the response outcome expectations that guide behaviour. In contrast, only valence appeared to be represented in the system concerned with comparing response outcome expectations with actual reinforcement. It was suggested that this violation detection system is activated by valence changes because changes in valence generally indicate that a change in response is necessary. This account raises two empirical questions. First, will magnitude changes that unambiguously indicate that a change in response is required produce arousal? Second, will valence changes that do not indicate that a change in response is required produce arousal? The answer to these two questions will lend further insight into the representation of response outcome expectations. For example, it has so far been assumed that response outcome expectations are coded in terms of valence of reinforcement. However, it is also possible that they are coded in terms of motivational significance, that is, whether the stimulus should be approached or avoided. The hypothesis that motivational significance is represented predicts that unexpected reinforcements that unambiguously indicate that a change in response is required should trigger arousal and rapid behavioural change, even if no valence change is involved. The hypothesis that valence is represented predicts that valence changes will trigger arousal, even if they do not indicate that a change in response is necessary. Experiment 4 tested these two predictions.

Predictions were tested using a computer game format in which subjects won or lost points depending on whether they chose to approach or avoid stimuli. SCRs following four types of unexpected reinforcement were measured, in a repeated-measures design:

1. Response Change & Valence Change: valence changes and subjects need to change their response to the stimulus.

2. Response Change: subjects need to change their response to the stimulus but there is no valence change.
3. Valence Change: valence changes but subjects do not need to change their response to the stimulus.
4. Control: there is neither a valence change, nor do subjects need to change their response to the stimulus.

2.13: Method

2.13.1: Design

The experiment involved a 2x2 repeated measures factorial design. The independent variables were Response Change (Response Change vs. no Response Change) and Valence Change (Valence Change vs. no Valence Change). The dependent variable was SCR.

2.13.1: Subjects

There were 30 subjects who received financial payment for their participation. This payment varied from £2 to £5 depending on how well the subject performed in the task. There were 21 female subjects and nine male subjects. The mean age was 25 years (s.d. = 6).

2.13.2: Apparatus

As for Experiments 1-3.

2.13.3: Procedure

The procedure was similar to Experiments 1-3. A computerised task was used. For clarity, the task instructions are given verbatim below:

> *Your small business is failing and you desperately need a short-term cash investment to get it on its feet again. The good news is, you have successfully obtained a £10,000 loan from the bank. The bad news is, you've already borrowed large sums of money from your rich relatives in order to set up the business. They've heard that you have a bank loan and they want to be repaid.*
>
> *You've agreed to pay back your relatives in regular instalments. The size of this instalment is different for each relative. The problem is, the amount that you owe all your relatives is greater than the size of your bank loan. You need to get your business making profit before your bank*

> *loan runs out. This means that if you pay all your relatives back at the agreed rate, your business will definitely go bust.*
>
> *What's the alternative?*
>
> *When the time comes to pay the instalment, you can instead ask that relative to lend you some more money. Sometimes when you do this your relative will take pity on you, and will lend you money. Sometimes the relative will agree to let you pay less than the set instalment level. Sometimes the relative will respond to this request with anger, and force you to pay more than the agreed instalment.*
>
> *Different relatives will respond differently to your plea for more money. It is ESSENTIAL that you learn how your relatives respond so that you know whether it is best to ask them for money, or pay the instalment.*
>
> *At various points in the game you will be given the opportunity to ask any of your relatives for an extra cash bonus. You won't find out how much the relative you choose gives you, if anything. You will therefore have to base your choice on their generosity generally. So make sure you know which of your relatives is the most generous!*
>
> *Remember, every pound is essential in order to save your business, which is your only livelihood. Your starving family are counting on you. As an additional motivation, for every £1000 you have by the end of the game, you will receive £1 payment (minimum payment is £2, maximum payment is £5).*

On each trial, a picture of a relative appeared on the screen (see figure 2.8). Subjects could pay that person the agreed instalment by clicking the button marked, "Pay £x". Alternatively, subjects could ask for money by clicking on the picture of the relative. A message then appeared on the screen, telling the subject how much money that relative had given them, or taken away from them (see figure 2.8). The total score message was updated, and the screen was frozen for four seconds, with the reinforcement message, to allow recording of SCRs following reinforcement. At certain points in the game, every relative appeared on the screen, and the subject was instructed to, "Click on the two relatives that you would like to ask for money". No reinforcement was given following these decisions.

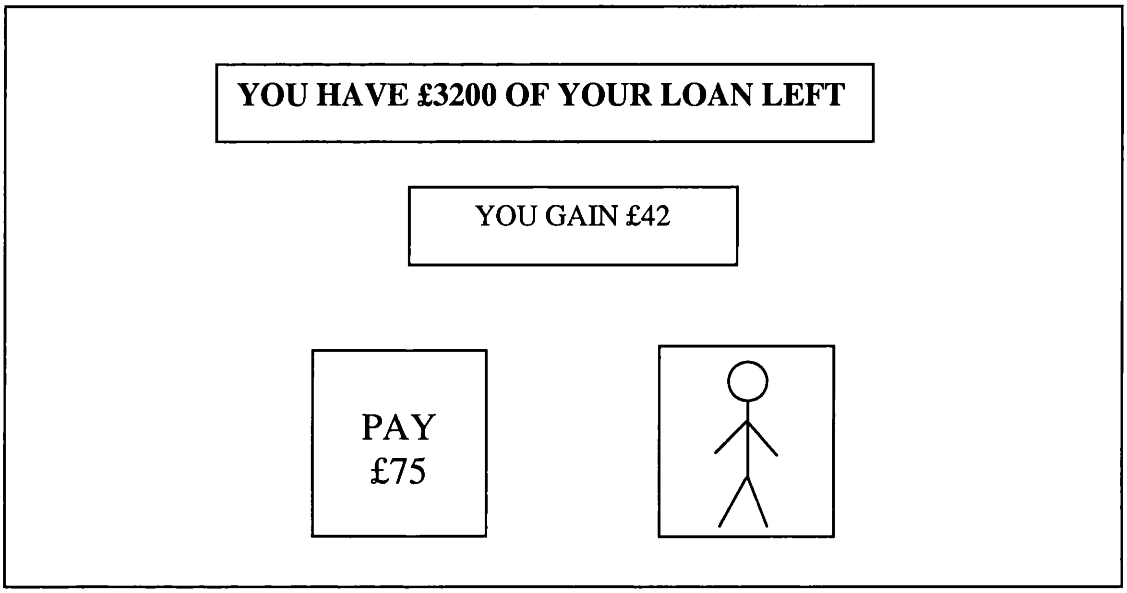

Figure 2.8: Schematic diagram of computer screen layout for Experiment 4

There were eight different relatives in the game, depicted by distinctive bitmap pictures. Relatives were randomly assigned to conditions at the beginning of the game. There were two relatives in each of the four conditions: Response Change & Valence Change; Response Change; Valence Change; and Control. Each relative was associated with a particular instalment payment (-£x), shown in the "Default payment" column of table 2.6. Each relative gave (£x) or took (-£x) a certain amount of money if the subject asked them for money. This is shown in the Phase 1 column of table 2.6. In the game, a random number between –25 and +25 was added to these values. It can be seen that for all relatives, it was better to "ask for money" than to pay the instalment in phase 1.

Condition	**Default payment**	**Reinforcement following "Ask for money" response**	
		Phase 1	Phase 2
RC+VC	-£75	£50	-£150
RC	-£125	-£50	-£150
VC	-£100	-£50	£50
Control	-£200	-£150	-£150

Table 2.6: Default payments, and amounts of money given or taken away by relatives when asked for money, in phases 1 and 2 of Experiment 4. RC+VC = Response Change & Valence Change, RC = Response Change, VC = Valence Change.

The phases were made up of blocks. Each block comprised eight trials. Each relative was presented once every block, in a random order. After the first block, and then every other block, the subject was presented with all eight relatives and asked to click on the two relatives he or she would like to ask for money. The experiment continued to the second phase when the subject had performed correctly in two consecutive blocks (i.e., asked every relative for money), and had chosen the correct two relatives to ask for money. Discrepancy was produced in the second phase by changing how much the relatives gave or took away when the subjects asked them for money, as shown in the Phase 2 column of table 2.6.

It can be seen from table 2.6 that in the Response Change & Valence Change condition, the response of the relatives changed from positive to negative and subjects did better if they changed their response from "ask for money" to paying the instalment. In the Response Change condition, the response of the relatives did not change valence, but subjects did better if they changed their response. In the Valence Change condition, the response of the relatives changed valence, but subjects still did best if they continued to "Ask for money".

2.13: Results & Discussion

2.13.1: Data analyses

SCRs following reinforcement were measured using a custom-written Matlab program, blind to experimental condition. All analyses were performed using SPSS software.

2.13.2: SCRs following Response Change and Valence Change stimuli

It was investigated whether valence change reinforcements and/or response change reinforcements produced SCR increases. Thus SCRs following reinforcement changes were compared with SCRs following expected reinforcements of the same value. For each of the four conditions, the mean SCR following reinforcement from "asking for money" was calculated using the first two blocks of phase 2. In the Response Change & Valence Change, and Response Change conditions, discrepant reinforcement in phase 2 was –£150. Mean SCR for these conditions was therefore compared with mean SCR following the expected reinforcement of –£150 associated with the two Control relatives in phase 1. In the Valence Change condition, discrepant reinforcement in phase 2 was £50. Thus mean SCR following Valence Change was compared with mean SCR following the expected reinforcement of £50 paid out by the two Response Change & Valence Change relatives in phase 1 (see table 2.6).

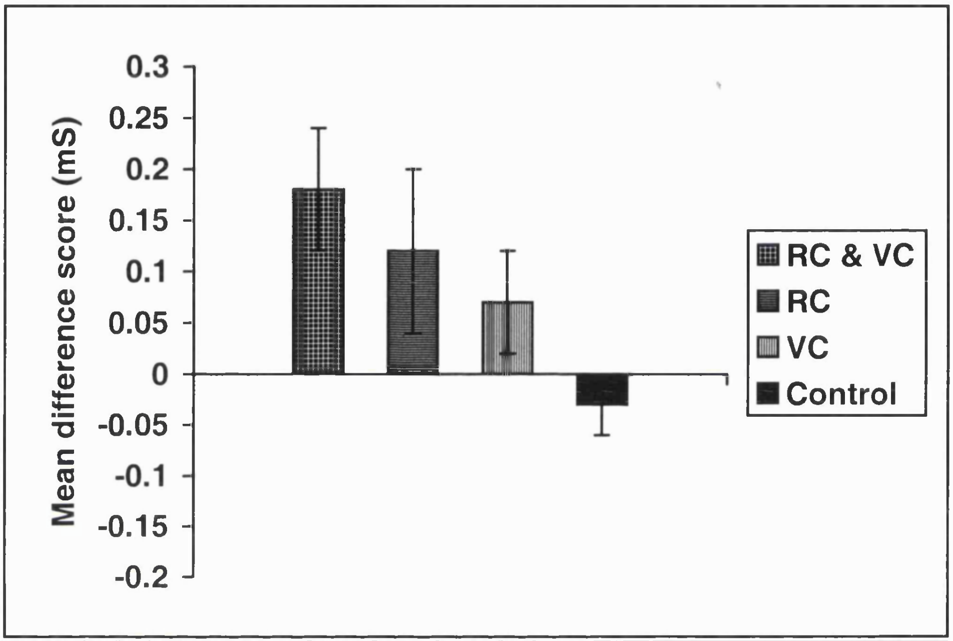

Figure 2.9: Response Change & Valence Change (RC & VC), Response Change (RC), Valence Change (VC), and Control difference scores.

Response Change & Valence Change, Response Change, Valence Change and Control difference scores were calculated. Figure 2.9 shows these difference scores. The data were analysed using a 2x2 repeated measures ANOVA with factors Response Change (Response Change vs. no Response Change) and Valence Change (Valence Change vs. no Valence Change). The main effect of Response Change was significant, F(1, 29) = 4.7, p < .05; one-tailed. There was also a significant main effect of Valence Change, F(1, 29) = 4.0, p < .05; one-tailed. The Valence Change by Response Change interaction was not significant, F(1, 29) = 0.37, p = ns. Paired samples t-test revealed that the Response Change & Valence Change, Response Change, and Valence Change difference scores were all significantly greater than the Control difference score, t(29) > 1.9, p < .05; one-tailed. That is, unexpected information indicating that a change in response was required, and changes in valence triggered SCR increase, both separately and in combination.

It is possible that the SCR increase observed following Response Change reinforcements was due to the magnitude change in reinforcement. A better control condition would have been one in which there was a magnitude change but no need to change response. However, both Experiments 1 and 2 failed to show SCR increases following magnitude changes, thus it is unlikely that the magnitude changes *per se* in the present experiment contributed to the SCR increase.

Thus in summary, the results support the idea that response outcome expectations are coded both in terms of valence, and in terms of motivational significance, that is, whether the stimulus should be approached or avoided.

2.13.3: Behavioural change following Response Change stimuli

It was next investigated whether the SCR increases following Response Change & Valence Change, and Response Change reinforcements were associated with rapid behavioural change. The dependent variable was frequency of (correct) "Pay instalment" responses in the second block of phase 2 (violation). At the beginning of this block, subjects had observed the new reinforcement value of each stimulus once. A one-sample t-test revealed that the frequency of correct responses following the second presentation of the two Response Change & Valence Change stimuli (mean = 60%, s.d. = 40) was significantly greater than zero, t(29) = 8.1, $p < .001$. A second one-sample t-test revealed that the frequency of correct responses following the second presentation of the two Response Change stimuli (mean = 42%, s.d. = 30) was also significantly greater than zero, t(29) = 7.7, $p < .001$. These findings are consistent with the idea that expectation violation detection triggers rapid behavioural change. An unexpected finding was that significantly more correct responses were made in the Response Change & Valence Change condition than in the Response Change only condition, t(29) = 2.2, $p < .05$. Possible implications of this result are discussed in the General Discussion section.

2.14: General Discussion

2.14.1: Summary of findings

The present experiments investigated autonomic arousal responses and instrumental relearning following violations of response outcome reinforcement expectations. As

Sokolov (1960) noted, it is possible to investigate which stimulus parameters are represented in an internal predictive model by manipulating different stimulus parameters and measuring subsequent arousal. An arousal response following manipulation of a certain parameter indicates that that parameter is represented in the model. The current experiments used this methodology to explore how instrumental response outcome expectations are represented.

It was found that valence changes and changes in what appeared to be the motivational significance of a stimulus (i.e., whether it should be approached or avoided) triggered increases in arousal. In contrast, magnitude changes did not influence arousal. It was therefore suggested that the response outcome expectations compared with actual reinforcement are coded in terms of valence and motivational significance but that magnitude information is not represented. In addition it was found that magnitude changes did not trigger rapid instrumental re-learning, whereas valence changes generally did. The association of arousal increases and rapid behavioural change following expectation violations is in line with suggestions that the processing aim of the expectation violation detection system is rapid behavioural change or inhibition (e.g., Gray, 1982; Grossberg, 1982; Mandler, 1984; Rolls, 1990).

2.14.2: Implications for response outcome expectation violation detection

From the pattern of arousal responses following changes in expected reinforcement, it appeared that magnitude information is not represented in response outcome expectations. However, subjects clearly were representing magnitude information in their response outcome expectations, as their development of instrumental responses was sensitive to differences in reinforcement magnitude between the stimuli. The findings are therefore paradoxical unless one proposes two different representational systems. It seems that the system that initially learns response outcome expectations represents these expectations more richly than does the system that detects violations of those expectations. The first system, that supports instrumental learning, represents magnitude and valence information. In contrast, the system that responds to violations of those expectations appears to represent valence, but not magnitude. Interestingly, a dissociation between the mechanism that learns expectations and the mechanism that detects expectation violations is seen in the influential Adaptive

Resonance Theory computational models, developed by Grossberg and colleagues (e.g., Grossberg, 1982; Grossberg & Gutowski, 1987; Grossberg & Levine, 1987; Carpenter & Grossberg, 1988; Grossberg, 2000).

It is possible that larger magnitude manipulations than those used in the current experiments would produce arousal increases. Alternatively, if the stakes had been higher, subjects may have responded autonomically to magnitude changes. However, the results do suggest that there is something different about the processing of valence changes. A change from 100 to –100 results in the same loss of expected winning as a change from 300 to 100, yet only the former condition produced an arousal response.

2.14.3: Implications for the processing aim of the expectation violation detection mechanism

Why might only valence be represented in the expectation violation detection mechanism? One possible answer lies in the stability-plasticity dilemma. Grossberg (e.g., Grossberg, 1982; Carpenter & Grossberg, 1988) has argued that every learning system must solve the stability-plasticity dilemma. This is the problem of maintaining a balance between maintaining stable representations in the face of irrelevant fluctuations in the environment, yet avoiding perseveration as a result of ignoring important changes in the environment. Grossberg has demonstrated that the stability-plasticity dilemma can be successfully resolved in a computational model in which input stimuli that approximately match predictions allow slow expectation learning, whereas input stimuli that badly mismatch predictions result in fast re-learning (see Grossberg, 1982; Grossberg & Gutowski, 1987; Grossberg & Levine, 1987; Grossberg & Schmajuk, 1987; Carpenter & Grossberg, 1988; Grossberg, 2000). The current experiments provide an insight into what constitute "matches" and "mismatches" of expectations. The findings suggest that the neural circuits that underlie instrumental behaviour may have evolved such that variations in magnitude contribute to slow learning of predictions, whereas variations in valence trigger mismatches and rapid re-learning of expectations.

2.14.4: Multiple representations of reinforcement?

One finding did not fit with the hypothesis that valence is represented in the expectation violation detection system. This was the finding that magnitude changes that indicated that a change in response was required also triggered arousal.[1] This finding suggested the idea that response outcome expectations are also coded in terms of their motivational significance, that is, whether they should be approached or avoided. It is not known whether the stimuli in Experiments 1-3 were also represented in terms of their motivational significance. It is possible that they were not, as even the most positive and negative stimuli were frequently paired with tokens of the same value. Thus it is possible that no stimulus was ever strongly associated with an approach or an avoid response.

In summary, it is proposed that the findings from the four experiments reported here suggest the existence of three distinct systems for the representation of reinforcement expectations. First, an instrumental learning system represents all sensory aspects of the expected response outcome, and guides instrumental learning. This is supported by demonstrated ability of subjects to successfully choose between tokens of the same valence but different magnitude (e.g., choosing a token worth 300 points in preference to a token worth 100 points). Second, an instrumental re-learning system compares expected valence of reinforcement with actual reinforcement. Detection of unexpected valence of reinforcement results in rapid re-learning of expected response outcomes in the instrumental learning system. This idea is consistent with the association of arousal increases and behavioural change following valence changes. A third system represents the motivational significance of the stimulus. This was suggested by the finding of an arousal increase following unexpected reinforcements that did not change valence but indicated that a change in response was required. Interestingly, inspection of figure 2.9 shows that the arousal increase following violations of both response outcome and motivational significance expectations (that is, Valence Change and Response Change) was approximately equal to the sum of

[1] It is possible that in this experiment, reinforcements following responses were coded in a relative rather than absolute fashion. For example, a loss of £50, when compared with the alternative response outcome of a loss of £100, may have been coded as positive rather than negative. However, if response outcomes were being coded in a relative fashion, then valence changes that did not require a change in response should not have produced arousal, yet they did. This suggests that absolute valence was being represented.

the arousal increases following violations of response outcome and motivational significance expectations separately. This is also suggestive of two separate systems contributing to arousal in the Valence Change and Response Change condition.

This interpretation of the current findings is clearly speculative. However, it is consistent with recent animal lesion work that suggests the existence of more than one system that represents reinforcement expectations. For example, Robbins and colleagues have found evidence that one neural circuit directs instrumental behaviour while a second separable system mediates reflexive, automatic conditioned responses (Killcross, Robbins, & Everitt, 1997; Parkinson, Robbins, & Everitt, 2000). The instrumental learning system expectation representations hypothesised here may correspond to the neural circuit thought to mediate instrumental learning identified by Robbins and colleagues. The representation of motivational significance may perhaps correspond to a neural circuit including the anterior cingulate and nucleus accumbens core that Everitt, Cardinal, Hall, Parkinson & Robbins (2000) have proposed is involved in giving direction to behavioural responding.

2.14.5: Expectation violation detection and instrumental re-learning

Rapid instrumental re-learning was always associated with an arousal increase in the current experiments. This is consistent with the hypothesis that the role of expectation violation detection is to trigger rapid response change, although it does not directly support the hypothesis. Both the instrumental re-learning and the motivational significance systems would be plausible candidates for a role in rapid instrumental re-learning. Detection of an expectation violation detection within the instrumental re-learning system would be important for rapid re-learning of the response outcomes expectations thought to guide instrumental learning. Detection of an expectation violation within the motivational significance system might also be important for a rapid adaptation of approach or avoidance behaviour. In fact, the behavioural data suggest that both systems may have been contributing towards rapid behavioural change. This is suggested by the finding in Experiment 4 that while valence changes were not necessary to evoke rapid instrumental re-learning change, they appeared to facilitate re-learning.

2.14.6: Implications for models of emotional learning

Several models of emotional behaviour, some of which were discussed in Chapter 1, implicitly or explicitly propose an expectation violation detection mechanism, that is, a mechanism that compares expected response outcome with actual reinforcement (e.g., Gray, 1982; Grossberg, 1982; Mandler, 1984; Oatley & Johnson-Laird, 1987; Rolls, 1990; Amsel, 1992; Schmajuk, Lam, & Gray, 1996). The present data suggest that magnitude is not represented in the system that detects violations of response outcome expectations.

2.14.7: Summary

It is proposed that there are at least three dissociable systems concerned with representing response outcome expectations. The instrumental learning system develops response outcome expectations that guide instrumental behaviour. These expectations represent both valence and magnitude (and probably more generally the nature of the reinforcement, e.g., food, water, shock). The second instrumental re-learning system compares response outcome expectations with actual reinforcement. Only valence is represented. Violations of expectation result in rapid re-learning of response outcome expectations in the instrumental learning system. In addition, a third system, the motivational learning system, represents the motivational significance of the stimulus – whether it should be approached or avoided. These three systems and their proposed interaction are shown below in figure 2.10.

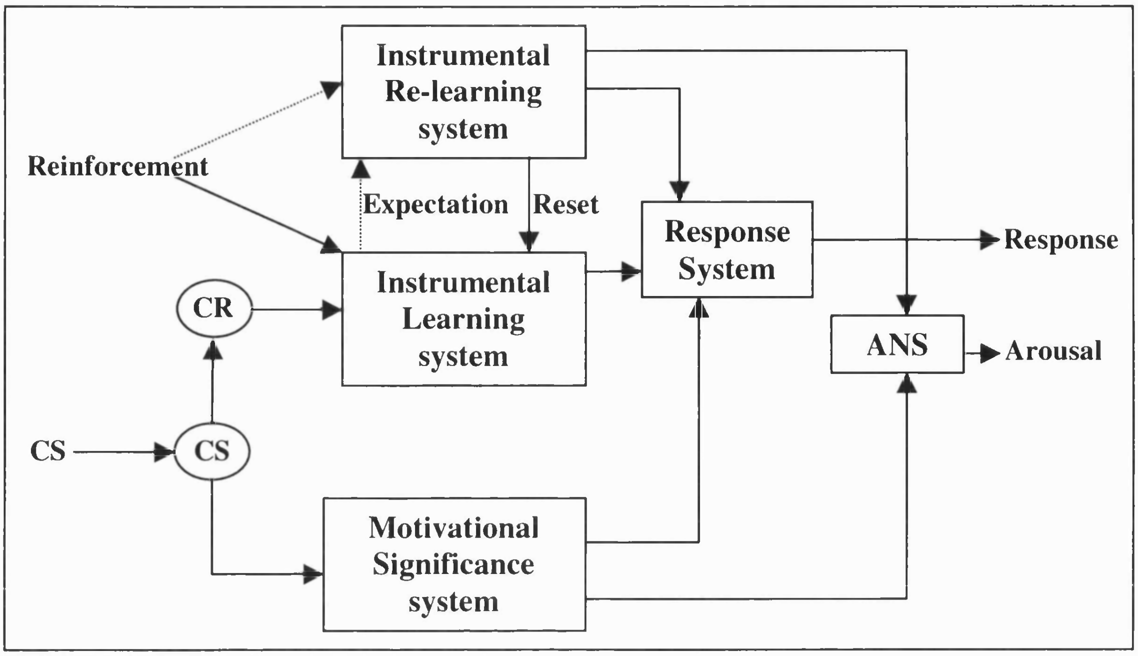
Instrumental Re-learning system
Reinforcement
Expectation
Reset
Response System
Response
Instrumental Learning system
CR
ANS
Arousal
CS
CS
Motivational Significance system

Figure 2.10: Schematic diagram of the Instrumental Learning system, the Instrumental Re-learning system, and the interaction between the two. Dotted line indicates that only valence information is represented. "Reset" connection between the two systems indicates that the Instrumental Re-learning system is involved in changing inaccurate response outcome expectations in the Instrumental Learning system, following valence violations. The Motivational Significance system is also shown. It is not clear how this system receives feedback on response outcomes. CS = conditioned stimulus, CR = conditioned response, ANS = autonomic nervous system.

The instrumental learning and re-learning systems are of primary interest in this thesis. These two systems are implemented in a computational model in the next chapter.

Chapter 3

The Valence Change Reset model: A computational model of instrumental learning and re-learning

3.1: Introduction

In this chapter, the hypotheses proposed in Chapter 2 are implemented in a computational model of instrumental learning and re-learning; the Valence Change Reset model. This model is unique in that unexpected changes in the valence of reinforcement are processed differently to unexpected changes in the magnitude of reinforcement. This chapter shows that the Valence Change Reset model can simulate the human behavioural data from Experiments 1-3 in Chapter 2.

3.2: Computational models of emotional learning

The current model follows the tenet of emotional learning proposed by Rescorla & Wagner (1972), that: *"organisms only learn when events violate expectations."* Rescorla et al., 1972: p75. The hypothesis that learning is proportional to the discrepancy between predicted and actual events is expressed mathematically in the three major computational models of emotional learning (Rescorla & Wagner, 1972; Pearce & Hall, 1980; Mackintosh, 1975). These models differ with regard to how expectation violations affect learning. Using the terminology of the current model, the equation:

$$\Delta w_x = \varepsilon_x \, [f(r) - w_x]$$

expresses the hypothesis that Δw_x, the amount that a system learns about stimulus x and its association with reinforcement, f(r), is inversely related to how well reinforcement, f(r), is predicted by w_x, the learnt association between stimulus x and f(r). Here, ε_x is termed associability (α in some models), that is, the extent to which stimulus x is processed.

Rescorla & Wagner (1972), Pearce & Hall (1980) and Mackintosh (1975) have all hypothesised that learning is partially determined by the discrepancy between expected and received reinforcement. However, these models differ in two main ways. The first difference between the models lies in the associability of stimulus x,

ε_x. The second difference arises from how the models formalise learning when more than one stimulus is involved in each trial of associative learning (see Dickinson, 1980). However, most of the distinctions between the models were not applicable to the learning situation simulated in this chapter. This is because there was only ever one stimulus present during learning. However, one aspect of the Pearce & Hall (1980) rule was relevant. In their model, the main way in which expectation violations affect learning is by changing the associability, ε_x. They suggested that the degree to which stimulus x is processed on a particular trial is the extent to which it was paired with unexpected reinforcement on the previous trial. This hypothesis was implemented in the current model.

Thus the current model arises from previous work in which emotional learning is proposed to be driven by differences between predicted and actual events (Rescorla & Wagner, 1972; Pearce & Hall, 1980; Mackintosh, 1975). It should be noted however that the learning rules proposed in these models concern prediction learning, or classical conditioning. Since the data to be explained here are instrumental responses, it was necessary to relate predictions to instrumental responses. To do this, the conventional step was taken of choosing responses probabilistically, with probabilities dependent directly on the relative predictions of reward (Sutton & Barto, 1998). The Valence Change Reset model also diverged from the classical conditioning models in one other important respect. In the Valence Change Reset model, unpredicted valences of reinforcement were processed differently from unpredicted magnitudes of reinforcement. In terms of the equations presented above, the effect of an unexpected valence of reinforcement was not to decrease or increase w_x, but to set $w_x = 0$.

3.3: Implementation of hypotheses in the connectionist model

3.3.1: Overview of model

In the current model, two interacting systems are proposed to underlie instrumental learning and re-learning processes (see figure 3.1):

1. The Instrumental Learning system

This system represents both valence and magnitude. It learns expected response outcomes. It projects information about expected valence (but not magnitude) to the Instrumental Re-learning system.

2. The Instrumental Re-learning system

This system receives information about the expected and received valence of reinforcement. However, it does not represent magnitude. When an unexpected valence is detected, the ANS is triggered. Projections to the Instrumental Learning system reset the expectations associated with that conditioned response.

Connectionist modelling provides a means for testing whether an explicit implementation of these two systems can successfully simulate the behavioural data presented in Chapter 2.

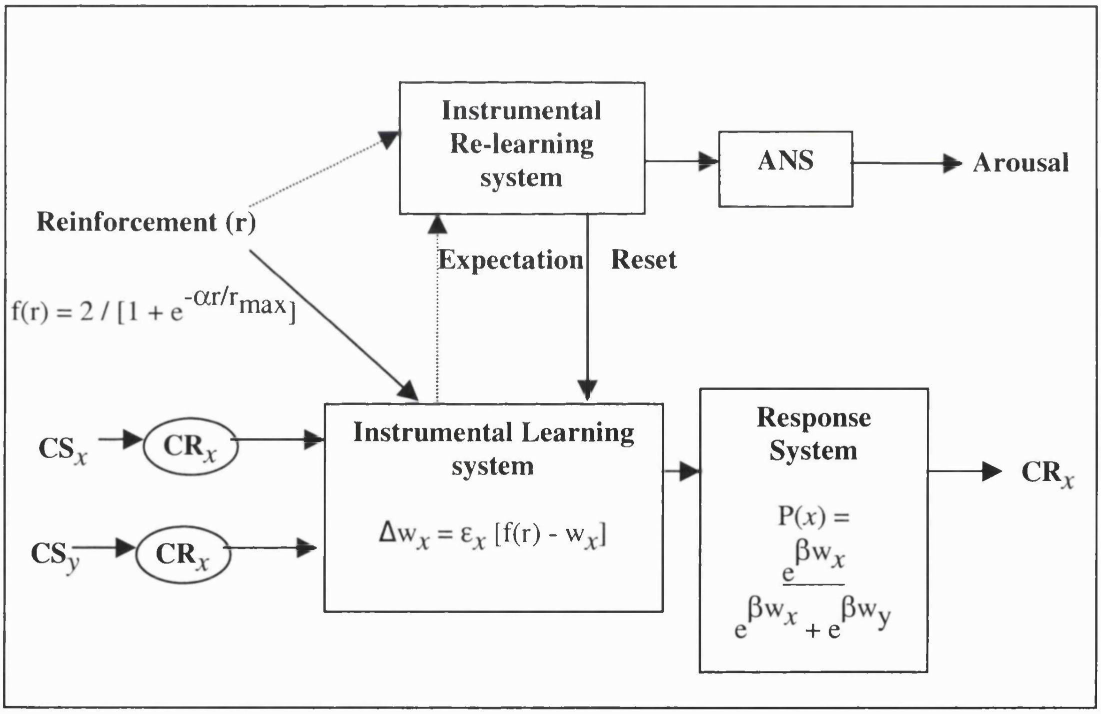
Instrumental Re-learning system
ANS
Arousal
Reinforcement (r)
$f(r) = 2 / [1 + e^{-\alpha r/r_{max}}]$
Expectation
Reset
CS_x
CR_x
CS_y
CR_x
Instrumental Learning system
$\Delta w_x = \varepsilon_x [f(r) - w_x]$
Response System
$P(x) = \frac{e^{\beta w_x}}{e^{\beta w_x} + e^{\beta w_y}}$
CR_x

Figure 3.1: Schematic overview of connectionist model of instrumental learning and re-learning. CS_x = Token x, CS_y = Token y. CR_x = selection of Token x,. CR_y = selection of Token y. Dotted arrow means that only valence information is represented. See text for further details.

3.3.2: The Instrumental Learning system:

A schematic diagram of the Instrumental Learning system of the model is shown in figure 3.2. Six units, T1- T6, correspond to representations of the six tokens in the experimental task. On each trial, two of these token representations are activated (corresponding to the two tokens that are presented in the game). The strength of the activation is determined by the token unit's weight, $w_1 - w_6$. The size and sign of each weight reflects the previous reinforcement history of the token. That is, a positive weight indicates that the choice of that token has been associated with reward more than punishment, and a negative weight indicates that the token has been associated with punishment more than reward.

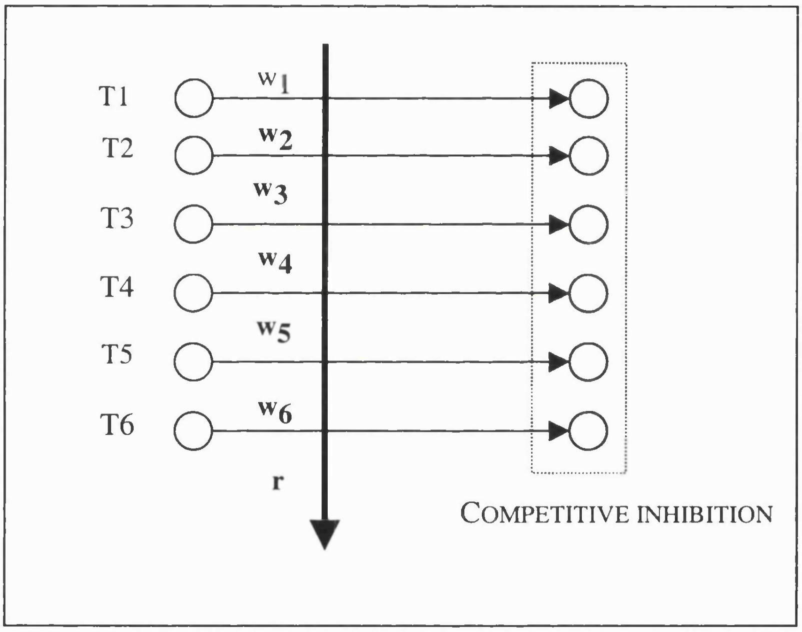

Figure 3.2: Schematic diagram of Instrumental Learning system.

3.2.2.1: Decision rule

In order for the model to select a token, only one of the two token units can remain activated. This can be achieved by using mutual inhibition between the two activated units and self-excitation of each unit such that the activity of the least activated unit is rapidly inhibited. A common method for implementing this competitive process is the hardmax / softmax activation function (see Bishop, 1995). Here, the firing rate of each unit is specified as an exponential function of the activation, scaled by the sum of all activations. Because the activation function increases greater than linearly, following competition the activation of initially more activated units increases relatively much faster than the less activated units. Thus, the probability of the system choosing Token x when presented with Tokens x and y is a function of the relative size of the weights of x and y, and parameter β:

$$P(x) = \frac{e^{\beta w_x}}{e^{\beta w_x} + e^{\beta w_y}}$$

It can be seen that when $w_x = w_y$, the probability of choosing Token x = 0.5. Where $w_x \neq w_y$, β determines the strength of the competition between the two units. As β increases, the competition becomes "winner-take-all", that is, the size of the difference between the two weights needed for the model to reliably choose the token with the largest weight approaches zero.

3.3.2.2: Reinforcement function:

The winning unit then receives input in the form of reinforcement activity. The reinforcement function scales this activity between +/-1 (see figure 3.3). Specifically, activity is scaled such that the maximum number of points in the game results in an activity of +1, and the minimum number of points in the game results in an activity of –1. A transformed sigmoidal reinforcement function was used:

$$f(r) = 2 / [1 + e^{-\alpha r / r_{max}}]$$

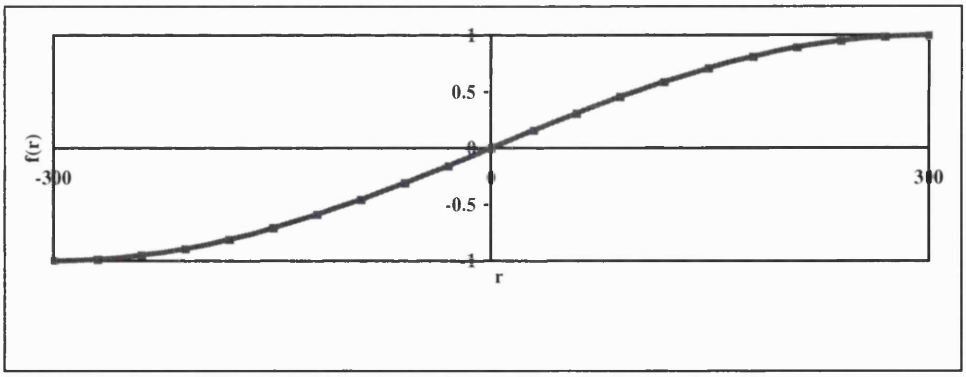

Figure 3.3: Transformed sigmoid reinforcement function

The variable r is the reinforcement received in the game. The parameter α determines the slope of the curve: as α increases, so too does the gradient of the slope. Thus α has to be set to maximise sensitivity in the relevant reinforcement range. The variable r_{max} is the maximum absolute reinforcement that the model expects to receive. At the beginning of the experiment, the model has no information about what the absolute maximum number of points in the game will be. r_{max} is therefore initially set to an arbitrary value. r_{max} approaches its true value using the following rule, where maximum_absolute_value (r) is the maximum absolute reinforcement received so far in the game:

$$\Delta r_{max} = k\ (\text{maximum\_absolute\_value}\ (r) - r_{max})$$

The variable k determines the rate at which r_{max} approaches maximum_absolute_value (r).

3.3.2.3: Learning rule

Following Rescorla & Wagner (1972), weight changes are specified by a learning rule that updates the weight according to the difference between received reinforcement, f(r), and expected reinforcement. Since the size and sign of a weight reflects the previous reinforcement history of a token unit, expected reinforcement is represented by the activated token unit's weight. Thus weight changes are specified by:

$$\Delta w_x = \varepsilon_x\ [f(r) - w_x]$$

It can be seen that underestimation of reinforcement will result in an increase in the weight, and an overestimation of the reinforcement will result in a decrease in the weight. The parameter ε_x determines the rate of learning and, following Pearce & Hall (1980), varies according to how well the model is able to predict reinforcement from the occurrence of the conditioned stimulus. Thus on the nth trial of stimulus x:

$$\varepsilon_{x_n} = m\ |f(r)_{(n-1)} - w_{x(n-1)}|$$

where m is a constant.

3.3.3: Instrumental Re-learning system

The Instrumental Re-learning system receives information about the valence of expected and received reinforcement. The strength of the expectation is reflected in the size of the learning rate, ε_x. If expectation is high (i.e., ε_x is small) and expected and received reinforcement are of different valences, then the Instrumental Re-learning system is triggered. In the model, the Instrumental Re-learning system is only triggered if $\varepsilon_x \leq m$. The effect of this triggering is to erase previous response outcome expectations for that conditioned stimulus. This is achieved by re-setting the weight for the activated token unit to zero. It is hypothesised that the arousal response recorded in the human experiments reflects the action of the Instrumental Re-learning system. Detecting a violation of valence essentially requires the same processing as providing a solution to exclusive OR (XOR) (see, for example, Rolls & Treves, 1998: p40). This is because there should only be an output when the two inputs are different.

In the following sections, the model is used to attempt to simulate the human behavioural data from Experiments 1-3 in Chapter 2.

3.4: Method

3.4.1: Procedure and model parameter values

The same model was used in each of the three experiments. For each experiment, 30 simulations were run. On each simulation, pre-learning weights were set randomly to values between 0.4 and 0.6. For the simulations reported here, parameter values were:

Slope of reinforcement function, $\alpha = 0.5$ ($\alpha / r_{max} = 0.0017$)
Strength of softmax activation function, $\beta = 10$
Learning rate, $m = 0.5$
Rate of change of r_{max}, $k = 0.9$
Initial maximum reward expected, $r_{max} = 10000$

The robustness of the model to changes in the five parameters was explored. One parameter value was manipulated, while the other parameters were held constant at the values reported above. It was investigated over what range of values the model performed within one standard deviation of the human performance for the three token combination categories (positive-negative, positive-positive, and negative-negative). α, which determined the slope of the reinforcement function, could take values between 0.1 and 0.7. When $\alpha < 0.1$, learning was too slow, and when $\alpha > 0.7$, performance on positive-negative token combinations was at ceiling. β, which determined the strength of the softmax activation function, could take values between 6 and 12. m, the learning rate, could take values between 0.3 and 0.95. k, the rate of change of r_{max}, had a lower limit of 0.4. The value of r_{max}, the initial maximum reward expected, did not appear to affect the performance of the model significantly. The model performed within 1 s.d. of the human data with values from 1 to 1 x 10^9. Thus successful performance of the model did not appear to be dependent on a very narrow range of parameter values.

3.4.2: Data analysis

Scoring of token choices in phase 1 (expectation acquisition) and phase 2 (violation) was identical to the scoring procedure used in Experiments 1-3, Chapter 2. A trial was scored as correct if the model "chose" the token with the largest value. A trial was scored as incorrect if the model "chose" the token with the least value. If the tokens were of equal value, the trial was not scored. In phase 2, if the model had not had the opportunity to learn the new reinforcement value associated with both tokens, the trial was excluded from the analysis.

3.5: Results and Discussion

3.5.1: Results of simulations of Experiment 1

In phase 1, the mean number of correct token choices was 23/26 (s.d. = 2). Figure 3.4 shows the model's token choice performance categorised in terms of percentage correct choices when a positive and a negative token were presented together (mean = 95%, s.d. = 4), two positive tokens of different values (mean = 78%, s.d. = 21), and two negative tokens of different values (mean = 61%, s.d. = 25).

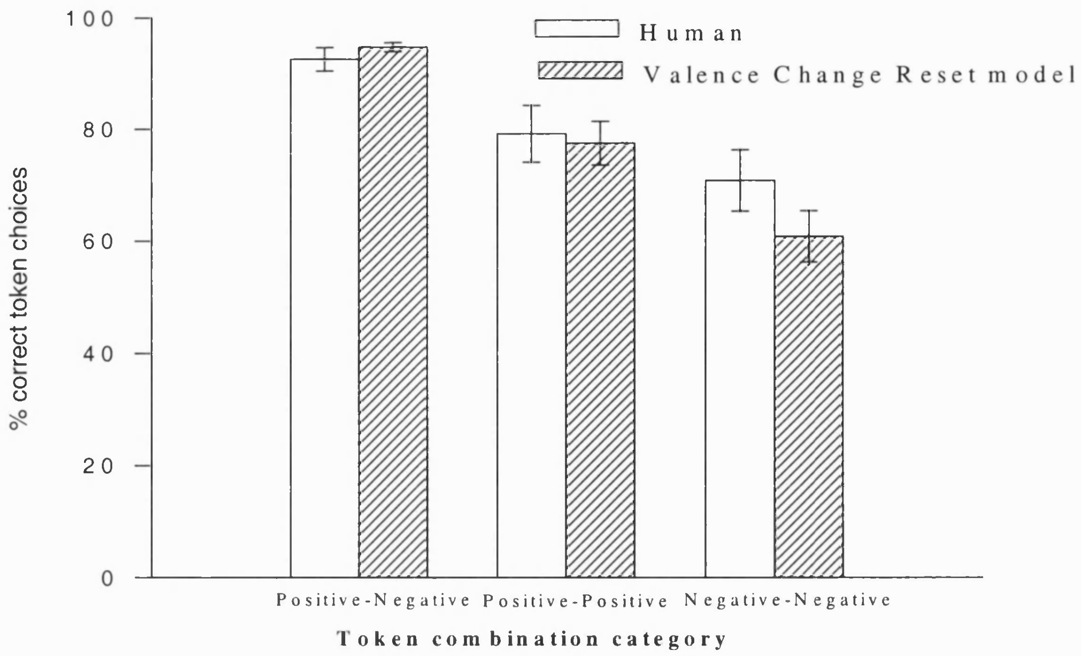

Figure 3.4: Performance of model and humans in phase 1 of Experiment 1.

Human performance is also shown in figure 3.4, for comparison. Model and human performance was compared using a split plot ANOVA with within-subjects factor Token Combination (Positive-Negative, Positive-Positive, Negative-Negative) and between-subjects factor Group (model vs. human). This revealed a significant main effect of Token Combination, F(1, 58) = 58.8, p < .001, reflecting the fact that there were differences in how successful subjects and the model were for the three different token combination categories. The main effect of Group was not significant, F(1, 58) = 0.76, p = ns, and the Token Combination by Group interaction was also not significant, F(1, 58) = 2.8, p = ns. These findings demonstrate that the model successfully simulated human performance in phase 1 (expectation acquisition) – the two populations could not be distinguished statistically.

The next analysis investigated whether the model adapted its token choices following magnitude changes. It was investigated whether, like humans, performance in phase 2 would not be significantly better than chance for token combinations in which the correct token to choose had changed. In phase 2, there were two token combinations for which the correct token to choose changed. Tokens 1 and 3 changed from 100 to 300 and 300 to 100 respectively. Thus Token 3 was the correct token to choose in phase 1, and Token 1 in phase 2, for that combination. Tokens 2 and 4 changed from

-100 to –300 and –300 to –100 respectively. Thus Token 2 was the correct token to choose in phase 1, and Token 4 in phase 2, for that combination.

The dependent variable was the frequency of correct token choices in the first and second blocks of phase 2. Data points were excluded if on a simulation the model had not yet had the opportunity to learn the new reinforcement values of both the tokens. This meant that for the model, in 18 simulations one trial was excluded. For five simulations, both trials were excluded. The mean frequency of correct token choices was 30% (s.d. = 41). This performance was significantly worse than chance, t(24) = 2.4, $p < .05$. An independent samples t-test revealed that the model's performance was not significantly different to the performance of the human subjects, t(50) = 1.9, p = ns. Thus, as with the human subjects, the model was sensitive to magnitude differences in the initial learning phase of the experiment. However, neither the model nor the humans responded behaviourally to magnitude changes, although there was a non-significant trend for the humans to perform better than the model.

3.5.2: Results of simulations of Experiment 2

In phase 2, the mean number of correct token choices was 22/26 (s.d. = 2). Figure 3.5 shows the model's token choice performance categorised in terms of percentage correct choices when a positive and a negative token were presented together (mean = 92%, s.d. = 7), two positive tokens of different values (mean = 80%, s.d. = 20), and two negative tokens of different values (mean = 65%, s.d. = 25). Human performance is also shown in figure 3.5, for comparison. Model and human performance was compared using a split plot ANOVA with within-subjects factor Token Combination (Positive-Negative, Positive-Positive, Negative-Negative) and between-subjects factor Group (model vs. human). This revealed a significant main effect of Token Combination, F(1, 58) = 59.0, $p < .001$, reflecting the fact that there were differences in how successful subjects and the model were for the three different token combination categories. The main effect of Group was not significant, F(1, 58) = .04, p = ns, and the Token Combination by Group interaction was also not significant, F(1, 58) = .03, p = ns. These findings demonstrate that the model successfully simulated human performance in phase 1 – the two populations could not be distinguished statistically.

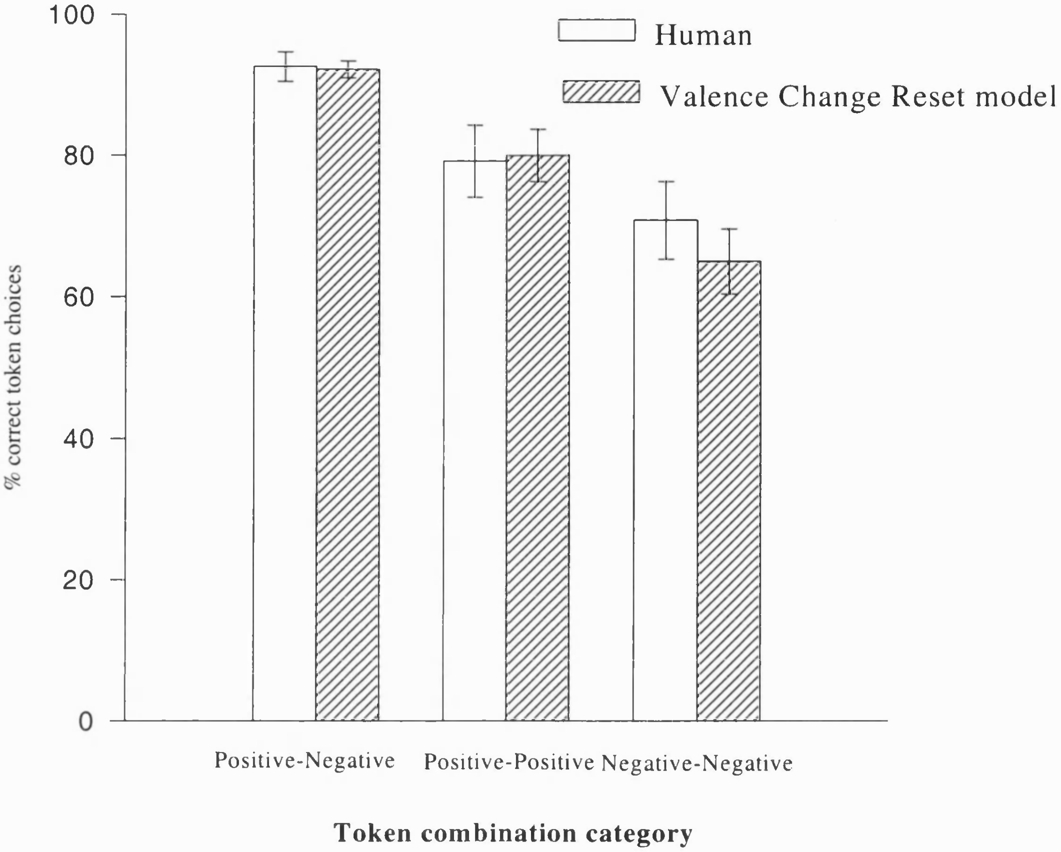

Figure 3.5: Model and human performance in phase 1, Experiment 2.

The next analysis investigated whether the model adapted its token choices following valence changes. In humans, performance in phase 2 for these token combinations was not significantly better than chance, but over half of the data points had to be excluded. In phase 1, there was one token combination for which the correct token to choose changed. Tokens 1 and 2 both changed valence in the phase 2 (100 to –100 points, and vice versa). Thus for the combination of Tokens 1 and 2, Token 1 was the correct token to choose in phase 2, but Token 2 was the correct token to choose in phase 2.

The dependent variable was the frequency of correct token choices. Data points were excluded if on a simulation the model had not yet had the opportunity to learn the new reinforcement values of both the tokens. This meant that for seven simulations, the trial was excluded. The mean frequency of correct token choices was 61% (s.d. = 50). This performance was not significantly better than chance, t(22) = 1.05, p = ns. An independent samples t-test revealed that the model's performance was not significantly different to the performance of the human subjects, t(35) = .22, p = ns. As with the humans, the small number of trials available for analysis may account for the absence of an effect of valence change on instrumental re-learning.

3.5.3: Results of simulations of Experiment 3

In phase 1, the mean number of correct token choices was 23/26 (s.d. = 2). Figure 3.6 shows the model's token choice performance categorised in terms of percentage correct choices when a positive and a negative token were presented together (mean = 97%, s.d. = 4), two positive tokens of different values (mean = 77%, s.d. = 22), and two negative tokens of different values (mean = 64%, s.d. = 28). Human performance is also shown in figure 3.6, for comparison. Model and human performance was compared using a split plot ANOVA with within-subjects factor Token Combination (Positive-Negative, Positive-Positive, Negative-Negative) and between-subjects factor Group (model vs. human). This revealed a significant main effect of Token Combination, F(1, 58) = 58.5, p < .001, reflecting the fact that there were differences in how successful subjects and the model were for the three different token combination categories. The main effect of Group was not significant, F(1, 58) = .03, p = ns, and the Token Combination by Group interaction was also not significant, F(1, 58) = 3.09, p = ns. These findings demonstrate that the model successfully simulated human performance in phase 1 – the two populations could not be distinguished statistically.

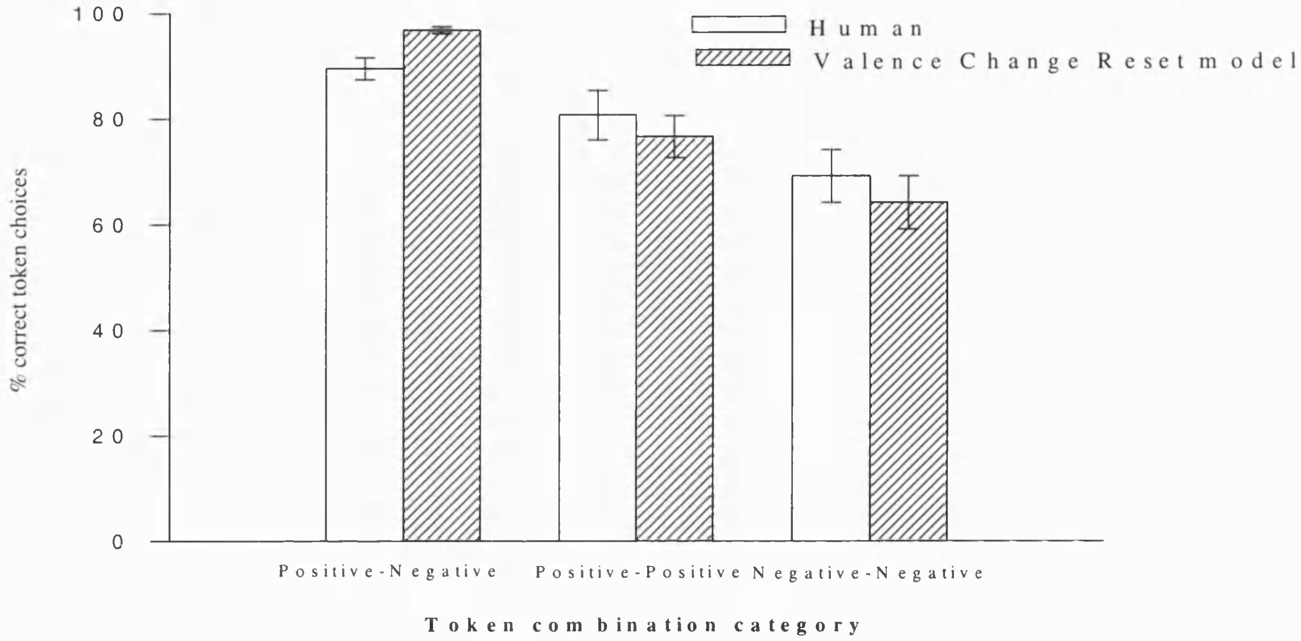

Figure 3.6: Performance of model and humans in phase 1, Experiment 3.

The next analysis investigated whether, as in humans, the valence changes were associated with an adaptive change in token choice behaviour. Tokens 1 and 2 both changed valence in phase 2 (100 to –100 points, and vice versa), as did Tokens 3 and 4 (300 to –300 points, and vice versa). Thus in phase 1, for the combinations of Tokens 1 and 2, 3 and 4, 1 and 4, and 2 and 3, Tokens 1 and 3 were the correct tokens to choose. However in phase 2, Tokens 2 and 4 were the correct tokens to choose. The performance of the humans was significantly better than chance for these token combinations in phase 2.

The dependent variable was the frequency of correct token choices. Data points were excluded if the model had not yet had the opportunity to learn the new reinforcement values of both the tokens. Thus for 25 simulations, one or more trials were missing. However, all data points were only excluded on one simulation. The mean frequency of correct token choices was 78% (s.d. = 29). A one-sample t-test revealed that this performance was significantly better than chance, t(28) = 5.14, p < .001. An independent samples t-test revealed that the model's performance was not significantly different to the performance of the human subjects, t(55) = .13, p = ns.

3.6: Discussion

The aim of this chapter was to implement the instrumental learning and re-learning systems proposed in Chapter 2 in a computational model, the Valence Change Reset model. In the model, the Instrumental Learning system was sensitive to differences in magnitude between different reinforcements (see figure 3.3). Thus magnitude as well as valence information was processed in this system. The Instrumental Re-learning system detected unexpected valences of reinforcement, but did not represent magnitude information. Violations of expected valence reset response outcome expectations in the Instrumental Learning system. The action of the Instrumental Re-learning system was hypothesised to trigger the ANS. The Valence Change Reset model successfully simulated instrumental learning in Experiments 1 to 3. Moreover, the model showed the same rapid instrumental re-learning seen in humans following valence changes, and an absence of rapid instrumental re-learning following magnitude changes. Thus the current model appears to provide a reasonable account of the three main observations in Experiments 1-3: first, that instrumental learning is sensitive to magnitude differences between conditioned stimuli; second, that only valence changes trigger autonomic arousal; third, that valence changes appear to be necessary to trigger rapid instrumental re-learning in the task used.

The current model draws on aspects of other computational models of emotional learning and re-learning in the literature. Learning in the Instrumental Learning system was driven by differences between expected and actual reinforcement (e.g., Rescorla & Wagner, 1972; Pearce & Hall, 1980; Mackintosh, 1975; Sutton & Barto, 1981). The use of learning rules that exploit differences between expected and actual events have been used with success in models such as Montague et al's model of the prediction of future reward in dopaminergic neurons (Montague, Dayan, & Sejnowski, 1996), and Schmajuk et al's model of latent inhibition (Schmajuk, Lam, & Gray, 1996).

Following the work of Gray (e.g., Gray, 1982), partially implemented by Schmajuk et al. (1996), mismatches between expected and actual events were proposed to have an inhibitory effect on conditioned responding, and to trigger the ANS. However, in the current model there was a difference between the processing of unexpected magnitudes of reinforcement, and unexpected valences. Unexpected magnitudes of reinforcement were processed in the Instrumental Learning system. In contrast, unexpected valences of reinforcement were processed in the Instrumental Re-learning system. In this respect, the current model was similar to Grossberg's Adaptive Resonance Theory model in which separate systems process expected and unexpected events (e.g., Grossberg, 1982; Grossberg & Gutowski, 1987; Grossberg & Levine, 1987; Grossberg & Schmajuk, 1987; Carpenter & Grossberg, 1988; Grossberg, 2000). As in the Instrumental Re-Learning system of the current model, the role of Grossberg's expectation violation detection system is to reset expectations in the learning system. However, the current model differs to Grossberg's model as it implements the hypothesis that only valence changes trigger the Instrumental Re-learning system. Other discrepancies between predicted and actual reinforcement are processed in the Instrumental Learning system. In Grossberg's model, a vigilance parameter determines the tolerance of the system to mismatches between expected and actual events.

The hypotheses implemented in the computational model presented here could have been implemented using a number of other architectures and learning rules. However, the specific details of how the two systems – the Instrumental Learning and Instrumental Re-learning systems - are implemented is relatively unimportant. Rather, the simulations from the current model are important in that they show that the two separate systems proposed in Chapter 2 can account for the behavioural and psychophysiological findings of that chapter. It should be noted that Experiment 4 was not simulated. The main reason for this was that the Valence Change Reset model did not implement the hypothesised Motivational Significance system proposed in Chapter 2. This system was thought to contribute to instrumental re-learning and ANS activation in Experiment 4. Thus it was not appropriate to attempt to simulate the human data from Experiment 4 with the Valence Change Reset model.

The Valence Change Reset model implemented hypotheses based on the data collected in Chapter 2, in particular, the finding that valence changes, but not magnitude changes, produced arousal increases. Future work is necessary to test predictions generated from the model. For example, two predictions can be made that would provide a critical test of the Valence Change Reset model. First, the model predicts that any valence change, however small, will trigger an arousal increase. Second, it is also predicted that no magnitude change, however large, will trigger an arousal increase. Thus, as noted in the previous chapter, further research is necessary to ascertain under what conditions the dissociation between arousal responses to valence changes and magnitude changes holds.

Although the Instrumental Learning and Re-learning systems proposed here interact importantly, the two systems are hypothesised to use different expectation representations. The two systems must therefore be hypothesised to be separable. Thus it is possible that certain neurological populations may have suffered damage to the Instrumental Learning system or the Instrumental Re-learning systems. Clearly, damage to either of these systems would result in very different performances on instrumental learning and re-learning tasks. It may be possible to simulate such performance using the Valence Change Reset model, with one or other system damaged. Chapters 4 to 6 are concerned with assessing instrumental learning and re-learning in three different neurological populations. It will be investigated whether their impairments might be accounted for in terms of damage to instrumental learning or re-learning systems.

3.7: Summary

In this chapter, the Valence Change Reset model implemented the hypothesis that an instrumental learning system interacts with a separable instrumental re-learning system. In the Valence Change Reset model, the instrumental learning system represented information about the valence and magnitude of reinforcement. In contrast, the instrumental re-learning system only represented valence information. The model successfully simulated the performance of human subjects on three variants of an instrumental learning and re-learning task, in particular, the observation that only valence changes appear to trigger rapid behavioural change. The Valence Change Reset model predicts that the instrumental learning and re-learning systems can be independently impaired.

Chapter 4

Investigating the role of the human amygdala in instrumental learning: A case study

4.1: Introduction

In Chapter 2, it was proposed that two different systems represent reinforcement expectations. The first system was proposed to underlie instrumental learning. The second system was proposed to be involved in instrumental re-learning. These two systems were implemented in a computational model in Chapter 3. In this computational model, the Instrumental Learning system represented both the magnitude and valence of expected and actual reinforcements. In contrast, the Instrumental Re-learning system represented only valence, and responded to unexpected valence changes by resetting response outcome expectations in the Instrumental Learning system. The model was able to accurately simulate human behavioural data on three variants of an instrumental learning and re-learning task. It was suggested in the previous chapter that the two hypothesised systems could be independently damaged. This chapter assesses the possibility that human amygdala damage may result in damage to part of the system that underlies instrumental learning. Previous research that supports this claim is briefly reviewed. The hypothesis is then tested with a patient with left amygdala damage.

4.2: The basolateral amygdala and response outcome expectation representation

4.2.1: Animal lesion and electrophysiological evidence

It has been suggested that the basolateral amygdala forms part of a neural circuit that supports instrumental conditioning (e.g., Everitt & Robbins, 1992; Whitelaw, Markou, Robbins & Everitt, 1996; Killcross, Robbins & Everitt, 1997; Burns, Everitt & Robbins, 1999; Schoenbaum, Chiba & Gallagher, 1999; Parkinson, Robbins & Everitt, 2000). This circuit is thought to be dissociable from conditioned emotional responses mediated by the central nucleus (e.g., Killcross et al., 1997; Hitchcott & Phillips, 1998; Parkinson et al, 2000). For example, Parkinson et al. (2000) suggest that:

"... the central nucleus may be interpolated within circuitry that underlies conditioned reflexive and motivational influences on behaviour, whilst the basolateral amygdala may subserve a more complex representational role in emotionally charged decisions and voluntary behaviour." Parkinson et al., 2000: p412).

In line with a role for the basolateral amygdala in instrumental learning, lesions or drug manipulations of this region affect the development of conditioned responses supported by secondary reinforcers (e.g., Cador, Robbins & Everitt, 1989; Burns, Robbins & Everitt, 1993; Hatfield, Han, Conley, Gallagher & Holland, 1996; Whitelaw et al., 1996; Killcross et al, 1997; Hitchcott & Phillips, 1998). Devaluation experiments suggest that the role of the basolateral amygdala in instrumental learning may include representing reinforcement expectations associated with conditioned stimuli. For example, in a study conducted by Hatfield et al. (1996), rats learnt an association between light and food. The food was then paired with lithium chloride. When presented with the light, the control rats were significantly less likely than the lesioned rats to show conditioned responding. This suggests that the conditioned stimulus activates a representation of expected (and in this case devalued) response outcome reinforcement, and that this representation is mediated by the basolateral amygdala. A similar result was found in a devaluation study using monkeys, and again the basolateral amygdala was implicated in the representation of reinforcement expectations (Malkova, Gaffan & Murray, 1997).

Recordings from basolateral amygdala cells indirectly suggest that the basolateral amygdala contributes to the representation of reinforcement expectations that underlie instrumental behaviour. Schoenbaum et al. (1999) measured the activity of basolateral amygdala cells during a series of go/no-go odour discrimination tasks. The basolateral amygdala cells developed selectivity to positive and negative odours considerably earlier than accurate go/no-go behaviour was achieved. The activity of the basolateral amygdala in anticipation of an expected reward was, for many neurons, similar to their activity when presented with the reinforcer itself. This suggests that the basolateral amygdala may be representing response outcome expectations. During the reversal-learning phase of the task, the cells reversed their responses – for example, cells that previously responded to the odour that was associated with reward in the first phase of the task, now responded to the odour that was previously associated with punishment. Again, this new selectivity developed before the behavioural criterion for reversal learning was achieved. The authors note that these findings suggest that information about reinforcement expectations encoded in the basolateral amygdala may support the development of instrumental behaviour (Schoenbaum et al., 1999).

4.2.2: Evidence from humans with amygdala damage

Amygdala lesions in humans are not circumscribed to particular nuclei as far as is known. Neuropsychological work therefore cannot attribute a cognitive function to a particular nucleus of the amygdala. Nonetheless, neuropsychological findings suggest that the human amygdala plays a similar function as the non-human amygdala. For example, rats with basolateral or lateral amygdala lesions show attenuated Pavlovian fear responses (e.g., Kim, Rison & Fanselow, 1993; Maren, 1998; LeDoux, Cicchetti, Xagoraris & Romanski, 1990). A similar effect has been seen in patients following amygdala damage. An impairment was shown first by Bechara, Tranel, Damasio & Damasio (1995) in patient SM, who had bilateral amygdala damage due to Urbach-Wiethe disease. In contrast with controls and a patient with bilateral hippocampal damage, the patient did not show a conditioned arousal response (indexed by SCR) to a conditioned stimulus that predicted the occurrence of a startling sound. A fear conditioning impairment was also seen in a patient, SP, who had a right medial temporal and left amygdala lesion. SP failed to show a conditioned autonomic response to a CS paired with shock (Phelps, LaBar, Anderson et al., 1998). Fear conditioning impairments have also been observed in two group studies of patients with damage that included the amygdala. A study of 26 patients with unilateral anteriomedial temporal lobe resections revealed an impairment in acquiring a conditioned autonomic response in simple and conditional fear conditioning tasks (LaBar, LeDoux, Spencer & Phelps, 1995). Five bilateral amygdala patients studied by Bechara, Damasio, Damasio & Lee (1999) also failed to develop a conditioned autonomic response to loud sound.

4.2.3: Evidence from functional imaging studies in humans

In line with the animal and human literature, some (although not all) functional imaging studies have revealed amygdala activity in response to the presentation of conditioned fear stimuli (e.g., LaBar, Gatenby, Gore, LeDoux & Phelps, 1998; Büchel, Dolan, Armony & Friston, 1999; Ploghaus, Tracey, Gati et al., 1999; for review, see Büchel & Dolan, 2000). However, as with the neuropsychological literature, this does not directly support a role for the basolateral amygdala in representing reinforcement expectations, since no current imaging technique has sufficiently good spatial resolution to identify individual nuclei in the amygdala.

4.3: Aims of investigation

A wealth of evidence supports a role for the amygdala (in particular, the basolateral amygdala) in representing reinforcement expectations. In animals, the importance of these representations in guiding instrumental behaviour has been demonstrated. Although the role of the human amygdala in instrumental learning has not been directly assessed, two lines of evidence suggest that an impairment might be seen. First, Gray and colleagues demonstrated an impairment in conditional discrimination instrumental learning in patients with temporal, in particular, right temporal lesions which included the amygdala (Daum, Schugens, Channon, Polkey & Gray, 1991; Channon, Daum & Gray, 1993). However, the hippocampal damage suffered by these patients could also explain their impairments. Second, anecdotal reports from the clinical literature describe behaviour that seems to reflect an absence of knowledge concerning the possible negative outcome of behaviour (e.g., Sprengelmeyer, Young, Schroeder et al., 1999). It was predicted that human amygdala damage would result in an instrumental learning impairment. The first aim of this chapter was to test this prediction using a single-case study of a patient with unilateral left amygdala damage. The second aim was to show that the Valence Change Reset model could simulate the instrumental learning performance of the control subjects on a fourth variant of the instrumental learning and re-learning task described in Chapter 2. The third aim was to simulate the performance of the amygdala patient using the Valence Change Reset model with a damaged Instrumental Learning system.

4.4: Case report

4.4.1: Case Description

BM is a 32-year-old, right-handed man who worked as a caterer. He was arrested in 1994 and subsequently convicted for murder and rape. BM was seen in prison for psychiatric assessment before the trial and was diagnosed as suffering from schizophrenia on the basis of formal thought disorder and persecutory and grandiose delusions. Since his admission to the hospital, BM has received anti-psychotic medication which at the time of testing was flupenthixol depot 400 mg fortnightly and 10 mg procyclidine twice daily.

Hospital file information revealed that BM shows profound social isolation, never makes phone calls, and does not write to anyone inside or outside of the hospital. BM has no visitors, and has indicated that he doesn't wish to have any. He spends most of the time in his room. He never attends hospital social functions, and rarely attends sports events. BM is always polite but finds it difficult to approach people. He does not have any hobbies or interests. BM's mother reports that as a child he was slow in walking and talking, and was rather clumsy. His language developed normally, but his use of it was "slow and ponderous". She recalls that he preferred to be on his own, was isolated from his siblings and other children, and showed little imaginative play. He was also often aggressive without provocation.

A psychiatric opinion based on file information suggested that BM suffered from either Asperger's syndrome or a schizotypal personality disorder or schizo-affective disorder. More recently, the diagnosis of Asperger's syndrome has been confirmed by an independent assessment carried out by an experienced psychiatrist. It should be noted that DSM-IV stipulates that if an individual has received a diagnosis of schizophrenia they cannot subsequently receive a diagnosis of Asperger's syndrome (American Psychiatric Association, 1994). This rule has been relaxed by the clinicians involved because from the clinical records the patient's Asperger's syndrome preceded his schizophrenia. Moreover, on formal assessment of psychiatric symptomatology he showed little evidence of active schizophrenic illness. On the Brief Psychiatric Rating Scale he scored moderate levels for emotional withdrawal and blunted affect. All other scores were mild or not present. For the Schedule for Assessment of Negative Symptoms, the global rating of affective flattening was mild, physical anergia was moderate, but all others were rated 1 or 2 (questionable or mild). On the Comprehensive Psychiatric Rating Scale, nothing of significance was noted except for a "severe or incapacitating illness" in the global rating of illness. Additionally, BM completed the Personality Disorders

Questionnaire. The completion of the questionnaire is rated "too good" which suggests that BM may be under-reporting his psychopathology. All other ratings were below threshold for DSM-IV personality disorders.

BM gave informed consent to participate in all testing sessions. His co-operation throughout the testing sessions was good.

4.4.2: Lesion Localization

Figure 4.1 shows a MRI scan of BM's brain taken in 1996. The scan was performed on a transportable 1GE Sigma MRI scanner (1 Tesla), operated by Alliance Medical Ltd. The scan reveals an abnormal signal return on T2 and possible low intensity lesion in T1 in the lateral part of the basal nuclei of the left amygdala. There was no generalised atrophy, and the frontal areas gave normal signal return. The scan is consistent with a dysembryonablastic neuroepithelial tumour of longstanding or congenital origin.

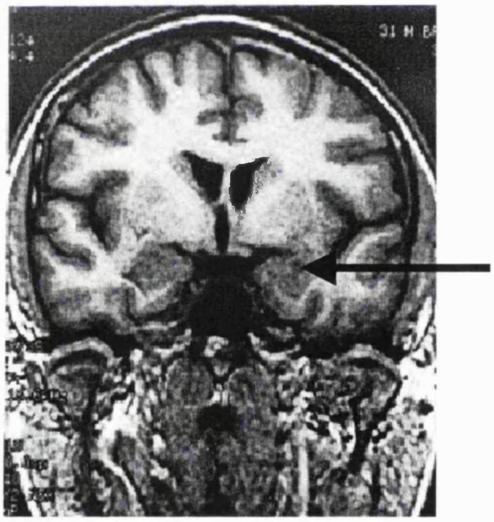

Figure 4.1: MRI scan showing abnormal signal return on T2 and possible low intensity lesion in T1 in the lateral part of the basal nuclei of the left amygdala.

Test	Score / Percentile
WAIS-R Full Scale IQ	103
Age-related Subtest Scaled Scores (AV = 10)	
Comprehension	14
Digit Span	11
Similarities	9
Block Design	10
Digit Symbol	10
National Adult Reading Test – Revised	108
Weschler Memory Scale – Revised	
Prose Recall: Immediate	11/22 : 31%ile
Delayed	11/22 : 50%ile
Design Recall: Immediate	34/41 : 56%ile
Delayed	33/41 : 69%ile
Paired Associates	
T1:	23/24 : 90%ile
T2:	24/24 : >90%ile
Recognition Memory Test	
Faces	39/50 : 25%ile
Buildings	46/50 : High average
Landscapes	27/30 : 75%ile
Words	42/50 : 10-25%ile
Rey Complex Figure Test	
Copy	36/36 : 100%ile
Recall	18/36 : 20-30%ile
Adult Memory & Information Processing Battery	
Information Processing, Form 1	
Motor Speed	60/90 : 90%ile
Cognitive Speed A	79/105 : 75-90%ile
Cognitive Speed B	81/105 : 75-90%ile
Accuracy A	2 : =50%ile
Accuracy B	0 : >90%ile
(all prorated from half-administrations)	
Graded Difficulty Naming Test	21/30 : 50%ile
Concrete Word Synonym Test	22/25 : 50-75%ile
Abstract Word Synonym Test	21/25 : 50%ile

Table 4.1: BM's neuropsychological performance

4.4.3: Neuropsychological assessment

BM was assessed on the WAIS-R and obtained a Full Scale IQ in the average range (see table 4.1). In particular, his performance on the Comprehension subtest was superior. Reading performance on the NART (Nelson & Willison, 1991) indicated a comparable level of ability. On the Weschler Memory Scale – Revised, BM was within the normal range for both prose and design recall. His memory for Paired Associates was in the 90th %ile or above. BM's performance on the Recognition Memory test for faces and words (Warrington, 1984) was at the lower end of the normal range. However, his recognition memory for buildings (Whiteley & Warrington, 1978) was superior, as was topographical recognition memory (Warrington, 1996). On the Rey Complex Figure Test, he was unimpaired in Copying, but in the low average range for Recall.

His performance on all tests in the Adult Memory & Information Processing Battery was in the 50th %ile or above. His naming skills were intact. His performance was within the normal range on the Graded Difficulty Naming Test (Warrington, McKenna & Orpwood, 1998). BM's single word comprehension was within the average range on a stringent Synonym Test (Warrington, McKenna & Orpwood, 1998).

Overall, BM's neuropsychological assessment shows that his IQ and reading ability are all in the average range. In addition, BM had no clinically significant intellectual, memory, language or speed of processing difficulties. BM's executive functions were also intact. These data are provided in detail in Chapter 8 (section 8.10).

4.5: Experiment 1 – The Four-Token Snake task

The aim of this experiment was to test the prediction that BM would show an impairment in instrumental learning. Some of the subjects in the experiments presented in Chapter 2 performed poorly (and were excluded from the analyses). For this reason, it was decided that a simpler version of the task would be more appropriate for investigating the effects of neurological damage. The prediction was therefore tested using a simplified four token version of the instrumental learning and re-learning tasks presented in Chapter 2; the Four Token Snake task.

4.6. Method

4.6.1: Control subjects

BM's performance was compared with that of five healthy males, matched for educational level, age, and IQ, with mean age of 30 years (s.d. = 4) and a mean WAIS-R IQ subtest score of 11.2 (s.d. = 2).

4.6.2: Procedure

As with the experiments in Chapter 2, the task was presented on a computer. The subject read the game instructions from the computer screen. These instructions were identical to those in Chapter 2. However, neither BM nor the controls were given performance-related chocolate bonuses. This was because hospital security prohibited bringing confectionery into the hospital. Informed consent was given by all subjects.

The format of the game was identical to the tasks described in Chapter 2. The subject moved a snake around the playing field of the computer screen, using the keyboard cursor keys. At the beginning of each trial, two small coloured tokens appeared on the screen simultaneously, equidistant from the snake's head. The screen then froze for four seconds. The subject was instructed to decide which token they were going to eat during this period. When the screen unfroze, the subject moved the snake to the token of choice. A message then appeared on the screen telling the subject how many mice had been won or lost. The screen was frozen for four seconds, with the reinforcement message. The playing field then cleared for the next trial, and the total score message at the top of the screen was updated. The game lasted approximately 20 minutes.

Four colours of token were used. Each coloured token was associated with a certain number of points, as shown in the "phase 1" column of table 4.2. The tokens were presented in pairs in nine blocks. Each block contained 10 trials, comprising the 10 possible token combinations (i.e., four same-colour token combinations and six different-colour token combinations). There were four phases to the experiment: familiarisation (Block 1), expectation acquisition (Blocks 2-4), reversal 1 (Blocks 5-7), and reversal 2 (Blocks 8-9). There were no breaks between any of the phases, nor was the subject informed that there were different phases in the experiment. Token pairs were presented on the screen randomly within each phase of the experiment. Discrepancy was produced in the first reversal phase by reversing the points associated with Tokens 1 and 2, as shown in the "phase 2" column of table 4.2. Tokens 3 and 4 kept the same points value during this phase of the experiment. Discrepancy was produced again in the second reversal phase by reversing the points associated with Tokens 3 and 4, as shown in the "phase 3" column of table 4.2. Tokens 1 and 2 this time remained unchanged.

Token	Phase 1 (expectation acquisition)	Phase 2 (reversal 1)	Phase 3 (reversal 2)
1	300	**-300**	-300
2	-300	**300**	300
3	300	300	**-300**
4	-300	-300	**300**

Table 4.2: Token values in phase 1 (expectation acquisition), phase 2 (reversal 1), and phase 3 (reversal 2) of Four-Token Snake task. Bold indicates a valence change.

4.6.3: Data treatment

Scoring of token choices in phase 1 was as follows. If presented with a positive and a negative token, the trial was scored as "correct" if the subject chose the positive token. The trial was scored as "incorrect" if the subject chose the negative token. If the tokens were of equal value (i.e., both positive or both negative) the trial was not scored.

4.7: Results and Discussion

BM was given the task twice, with 4 months between testing sessions. His performance in phase 1 (expectation acquisition) on the two administrations is shown in table 4.3, together with the control subjects' performance.

	BM		Controls	
Phase	Test 1	Test 2	Mean (s.d.)	Range
Phase 1 (expectation acquisition)	25%	58%	88% (16)	67-100%

Table 4.3: Percentage correct responses by BM and controls in phase 1 of the Four-Token Snake task.

It can be seen from table 4.3 that on both administrations of this task, BM's performance in phase 1, which assesses instrumental learning ability, was below the range of that of the controls. Indeed, his performance on the first administration was substantially below that of the worst control[2]. A binomial test revealed that BM's performance on the second administration was not significantly better than chance, p = .19. Thus, as predicted, BM was impaired in directing his instrumental behaviour on the basis of learnt response outcome expectations. As BM did not learn response outcome expectations, it was not possible to assess his instrumental re-learning following valence changes.

4.8: Experiment 2 – The Ask for Money task

Experiment 1 showed that BM was impaired on an instrumental learning task. In order to investigate this impairment further, a second different instrumental learning and re-learning task was developed. In this second task, the Ask for Money task, instead of choosing one of two conditioned stimuli, the subject had to choose one of two different instrumental behaviours in response to a conditioned stimulus.

4.9: Method

4.9.1: Control Subjects

BM's performance was compared with that of five healthy males, matched for educational level, age, and IQ, with mean age 31 years (s.d. = 2) and mean WAIS-R subtests score of 11.0 (s.d. = 1.5).

[2] A further nine control subjects were also given the Four Token Snake task, and their data are reported in Chapters 5 and 6. Like the control subjects presented here, none of these nine subjects scored less than 67% correct in phase 1 of the Four Token Snake task.

4.9.2: Procedure

This task was also presented on computer. The task was similar in design to Experiment 4 reported in Chapter 3. For clarity, the task instructions are given verbatim below:

> *You are penniless and desperately need some money. You are going to have to try to persuade your relatives to give you some money.*
>
> *When you see a relative, you can either **hint** that you would like some money, or **beg** them to give you money.*
>
> *Unfortunately, you already owe all of your relatives a lot of money, so sometimes a relative will demand that you repay them some of the money you owe them.*
>
> *Some relatives may respond better to hinting, and some relatives may respond better to begging. This means that you can maximise the amount of money that your relatives give you, and minimise the amount of money that you have to pay them back, by learning which is the best way to ask your different relatives for money.*
>
> *In the game, relatives will appear one at a time. You must decide whether to **hint** for money, or **beg**. Once you have chosen, a message will appear on the screen. If your relative has given you money, the message will say **"You have won x pounds"**. If your relative has demanded money from you, the message will say **"You have lost x pounds"**. A message will also appear on the screen, telling you how many pounds you would have won or lost if you had chosen the opposite way to ask for money (i.e., hinting instead of begging, or begging instead of hinting).*
>
> *Remember, you need all the money you can get! Good luck.*

On each trial, a picture of a relative appeared in the centre of the screen (see figure 4.1). Subjects could hint that they would like to borrow money by clicking the button marked, "Hint". Alternatively, subjects could click on the button marked, "Beg". A message then appeared on the screen, telling the subject how much money that relative had given them, or taken away from them (see figure 4.1), and how much they would have won or lost if they had asked for money in the opposite way. The total score message was updated. There was a pause of 4 seconds before the next trial.

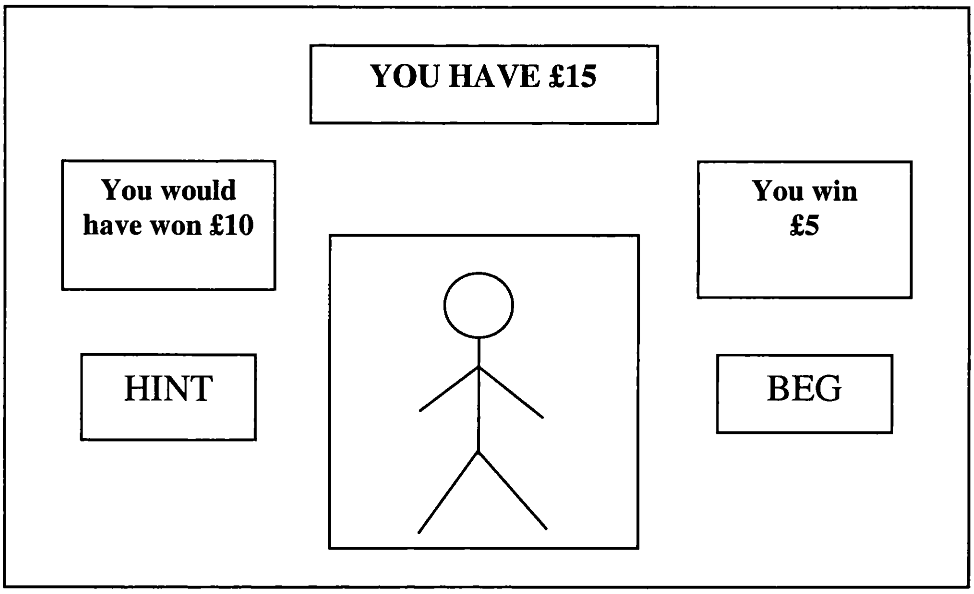

Figure 4.2: Schematic diagram of computer screen layout for Ask for Money experiment.

There were four different relatives in the game, depicted by distinctive bitmap pictures. Each relative gave or took away a certain amount of money when hinted at or begged for money. This is shown in the "phase 1" column of table 4.4. It can be seen that for two relatives it was better to hint, and for two relatives it was better to beg. In both phases of the experiment, there were two positive stimuli (i.e., relatives who gave money) and two negative stimuli.

	Phase 1		**Phase 2**	
Relative	**Hint**	**Beg**	**Hint**	**Beg**
1	5	**10**	**-5**	-10
2	**-5**	-10	5	**10**
3	**10**	5	**10**	5
4	-10	**-5**	-10	**-5**

Table 4.4: Pounds given or taken away by relatives in phase 1 (expectation acquisition) and phase 2 (reversal), when either hinted to or begged for money. Bold text indicates the optimal response for that relative.

The phases were made up of blocks. Each block comprised four trials. Each relative was presented once every block, in a random order. There were eight blocks in phase 1 (expectation acquisition) and six blocks in phase 2 (reversal). In phase 2, Relatives 1 and 2 changed both the valence of their response, and whether begging or hinting was the most optimal response, as shown in the "phase 2" column of table 4.4.

4.9.3: Data analysis

The first block of phase 1 and the first block of phase 2 were not scored, and were excluded from the following analyses. A trial was scored as correct if the subject made the optimal response, otherwise the trial was scored as incorrect.

4.10: Results and Discussion

BM's performance on the two phases of the task is shown in table 4.5. It can be seen that BM performed below the range of the controls in both phase 1 (18/28) and phase 2 (11/20).

	BM	**Controls**	
Phase	Mean	Mean (s.d.)	Range
Phase 1 (expectation acquisition)	18/28	24 (3)	19-26
Phase 2 (reversal)	11/20	17 (3)	14-20

Table 4.5: Number of correct responses by BM and controls in Ask for Money task.

Thus as in the Four-Token Snake task, BM's performance was below the range of that of the controls. A binomial test revealed that BM's performance in phase 1 (expectation acquisition) was significantly better than chance, p = .049. It is possible that BM was able to perform slightly above chance on the Ask for Money task because the design of this task may have made it possible for stimulus-response associations to be set up. These stimulus-response associations may have compensated for BM's impairment. For example, Burns et al. (1999) have suggested that learnt stimulus-response associations may have attenuated the impairment seen in rats with basolateral amygdala lesions on an appetitive conditional discrimination task. Stimulus-response learning has been shown to be dissociable from stimulus-reinforcement learning and is thought to be mediated by the dorsal striatum (McDonald & White, 1993).

4.11: Experiment 3 – performance of "intact" and "impaired" Valence Change Reset model

The first aim of Experiment 3 was to investigate whether the Valence Change Reset model could reasonably simulate the performance of the control subjects on the Four Token Snake task[3]. The second aim of Experiment 3 was to compare BM's poor performance in the instrumental learning phase of the Four Token Snake task with that of the Valence Change Reset model with a damaged Instrumental Learning system.

4.12: Method & Procedure

4.12.1: "Intact" Valence Change Reset model

The network was similar to that used to simulate performance in Chapter 2. It was necessary to make a slight change in architecture: for these simulations there were only four token units, not six. Parameter values were the same as those used in previous simulations, with the exception of ß, which determines the sensitivity of the softmax activation function to differences in the value of weights. ß was set at 5 in the current simulations (for simulations in Chapter 3, ß = 10). This was necessary in order to match the performance of the controls. This may reflect the difference in IQ between the population studied in this experiment, and the mainly student population used in the experiments reported in Chapter 2.

4.12.2: "Impaired" Valence Change Reset model

In the Valence Change Reset model, the change in weights represents the formation of response outcome expectations. An impairment in this process was therefore modelled by reducing the constant, m, that determines the learning rate for each token unit. The learning rate for the active unit was determined in the usual way by the equation:

$$\varepsilon_x = m\,[f(r) - w_x]$$

However, in the "impaired" model, m = 0.001, slowing the learning of response outcome expectations.

[3] The performance of the nine control subjects reported in Chapters 5 and 6 were included in this data set.

4.12.3: Procedure

Fourteen simulations of the “intact” model were run[4]. Ten simulations of the “impaired” model were run. As before, for each simulation pre-learning weights were set randomly to values between 0.4 and 0.6.

4.13: Results and Discussion

4.13.1: Performance of the “intact” Valence Change Reset model compared with controls

The token choice performance of the Valence Change Reset model is shown in table 4.6. A split-plot ANOVA with within-subjects factor Phase (phase 1, phase 2, phase 3) and between-subjects factor Group (human vs. model) was conducted in order to investigate how well the model simulated control performance. There was no significant main effect of Phase, F(2, 52) = .43, main effect of Group, F(2, 52) < 1, or Phase by Group interaction, F(1, 26) = .74, all p’s = ns. Thus the performance of the control subjects and the model were statistically indistinguishable.

4.13.2: Performance of the “Impaired” Valence Change Reset model compared with BM

The token choice performance of the “impaired” model in phase 1 (expectation acquisition), phase 2 (reversal 1) and phase 3 (reversal 2) is shown in table 4.6.

	Controls	“Intact” model		BM		“Impaired” model	
Phase	Mean (s.d.)	Mean (s.d.)	Range	Test 1	Test 2	Mean (s.d.)	Range
1	88 (14)	86 (10)	75-100	25	58	48 (14)	17-67
2	87 (18)	91 (8)	75-100	67	63	53 (11)	33-67
3	90 (14)	88 (12)	67-100	50	29	51 (21)	13-88

Table 4.6: Performance of the “intact” model, controls, “impaired” model, and BM in phase 1 (expectation acquisition), phase 2 (reversal 1) and phase 3 (reversal 2) of the Four-Token Snake task.

[4] This matched the total number of control subjects.

It can be seen from table 4.6 that BM's performance fell within the range of the network in all phases of the experiment, in both administrations of the task. Thus damaging the Instrumental Learning system of the Valence Change Reset model results in a reasonable simulation of performance following amygdala damage.

4.14: General Discussion

Following a unilateral left amygdala lesion of longstanding or congenital origin, BM presented with an impairment in instrumental learning. BM was given two tasks of instrumental learning and re-learning. On both tasks, he performed below the range of controls in the learning phase. These findings provide evidence for the importance of the human amygdala in supporting instrumental learning. On the first of the instrumental learning tasks, the Four-Token Snake task, BM's poor performance was simulated by the Valence Change Reset model in which the learning rate for learning response outcome expectations was greatly reduced.

4.14.1: Implications for the role of amygdala in instrumental learning

BM's poor performance on the two instrumental learning tasks is consistent with the hypothesis that the basolateral amygdala forms part of the neural circuitry that supports instrumental conditioning (e.g., Killcross et al., 1997; Burns et al., 1999). This impairment was simulated in a computational model by greatly reducing the ability of the model to develop expectations of reinforcement. The Valence Change Reset model therefore predicts impaired performance on both positive and negative instrumental learning tasks. It could not be determined from the current studies whether BM's impairment lay in learning negative reinforcement expectations, positive reinforcement expectations, or both. This was because both tasks involved both appetitive and aversive instrumental learning.

4.14.2: Alternative accounts of findings

An alternative explanation of BM's poor performance on the two instrumental learning tasks is that he had a general memory impairment, rather than an impairment specifically in learning the reinforcements associated with particular response options. Arguing against this account is BM's mostly normal performance on four tests of memory (see table 4.1). It is possible that in the Ask for Money task, BM's impairment was not due to an inability to develop behavioural responses according to reinforcement expectations, but in learning an arbitrary response according to a rule (e.g., "If relative 1, then beg"). However, as will be shown in Chapter 8 (see table 8.2), BM performed normally on two executive function tasks that also involved performing an arbitrary response in accordance with a rule. BM performed normally on the Rule Shift task (Wilson, Alderman, Burgess, Emslie & Evans, 1996) which involves responding verbally to playing cards according to an arbitrary verbal rule. BM was also intact on a non-spatial conditional learning task in which each of six stimuli had to be associated with one of six arbitrary responses (Petrides, 1990). As the Four-Task Snake task and the Ask for Money task involved only four stimuli, it seems unlikely that a deficit in a rule-learning component in the instrumental learning task underlies BM's severe impairment on these two tasks.

BM's performance on the Ask for Money task was better than his performance on the Four Token Snake task, although his performance was still impaired relative to controls. With regard to this, it is interesting to note that BM also performed normally on the Intra-dimensional / Extra-dimensional Shift task, reported in Chapter 8 (Dias, Robbins & Roberts, 1996; Hughes, Russell & Robbins, 1994). In the Intra-dimensional/Extra-dimensional Shift task, the subject must choose one of two stimuli on each trial, on the basis of the presence of a rewarded shape or line. Task difficulty effects may underlie BM's range of performance on the Four Token Snake, Ask for Money, and Intra-dimensional/Extra-dimensional Shift task. It is also possible that in the Ask for Money and Intra-dimensional/Extra-dimensional Shift task, stimulus-response associations may have supported correct instrumental responding (e.g., Burns et al., 1999).

4.15: Summary

The findings of the current experiments are consistent with animal literature suggesting a role for the (basolateral) amygdala in representing response outcome expectations and guiding instrumental behaviour. The impaired performance of a patient with left amygdala damage was successfully simulated by the Valence Change Reset model in which the capacity to develop reinforcement expectations in the Instrumental Learning system had been greatly reduced.

Chapter 5

Investigating the role of the human orbitofrontal cortex in instrumental learning and re-learning: Two case studies

5.1: Introduction

In Chapter 2, it was proposed that two different systems represent reinforcement expectations. One system was proposed to underlie instrumental learning, the other instrumental re-learning. These two systems were implemented in a computational model in Chapter 3. This model, the Valence Change Reset model, successfully simulated human behavioural data on three variants of an instrumental learning and re-learning task. It was suggested in Chapter 3 that the two hypothesised systems could be independently damaged. The previous chapter reviewed the evidence that amygdala damage might reflect damage to part of an instrumental learning system, and this hypothesis was tested with a single-case study of a patient with unilateral amygdala damage. Consistent with prediction, the patient was impaired in instrumental learning. This chapter assesses the evidence that human orbitofrontal cortex damage may result in damage to part of the neural circuit that mediates instrumental re-learning. This proposal is then tested with two patients with orbitofrontal cortex damage. The performance of the patients is compared with that of the Valence Change Reset model with a damaged Instrumental Re-learning system.

5.2: The orbitofrontal cortex and instrumental re-learning

A number of studies indicate that the orbitofrontal cortex is part of the neural circuitry that mediates rapid behavioural change following unexpected changes in the valence of reinforcement. Experimental animals with orbitofrontal lesions are impaired in extinction (e.g., Butter, 1969), and in the reversal component of object, place and visual discrimination reversal learning tasks (e.g., Iversen & Mishkin, 1970; Jones & Mishkin, 1972; Dias, Robbins & Roberts, 1996). Monkeys with lesions of the inferior convexity of the orbitofrontal cortex are impaired at withholding responses on no-go trials (Iversen & Mishkin, 1970). Similarly, it has been found that patients with ventral frontal lesions are impaired in the reversal component of reversal learning tasks (Rolls, Hornak, Wade & McGrath, 1994).

Electrophysiological studies also suggest that the orbitofrontal cortex is involved in responding to changes in reinforcement contingencies. For example, Thorpe, Rolls & Maddison (1983) measured neuronal activity in the orbitofrontal cortex of monkeys during a go/no-go visual discrimination reversal task. Some cells were found that responded strongly on the first error trial of the reversal when the monkey received saline instead of the expected juice. However, these cells did not respond to saline outside the context of the task, nor to other arousing stimuli. This differential responsiveness suggests that these cells were responding to the information that the monkey's expectation had been violated, rather than to the reinforcement *per se.* Other cells were found that responded to unexpected saline when the monkey was licking for juice. Again, this suggests that the orbitofrontal cortex responds to expectation violations. In a similar study using a go/no-go task, Watanabe (1989) revealed a population of orbitofrontal neurons that appeared to code for the correctness of the monkey's response, regardless of outcome.

Thus the evidence is consistent with a role for the orbitofrontal cortex in responding to unexpected changes in valence, as evidenced by the behavioural effects of orbitofrontal damage on reversal learning (e.g., Rolls et al., 1994; Dias et al., 1996), and by the neuronal activity seen in the orbitofrontal cortex following unexpected reinforcements (e.g., Thorpe et al., 1983; Watanabe, 1989). Rolls has argued that the orbitofrontal cortex is involved in stimulus-reinforcement re-learning (e.g., Rolls, 1990; 1996; 2000). An alternative, although not mutually exclusive, position is that the orbitofrontal cortex is involved in inhibiting previously rewarded or predominant responses (e.g., Elliott, Dolan & Frith, 2000; Roberts & Wallis, 2000; Shimamura, 2000). It is currently not possible to distinguish between these two different accounts, since in traditional reversal learning experiments, the need to learn a new stimulus-reinforcement association is confounded with the need to change response.

5.3: The orbitofrontal cortex and instrumental learning

Rolls has argued that the orbitofrontal cortex is also involved in stimulus-reinforcement learning (e.g., Rolls, 1996; 2000). This position predicts that patients and experimental animals with orbitofrontal cortex damage will be impaired in instrumental learning as well as re-learning. In contrast, the position presented here, which suggests that the instrumental re-learning system is separable from the instrumental learning system, predicts that orbitofrontal cortex damage will leave instrumental learning intact. Butter (1969) found no difference between controls and monkeys with orbitofrontal cortex damage in the acquisition of instrumental responses in a lever-pressing task. Iversen & Mishkin (1970) found no difference between monkeys with medial orbital lesions and controls in acquiring instrumental responses in object and visual pattern discrimination tasks. Similarly, while Jones & Mishkin, (1972) did find some evidence of impairment in an object discrimination task, they argued that the orbitofrontal lesioned monkeys' pattern of errors was "inconsistent with a major deficit in acquiring associations" (p373). Dias et al. (1996) found no impairment in either simple or conditional discrimination learning in monkeys with lesions of the orbitofrontal cortex. From the human literature, Rolls et al. (1994) found that patients with orbitofrontal cortex damage were not impaired in simple visual discrimination learning.

Recent electrophysiological data from Schoenbaum, Chiba & Gallagher (1999) also fail to support the hypothesis that the orbitofrontal cortex is involved in instrumental learning. Schoenbaum et al. (1999) measured the activity of both basolateral amygdala and orbitofrontal cells during a series of go/no-go odour discrimination tasks. A number of neurons in both the basolateral amygdala and orbitofrontal cortex fired selectively during the evaluation of odour cues. That is, some neurons in both brain regions fired more strongly during the evaluation of odour cues predicting sucrose solution, whereas other neurons fired more strongly during the evaluation of odour cues predicting aversive quinine solution. However, only the basolateral amygdala cells developed this selectivity before accurate go/no-go behaviour was achieved. In contrast, selective responses to the conditioned stimuli were only observed in the orbitofrontal cortex when the rat performed reliably well on the task. Further differences in cell activity emerged during the reversal-learning phase of the task. Over half of the basolateral amygdala cells reversed their responses. For example, cells that had responded most to the previously rewarded odour now responded to the previously punished odour. Again, this new selectivity again developed before the behavioural criterion for reversal learning was achieved. In contrast, significantly fewer orbitofrontal cortex cells reversed their responses.

Instead, a new population of orbitofrontal cortex cells developed selectivity to the different outcomes associated with the odours. The authors concluded that:

> *"selective activity in orbitofrontal cortex did not consistently represent the identity of particular odours, the motivational characteristics of the associated reinforcer, or preparation for the motoric response."* Schoenbaum et al., 1999: p1882.

However, one study has found evidence for a role for the ventromedial frontal cortex in instrumental learning. Gaffan & Murray (1990) found that bilateral lesions of the ventromedial frontal cortex resulted in impaired performance on a visual discrimination task, compared with preoperative performance.

Thus there is currently little evidence to support the case that the orbitofrontal cortex is involved in instrumental learning. However, a caveat to this conclusion is that the instrumental learning phases of the tasks used to assess patients and animals with orbitofrontal cortex damage are usually very simple, and therefore may be insensitive to instrumental learning deficits. For example, patient BM performed normally on both the simple and conditional discrimination learning phases of Dias et al.'s (1996) Intra-dimensional/Extra-dimensional Shift task (see table 8.2). However, as demonstrated in the previous chapter, BM does have an impairment in instrumental learning. Thus, instrumental learning in patients with orbitofrontal cortex damage needs to be assessed using a task sensitive to impairment in neurological patients.

5.4: Summary and aims

In Chapter 3 it was argued that instrumental re-learning involves a mechanism additional to that involved in instrumental learning. The experiments reported in Chapter 2 indicated that this mechanism only responds to certain types of reinforcement expectation violation, namely, valence changes and reinforcements that indicate that a change in response is required. Since valence changes normally indicate that a change in response is required, it was argued that the processing aim of this mechanism is response change. The data reviewed above suggest that the orbitofrontal cortex may be an important part of the neural circuitry involved in response change (e.g., Rolls, 1990; Shimamura, 2000). This suggests that orbitofrontal cortex damage may reflect damage to part of the neural circuitry that mediates instrumental re-learning.

The first aim of the current experiment was to test the prediction that patients with orbitofrontal damage will be impaired in instrumental re-learning. The second aim was to investigate whether patients with orbitofrontal cortex damage are impaired or intact in instrumental learning using an experimental paradigm with a demonstrated sensitivity to instrumental learning impairments - the Four Token Snake task. The third aim was to explore whether the performance of patients with orbitofrontal cortex damage can be simulated by the Valence Change Reset model with a "damaged" Instrumental Re-learning system.

5.6: Case reports

5.6.1: Case Descriptions

Patient CM

CM is a 48 year old patient at a special hospital, who suffered a traffic accident in 1978. His subsequent head injury required neurosurgery. This brain damage resulted in a change in personality, which was particularly evident with regard to his sexual behaviour. His wife found the change in personality increasingly difficult to cope with, and described him as needy, aggressive, and violent. They divorced in 1980. Since this marriage, CM has been married again three times. In all of CM's marriages, his spouses have noted his violent behaviour. CM was convicted of two sexual attacks in 1982, and a third in 1992. CM hit two patients in his first two years in the special hospital.

Patient DJ

DJ is a 53 year old patient at a special hospital. At the age of 14 he rode his bicycle into the back of a lorry and suffered a severe head injury in which he severed the optic nerve of his right eye such that it had to be enucleated. The bi-frontal damage led to severe personality change, with socially inappropriate behaviour, poor impulse control and loss of his friends. A diagnosis of acquired psychopathy was made following his brain damage. Hospital notes reveal that there has been little change in his behaviour since then.

Both patients gave informed consent to participate in all testing sessions.

	CM	DJ	Comparison data
Hayling Sentence Completion (errors Part B)	5	2	3.2 (2.4)
Verbal Fluency	25	60	48.6 (25.8)
Cognitive Estimates (errors)	-	1	4.8 (4.2)
* Temporal Judgements	1/4	-	2.2 (0.9)
Modified Six Elements Task	3/4	3/4	3.6 (0.5)
* Zoo Map	2/4	3/4	2.4 (2.0)
* Key Search	1/4	4/4	2.6 (1.3)
Rule Shift	3/4	-	3.4 (0.9)
* Brixton Spatial Anticipation (errors)	13	8	16 (5.7)
Non-spatial Conditional Learning (errors)	F	-	21.0 (12.0)

Control data from five male controls with mean age of 30 years (s.d. = 4) and mean WAIS-R subtest score of 10.7 (1.3), with exception of tasks marked with an asterisk (*), which were from published norms.

Table 5.1: CM's and DJ's performance on executive function tasks

5.6.2: Lesion Localizations

Patient CM

A CT scan taken in 1983 showed left anterior frontal damage with haemorrhage and oedema in the left frontal region. MRI scans taken in 1994 and 1998 revealed evidence of left frontal lobe damage with altered signal throughout the left frontal pole. The right frontal region appeared normal, as did the temporal lobes.

Patient DJ

DJ also suffered severe bifrontal brain damage. He required neurosurgery for the removal of necrotic brain tissue and reconstruction of his skull in the right frontal quadrant. A CT scan in 1992 showed severe bi-frontal damage being widespread over the orbital surface, with the rest of the brain relatively well preserved.

5.6.3: Neuropsychological assessment

Patient CM

CM's Full Scale WAIS-R IQ was 86. Memory assessment from hospital file information (including Rey Auditory Verbal learning and Rey Complex Figure tasks) concluded that his performance was appropriate to measured IQ. Hospital file information reveals that CM performed normally on their standard frontal battery. In addition, CM was given nine tests of executive function. His performance on these tasks is shown in table 5.1. Although detailed information concerning CM's brain damage is not available, his neuropsychological profile, together with the adverse personality changes detailed above, suggests orbitofrontal, rather than dorsolateral, frontal lobe damage. CM was marginally impaired (>1 s.d. but <1.65 s.d. less than control mean) relative to controls with a higher IQ on the Key Search and Temporal judgement tasks. CM was also severely impaired on the Non-Spatial Conditional learning task, which is sensitive to lateral frontal lobe damage (Petrides, 1990). Overall, considering CM's low IQ, his performance on the executive functions tasks is generally not less than might be expected.

Patient DJ

DJ's full scale WAIS-R IQ was 114. DJ was given seven tests of executive function. His performance is shown in table 5.1. It can be seen that his performance was unimpaired on all seven tests. This suggests that dorsolateral frontal function was preserved.

5.7: Experiment 1 – Intra-dimensional/Extra-dimensional Shift task

The first aim of the investigation was to test the prediction that both CM and DJ would be impaired in instrumental re-learning. This was investigated using the Intra-dimensional/Extra-dimensional Shift task. Four different capacities are assessed in this task: first, the ability to perform simple and compound discrimination learning; second, the ability to transfer this learning to a new exemplar of the same dimension (intra-dimensional shift); third, the ability to change response set from one dimensional to a second previously irrelevant dimension (extra-dimensional shift); and fourth, the ability to change response when reinforcement contingencies change (reversal). The reversal component of the task is sensitive to orbitofrontal cortex damage in non-human primates (Dias et al., 1996). It was therefore predicted that CM and DJ would show a selective impairment in the reversal component of the task.

5.8: Method

5.8.1: Control subjects

There were 12 control subjects with a mean age of 33 years (s.d. = 9) and a mean Ravens score of 9 (50th percentile), s.d. = 3. These subjects were inmates from Holloway and Grendon prisons.

5.9: Procedure

In this computerised task, on each trial the subject is required to choose one of two visual stimuli that appear randomly in two quadrants of the computer screen. The Intra-dimensional/Extra-dimensional Shift task was presented in nine stages. For all stages, the criterion for progressing onto the next stage was a run of eight correct choices. The stages were as follows:

1) *Simple discrimination* between two pink shapes.
2) *Simple reversal* using the same stimuli but with the contingencies reversed.
3) *Compound discrimination-separate.* A pair of white lines were introduced, but separate from the pink shape. The contingencies remained unchanged.
4) *Compound discrimination superimposed* using the same contingencies. The white lines were superimposed upon the pink shapes.
5) *Compound reversal* using the same cues but reversed contingencies.
6) *Intra-dimensional shift* – subjects were required to transfer learning to a new set of exemplars, still responding to the cue of shape.
7) *Intra-dimensional reversal* still using shape but with reversed contingencies.
8) *Extra-dimensional shift* using new exemplars. Subjects had to shift response set to the white lines.
9) *Extra-dimensional reversal* in which the reinforcement contingency for the newly relevant white lines was reversed.

5.10: Results & Discussion

The performance of CM, DJ and the controls on simple and compound discrimination learning, intra-dimensional shifts, extra-dimensional shifts, and reversals are shown in table 5.2. It can be seen that both CM and DJ were unimpaired in discrimination learning, intra-dimensional shifting, and extra-dimensional shifting. DJ was also unimpaired in the Reversal component of the task. However, CM's performance was more than 3 s.d.'s below the control mean.

Phase	CM	DJ	Controls Mean (s.d.)
Discrimination learning	2	0	0.8 (1.5)
Intra-dimensional shift	0	0	0.1 (0.3)
Extra-dimensional shift	6	1	3.6 (4.1)
Reversal	6†	0	1.0 (1.4)

† > 3 s.d.'s below control mean

Table 5.2: Performance (errors) of CM, DJ, and control subjects on Intra-dimensional and Extra-dimensional shifts, and reversal trials.

5.11: Experiment 2 – Four token Snake Task

Experiment 2 had two aims. The first aim was to explore whether CM and DJ would be impaired in instrumental re-learning in a more complex task, namely the Four Token Snake task. The second aim was to explore whether any impairment in re-learning observed was dissociable from instrumental learning performance.

5.12: Method

5.12.1: Control subjects

In addition to the five control subjects described in the previous chapter, nine incarcerated subjects were included in the control group. These subjects were inmates from Holloway and Grendon prisons. Since CM's IQ was below the normal range, and DJ's IQ was above the normal range, the control group was similarly categorised. CM's performance was compared with control subjects who performed below the 50th percentile on Ravens matrices or WAIS-R subtests. DJ's performance was compared with control subjects who performed above the 50th percentile on the Ravens matrices or WAIS-R subtests. Table 5.3 shows the mean, standard deviations, and ranges of age and estimated percentile IQ (based on either WAIS-R subtests or Ravens matrices) of the two control groups.

Group	Age Mean (s.d.) Range	Estimated %ile IQ Mean (s.d.) Range
CM's control group (n = 5)	29 (5) *24-36*	23 (12) *4-38*
DJ's control group (n = 9)	37 (11) *25-54*	86 (14) *53-96*

Table 5.3: Mean, standard deviation and range of age and estimated percentile IQs for CM's and DJ's control groups. CM's age was 48 years and his estimated IQ was in the 18th percentile. DJ's age was 53 years and his estimated IQ was in the 82nd percentile.

5.13: Procedure

The incarcerated subjects were tested in one of the interview rooms attached to whichever ward the subject was housed on. Informed consent was taken for all subjects.

Subjects were given the Four-Token Snake task. Experimental details of this task have been given previously in Chapter 4 (section 4.6.2).

5.13.1: Data Analysis

Scoring of token choices was as follows. A trial was scored as "correct" if the subject chose the token with the largest value. A trial was scored as "incorrect" if the subject chose the token with the least value. If the tokens were of equal value, the trial was not scored. In phases 2 and 3 (reversal 1 and reversal 2), data points were excluded if the subject had not had the opportunity to learn the points value of both tokens in that phase.

5.14: Results & Discussion

The performance of the two control groups was compared using a split plot ANOVA with within-subjects factor Phase (Phase 1, 2, 3) and between-subjects factor Group (CM Control Group vs. DJ Control Group). The analysis revealed no significant main effect of Group, F(1, 12) = .26, p = ns, and no significant Phase by Group interaction, F(1, 12) = 0, p = ns. The data for the two control groups were therefore collapsed together. Table 5.4 presents the means, standard deviations, and ranges of percentage correct token choices for CM, DJ, and the control group, in all three phases of the experiment. The mean percentage correct and standard errors are shown graphically in figure 5.1.

Phase	**CM**	**DJ**	**Control**	
			Mean (s.d.)	Range
Phase 1 (expectation acquisition)	100	75	88 (14)	67-100
Phase 2 (reversal 1)	73	60	87 (18)	46-100
Phase 3 (reversal 2)	25††	50**	90 (14)	67-100

** > 2 s.d.'s below control mean

†† > 4 s.d.'s below control mean

Table 5.4: Mean, standard deviation, and range of percent correct token choices in the three phases of the Four Token Snake task.

Qualitatively, both CM and DJ showed a progressive deterioration of performance across the three phases of the experiment that was not seen in the controls. Quantitatively, both CM and DJ were intact in phase 1 (expectation acquisition), but significantly impaired relative to controls in the second reversal phase of the task. Thus it appears that their initial instrumental learning was intact, but that instrumental re-learning was impaired.

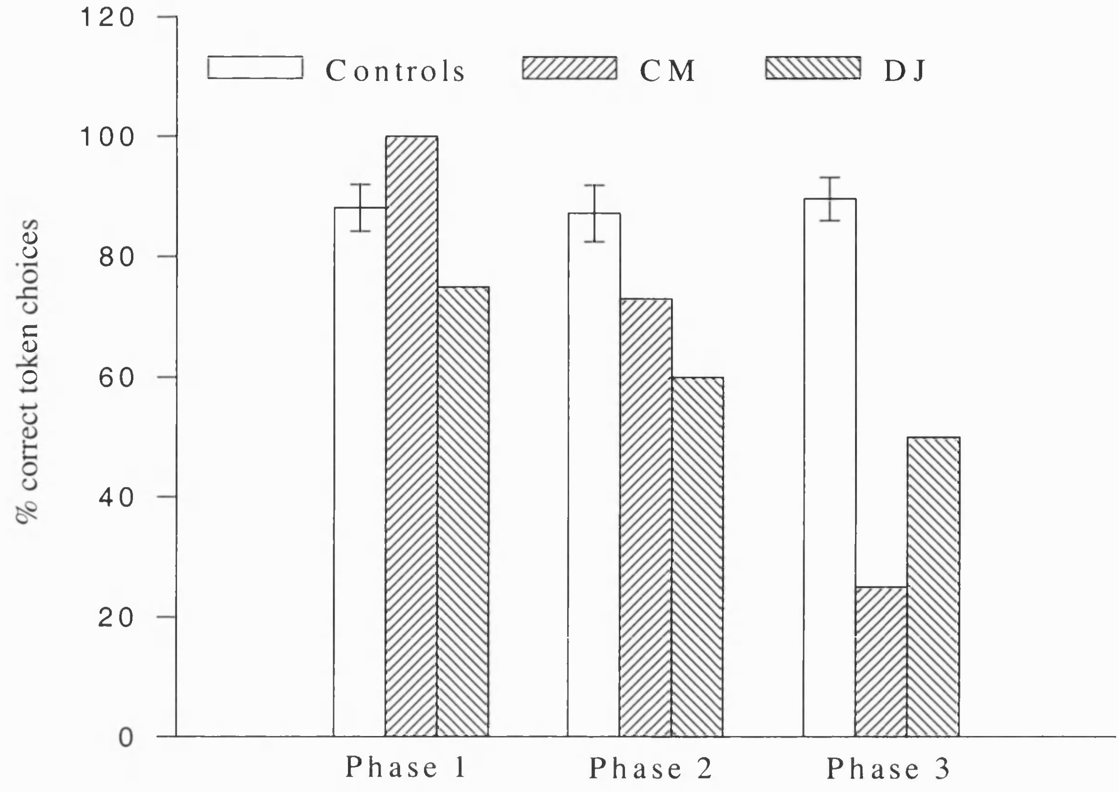

Figure 5.1: Performance of controls, CM and DJ in phase 1 (expectation acquisition), phase 2 (reversal 1) and phase 3 (reversal 2) of the Four-Token Snake task.

CM and DJ's instrumental re-learning impairment was explored further. In the two reversal phases of the task, there were two types of positive-negative token combinations. In the first type of combination, neither token had changed valence. In the second type of combination, one or both tokens had changed valence. The performance of CM, DJ, and the controls on these two types of token combination are shown in table 5.5.

Token combination	CM	DJ	Mean (s.d.)
No valence change	80	100	95 (10)
Valence change	47*	45*	85 (20)

* > 1.65 s.d. below control mean.

Table 5.5: Performance of CM, DJ and controls on token combinations involving no valence changes and token combinations involving one or two valence changes, in phases 2 and 3 of the Four Token Snake task.

It can be seen that both CM and DJ were unimpaired on no valence change token combinations. In other words, when token choice performance did not depend upon re-learning a new token value, performance was unimpaired. This supports the hypothesis that instrumental learning is intact in the patients. In contrast, CM and DJ both showed impairment on token combinations in which at least one of the tokens had changed valence. Thus a closer analysis of their errors in the reversal phases supports the notion that they have a specific impairment in responding to valence changes.

5.15: Experiment 3 – performance of "intact" and "impaired" network

The third aim of Experiment 2 was to compare CM and DJ's performance in the Four Token Snake task with that of the Valence Change Reset model with a damaged Instrumental Re-learning system.

5.16: Method & Procedure

5.16.1: "Impaired" model

In the "impaired" model, the Instrumental Re-learning system was removed. In other words, valence changes no longer resulted in a resetting of weights (reinforcement expectations) in the Instrumental Learning system. The parameter values used were identical to those used in the previous chapter. Ten simulations of the "impaired" network were run. As before, for each simulation pre-learning weights were set randomly to values between 0.4 and 0.6.

5.17: Results and Discussion

The token choice performance of the "impaired" model in phase 1 (expectation acquisition), phase 2 (reversal 1) and phase 3 (reversal 2) is shown in table 5.6.

Phase	CM	DJ	"Impaired" Model Mean (s.d.)
Phase 1 (expectation acquisition)	100	75	88 (9)
Phase 2 (reversal 1)	73	60	85 (9)
Phase 3 (reversal 2)	25	50	90 (12)

Table 5.6: Performance of CM, DJ, and the "impaired" model in phases 1, 2 and 3 of the Four-Token Snake task.

It can be seen from table 5.6 that the absence of the Instrumental Re-learning system in the network did not result in an accurate simulation of the pattern of performance observed in CM and DJ. Unlike the patients, the network did not show a progressive deterioration of performance. Indeed, the absence of the Instrumental Re-learning system did not appear to affect token choice performance.

5.18: General Discussion

This study investigated instrumental learning and re-learning abilities in two patients with orbitofrontal cortex damage. Two different tasks were used. In both tasks, the patients were unimpaired in learning instrumental responses. Thus both patients were unimpaired on a task assessing simple and compound discrimination learning (the Intra-dimensional/Extra-dimensional Shift task; Hughes et al, 1994). The patients were also unimpaired in learning instrumental responses in a second task (the Four Token Snake task). However, in the Intra-dimensional/Extra-dimensional Shift task, patient CM was impaired in reversing learnt instrumental responses following changes in the valence of outcomes associated with those responses. Moreover, in the Four Token Snake task, both patients showed an impairment in their ability to reverse instrumental responses following changes in the valence of stimulus reinforcement values. The findings support the hypothesis that instrumental re-learning involves a mechanism dissociable from that involved in instrumental learning.

5.18.1: The role of the orbitofrontal cortex in instrumental re-learning

The performance of CM and DJ on the Four Token Snake task is consistent with previous research indicating a role for the orbitofrontal cortex in reversing instrumental responses following valence changes (e.g., Jones & Mishkin, 1972; Rolls et al., 1994; Dias et al., 1996). Both patients performed substantially worse than the lowest scoring control subject in the second reversal phase of the experiment. Inspection of the patients' errors on token combinations in which one or more of the token values had changed valence also revealed an impairment relative to controls in the reversal phases of the experiment. In contrast, their performance on token combinations in which neither token had changed valence was unimpaired. Moreover, CM was also impaired in the reversal component of the Intra-dimensional/Extra-dimensional Shift task which also involves reversing an instrumental response following valence change.

It is not clear why DJ was unimpaired in the reversal components of the Intra-dimensional/Extra-dimensional Shift task. It is possible that task difficulty was a factor, especially given DJ's superior IQ. Control subjects made an average total of only one error on all the reversal phases of the task, thus ceiling effects may have occurred. In this respect, it is interesting to note that BM performed normally on the simple and conditional discrimination learning phases of the task, and yet he showed a severe impairment in a more difficult task, the Four Token Snake task.

5.18.2: Implications for the role of the orbitofrontal cortex in instrumental learning

It has been argued that the orbitofrontal cortex is involved in learning associations between stimuli and their reinforcement value (e.g., Rolls, 1990; 1996; 2000). However, the present findings are consistent with previous studies that have mostly failed to demonstrate an impairment in instrumental learning following orbitofrontal cortex damage (e.g., Rolls et al., 1994; Jones & Miskin, 1972; Butter, 1969; Dias et al., 1996; although see Gaffan & Murray, 1990). The present study differs in one important way with respect to these previous studies of extinction and reversal learning. This study used a novel task that was sufficiently complex that the control group did not perform at ceiling on the task. Moreover, the previous chapter found that a patient with amygdala damage was severely impaired in the instrumental learning component of the task. This shows that the instrumental learning phase of the experiment was sufficiently demanding to be sensitive to an instrumental learning impairment. Thus the current findings provide evidence to support the hypothesis that instrumental re-learning can be impaired in the absence of an instrumental learning deficit, and that the orbitofrontal cortex is particularly involved in instrumental re-learning.

5.18.3: Implications for the Valence Change Reset model

The performance of the patients on the Four Token Snake task supports the hypothesis that instrumental re-learning following valence changes depends upon a mechanism dissociable from that involved in instrumental learning. The current findings also lend support to the data reviewed previously that suggested a role for the orbitofrontal cortex in mediating changes in instrumental responding following changes in stimulus-reinforcement associations (e.g., Thorpe, Rolls & Maddison, 1983; Watanabe, 1989). However, the Valence Change Reset model did not provide an adequate account of the Instrumental Re-learning system and its interaction with the Instrumental Learning system. Removal of the Instrumental Re-learning system did not simulate the impaired performance of the patients following valence changes. Weight values attained appropriately positive or negative values following valence changes more slowly than in the intact network, yet these slow weight changes mediated by the Instrumental Learning system were sufficient to mediate correct token choices.

It is interesting to note that, all other things being equal, the network was as likely to choose a stimulus that has rarely been chosen in the past as a stimulus that has been frequently chosen. In other words, the network did not represent how often in the past a token had been responded to previously. The phenomenon of the "prepotent" or "dominant" response would suggest that this is a possible weakness of the model. Stimulus-response learning as well as stimulus-reinforcement association learning is thought to contribute to instrumental responding (e.g., McDonald & White, 1993; Balleine & Dickinson, 1998). Several authors have argued for a role for the striatum as well as the orbitofrontal cortex in mediating instrumental responses (e.g., Rolls, 1996; Cador et al, 1989; Everitt & Robbins, 1992; Gallagher & Schoenbaum, 1999).

For example, Schoenbaum & Gallagher (1999) suggest that there are two pathways for the expression of instrumental responses: via direct projections from the basolateral amygdala to the striatum; and indirectly via the prefrontal cortex to the striatum. Rolls (1996) has suggested that response reversal processes mediated by the orbitofrontal cortex may act via projections to the striatum. Thus, the striatum may be a neural site where response tendencies develop. The orbitofrontal cortex may act to reset these response tendencies rather than, or as well as, resetting expectations in the basolateral amygdala.

5.19: Summary

The findings presented here support the hypothesis that instrumental re-learning is separable from instrumental learning, and that the prefrontal cortex is necessary for the former but not the latter process. What remains unknown is what precise role the orbitofrontal cortex plays in instrumental re-learning. The Valence Change Reset model implemented the hypothesis that an instrumental re-learning system facilitates re-learning of stimulus-reinforcement associations by resetting response outcome expectations. However, removal of this mechanism did not adversely effect instrumental re-learning by the model. The addition or substitution of an inhibitory role to the Instrumental Re-learning system was therefore suggested. The current experiments did not test the prediction that patients with orbitofrontal cortex damage should be unimpaired in slow instrumental re-learning following changes in the magnitude of reinforcement. This is a direct prediction from the hypothesis that the Instrumental Re-learning system only processes valence information, and remains to be explored in future work.

Chapter 6

Developmental psychopathy and instrumental learning and re-learning

6.1: Introduction

It has been suggested that developmental psychopathy arises in part from early amygdala dysfunction (Blair, Morris, Frith, Perrett & Dolan, 1999; Blair & Frith, 2000). Chapter 4 reviewed some of the evidence for the importance of the amygdala in instrumental learning in animals (e.g., Killcross, Robbins & Everitt, 1997), and demonstrated a severe impairment in instrumental learning in BM, a patient with early left amygdala damage. Thus, the position that developmental psychopathy is associated with amygdala dysfunction predicts an impairment in instrumental learning in this population. Other researchers have proposed that orbitofrontal cortex dysfunction contributes to developmental psychopathy (Anderson, Bechara, Damasio, Tranel & Damasio, 1999; LaPierre, Braun & Hodgins, 1995). Chapter 5 reviewed some of the evidence for the importance of the orbitofrontal cortex in instrumental re-learning (e.g., Iversen & Mishkin, 1970; Rolls, Hornak, Wade & McGrath, 1994; Dias, Robbins & Roberts, 1996), and demonstrated a re-learning impairment in two patients with orbitofrontal cortex damage. Thus the orbitofrontal cortex dysfunction account of developmental psychopathy predicts an impairment in instrumental re-learning in this population. This chapter explores the hypotheses that developmental psychopathy is associated with amygdala and/or orbitofrontal cortex dysfunction by investigating whether the disorder is associated with an impairment in instrumental learning and/or instrumental re-learning.

6.2: Developmental psychopathy

Clinical descriptions of psychopathy characterise psychopathic individuals as callous, with a diminished capacity for remorse, reduced affect, and poor behavioural control (e.g., Cleckley, 1950; Hare, 1991). The Psychopathy Checklist-Revised (PCL-R) is an empirically based list of behavioural features of psychopathy commonly used for classification (Hare, 1991; see table 6.1). Psychometric analysis of the PCL-R consistently identifies two factors. Factor 1 corresponds to affective and interpersonal traits. This factor tends to be stable across the life-span, and is unrelated to socio-economic factors or IQ. This independence from socio-economic factors suggests a possible biological basis to this component of psychopathy. In contrast, Factor 2 (antisocial lifestyle) does correlate with socio-economic status, and varies with age (Hare, 1991; Harpur, Hare & Hakstein, 1989; Harpur & Hare, 1994).

Factor 1	**Factor 2**
• Glibness/ superficial charm	• Need for stimulation
• Grandiose sense of self worth	• Parasitic lifestyle
• Pathological lying and deception	• Poor behavioural control
• Conning/ lack of sincerity	• Early behavioural problems
• Lack of remorse or guilt	• Lack of realistic long term plans
• Lack of affect and emotional depth	• Impulsivity
• Callous/ lack of empathy	• Irresponsible behaviour
• Failure to accept responsibility for own actions	• Frequent marital relationships
	• Promiscuity
	• Juvenile delinquency
	• Revocation
	• Criminal versatility

Table 6.1: Hare's Psychopathy Checklist-Revised (Hare, 1991)

6.3: Evidence for amygdala dysfunction in developmental psychopathy

Three lines of evidence suggest that developmental psychopathy may be associated with amygdala dysfunction. First, psychopathic individuals show attenuated fear conditioning (e.g., Hare, 1965a; Hare, 1965b; Hare, Frazelle & Cox, 1978; Aniskiewica, 1979). Reduced fear conditioning is observed following amygdala damage in patients (Bechara, Tranel, Damasio & Damasio, 1995; Phelps, LaBar, Anderson et al., 1998; Bechara, Damasio, Damasio & Lee, 1999) and experimental animals (LeDoux, 1998). A second indicator of amygdala dysfunction in psychopathy is the reduction of the normal augmentation of the startle reflex response following a visual threat stimulus (Patrick, Bradley & Lang, 1993). Significantly reduced potentiated startle is seen following amygdala damage in humans (Angrilli, Mauri, Palomba, et al., 1996). In addition, animal work by Davis and colleagues has established a role for the central nucleus of the amygdala in the potentiation of the startle response (e.g., Hitchcock & Davis, 1986).

A third line of evidence for the amygdala dysfunction position comes from recent work identifying the role of the amygdala in the processing of fearful and sad emotional expressions. Neuropsychological studies have revealed that amygdala damage most frequently compromises the recognition of expressions of fear, and then sadness and anger (Adolphs, Tranel, Damasio & Damasio, 1994; Adolphs, Tranel, Damasio & Damasio, 1995; Calder, Young, Rowland & Perrett, 1996; Scott, Young, Calder et al., 1997; Broks, Young, Maratos et al., 1998; Adolphs, Tranel, Hamann et al., 1999; Adolphs & Tranel, 1999; Sprengelmeyer, Young, Schroeder et al., 1999). Functional imaging studies implicate the amygdala in the processing of fearful and sad expressions, although not angry expressions (e.g., Morris, Frith, Perrett et al., 1996; Whalen, Shin, McInerney & Rauch, 1998; Blair et al., 1999). Psychopathic individuals are impaired in the recognition of fearful facial expressions (Mitchell & Blair, in prep), and are selectively hyporesponsive autonomically to fearful and sad expressions (Blair, Jones, Clark & Smith, 1997). Children with psychopathic tendencies are impaired in both fear and sadness recognition (Blair, Colledge, Mitchell & Murray, 2000). Thus functional imaging and neuropsychological research suggests that the emotion processing impairments seen in psychopathic individuals may arise from amygdala dysfunction.

Thus there is increasing evidence that developmental psychopathy is associated with amygdala dysfunction. The amygdala dysfunction position predicts an impairment in instrumental learning. The following section outlines the current evidence for an instrumental learning impairment in psychopathic individuals.

6.4: Developmental psychopathy and instrumental learning

Assessment of instrumental learning in psychopathic individuals is based mostly on variants of two different tasks. The first is a card playing task developed by (Newman, Patterson & Kosson, 1987). In this task, the probability of punishment increases by 10% every 10 card plays, from 10% to 100% probability of punishment. Psychopathic individuals and children with psychopathic tendencies or conduct disorder play significantly more cards than controls, and thus do considerably worse on the task (Shapiro, Quay, Hogan & Schwartz, 1988; Newman, Patterson, Howland & Nichols, 1990; Fonseca & Yule, 1995; Newman et al., 1987; Fisher & Blair, 1998).

The second type of task on which psychopathic individuals have been frequently assessed are object discrimination go/no-go tasks. For example, in the original version of this task, developed by Newman, Widom & Nathan (1985), stimuli are two-digit numbers. Four of these numbers are positively reinforced (S+'s) and four are negatively reinforced (S-'s). On each trial, a stimulus is presented and subjects are rewarded for responding to S+ stimuli and punished for responding to S- stimuli. It has been frequently found that psychopathic individuals make significantly more errors of commission than controls, but with the same numbers of errors of omission as controls (e.g., Lykken, 1957; Schmauk, 1970; Newman & Kosson, 1986; Thornquist & Zuckerman, 1995; although see Schmauk, 1970; Arnett, Howland, Smith & Newman, 1993; for some exceptions to this finding).

It is possible to understand psychopathic individuals' impaired performance on the card-playing and go/no-go tasks in terms of an instrumental learning impairment. In these tasks, there are both approach and avoid contingencies. The rewards gained by the approach behaviours may be sufficient to establish a stimulus-response based approach tendency (McDonald & White, 1993; Balleine & Dickinson, 1998; Burns et al., 1999). An inability to learn the response outcome of punished approach responses could explain the psychopathic individuals' large number of commission errors.

6.5: The BIS dysfunction and response set modulation accounts of instrumental learning deficits

In fact, psychopathic individuals' impairments on the Newman card playing task and go/no-go tasks have not generally been interpreted in terms of an instrumental learning deficit. There are two main accounts of psychopathic individuals' impaired performance. The first account is that psychopathic individuals' impairments are passive avoidance errors, arising from a dysfunctional Behavioural Inhibition System (e.g., Fowles, 1980; Quay, 1993). The second account is that psychopathic individuals' errors arise from a deficit in response set modulation (e.g., Newman et al., 1987; Patterson & Newman, 1993; Wallace, Vitale & Newman, 1999). These two accounts are discussed briefly below.

The large number of commission errors made by psychopathic individuals compared with their normal performance on omission errors has been regarded as demonstrating an impairment in passive avoidance. An impairment in passive avoidance is predicted by the position that psychopathic individuals suffer from a dysfunction of the Behavioural Inhibition System (BIS) (e.g., Gray, 1982; Gray, 1985; Fowles, 1980; Quay, 1993). Thus, psychopathic individuals are predicted to be insensitive to cues that trigger the BIS, that is, signals of punishment and unexpected non-reward (e.g., Gray, 1982). In contrast, the Behavioural Approach System (BAS) is hypothesised to be intact, thus it is predicted that psychopathic individuals will show relatively heightened sensitivity to cues that trigger the BAS, that is, signals of reward. In line with this account, Scerbo, Raine, O'Brien, et al. (1990) found that adolescent psychopaths made significantly fewer errors of omission than did controls on Newman's go/no-go task.

A second account that extends upon the notion of an impaired BIS in psychopathy has been proposed by Newman and colleagues. They propose a "response set modulation" deficit account of psychopathy (e.g., Patterson & Newman, 1993; Newman, Schmitt & Voss, 1997). It is argued that, *"once psychopaths adopt a response set for reward, they have difficulty attending to competing response contingencies"* Newman et al., 1987: p145. In support, Newman and colleagues have found evidence that the passive avoidance impairment in psychopaths is constrained to conditions in which both approach and avoidance contingencies are present. For example, Newman & Kosson (1986) modified the object discrimination go/no go task so that subjects were punished for commission and omission errors, but were not rewarded for correct responses or inhibition of responses. On this modified task, although both psychopathic individuals and controls made significantly more commission errors than in the task involving both rewards and punishments, psychopathic individuals' performance was no worse than controls (Newman &

Kosson, 1986). Thus it was argued that psychopathic individuals are only impaired when they need to switch from a dominant approach response set to an avoid response set.

There is also evidence that reducing the development of a dominant response set by requiring subjects to focus on both the approach and avoid contingencies of the task also abolishes the psychopathic individuals' impairment (Newman et al., 1990). Moreover, Newman and colleagues have found that introducing a delay between trials eliminates the psychopathic individuals' impairment on passive avoidance tasks (Newman et al., 1987; Arnett et al., 1993). It is argued that the delay promotes the processing of information relevant to the need to change to the non-dominant response set (Wallace et al., 1999).

6.6: Evidence for orbitofrontal cortex dysfunction in developmental psychopathy

Orbitofrontal cortex damage, both early and in adulthood, often results in severely anti-social behaviour, including violent aggression (e.g., Grafman, Schwab, Warden et al., 1996; Anderson et al., 1999; Blair & Cipolotti, 2000). There are three main accounts of anti-social behaviour following orbitofrontal cortex damage. First, Damasio and colleagues have proposed that orbitofrontal cortex patients' abnormal behaviour arises from an impairment in somatic marker generation (Damasio, Tranel & Damasio, 1991; Bechara, Damasio, Damasio & Anderson, 1994; Damasio, 1994; Bechara, Damasio, Tranel & Damasio, 1997). It is argued that the patients' poor decision-making, including their anti-social behaviour, results from an absence of somatic states that in normal individuals signal the possible negative consequences of the behaviour being considered. Damasio and colleagues have suggested that early impairment of the somatic marker system might underlie developmental psychopathy (e.g., Damasio, 1994; Anderson et al., 1999). However, there are three difficulties with this account of developmental psychopathy. First, patients with both early and adulthood ventromedial frontal lobe damage are impaired on the Four Pack Gambling task described in Chapter 1 (section 1.6) (e.g., Bechara et al., 1994; Anderson et al., 1999). In contrast, psychopathic individuals have been found to perform as well as controls on this task (Schmitt, Brinkley & Newman, 1999). Second, ventromedial frontal lobe patients are generally autonomically hyporesponsive to emotional visual stimuli under passive viewing conditions (Damasio, Tranel & Damasio, 1990), and Bechara, Damasio & Damasio (2000) have argued that this inability to process the emotional attributes of a stimulus may contribute to their failure to generate somatic states. However, psychopathic individuals show normal autonomic arousal responses to emotional visual stimuli (Patrick et al., 1993). Thus there is no evidence of poor somatic marker generation in psychopathic individuals. Third, patients with early or late ventromedial frontal lobe damage show reactive aggression, whereas developmental psychopaths tend to show instrumental aggression (Cornell, Warren, Hawk, et al., 1996).

A second account of the behavioural disturbance in acquired sociopathy has been proposed by Blair & Cipolotti (2000). They suggest that the orbitofrontal cortex forms part of a system that is activated by other individuals' angry expressions and perhaps representations of situations associated with anger. The consequence of activating this system is the modulation of current behavioural responding. In line with this account, angry faces activate orbitofrontal cortex (Blair et al., 1999). Consistent with the hypothesis that the processing of angry faces is important in modulating aggression, both alcohol and diazepam selectively impair the ability of healthy individuals to process angry expressions (Borrill, Rosen & Summerfield, 1987; Blair & Curran, 1999), and both of these drugs are associated with increased risk for reactive aggression (e.g., Bond, Curran, Bruce, O'Sullivan & Shine, 1995; Dougherty, Bjork, Bennett & Moeller, 1999). However, the social response reversal system is not thought to be impaired in developmental psychopathy (Blair & Cipolotti, 2000; Blair & Frith, 2000). Developmental psychopaths are unimpaired in their autonomic responsiveness to, and recognition of, anger (Blair et al., 1997; Mitchell & Blair, in prep). In addition, as noted earlier, psychopathic individuals tend to display instrumental aggression rather than reactive aggression (Cornell et al., 1996).

A third account of anti-social behaviour in orbitofrontal cortex patients is that it arises from their instrumental re-learning impairment (Rolls, 1996; 2000). This failure is evidenced as a tendency to continue to respond when responses are no longer rewarded, and has been observed in formal testing of patients and experimental animals (e.g., Butter, 1969; Rolls et al., 1994; Dias et al., 1996). (Rolls has argued that this impairment reflects an inability to re-learn stimulus-reinforcement associations, but as noted in the previous chapter, the data are currently also consistent with an inability to inhibit a previously rewarded response). Two studies suggest that the instrumental re-learning function of the orbitofrontal cortex is impaired in adults with psychopathy. First, LaPierre et al. (1995) found that psychopathic individuals made significantly more commission errors than controls in the reversal phase of a visual discrimination go/no-go reversal task. Such an impairment is also seen in orbitofrontal cortex patients (Rolls et al., 1994). Second, Mitchell & Blair (unpublished data) found that psychopathic adults were impaired in the reversal component of the ID-ED task. In this component of the task, subjects must learn that a particular shape or line is no longer associated with reward. Reversal learning in this task has been shown to be sensitive to orbitofrontal cortex damage (Dias et al., 1996). There is therefore preliminary evidence that psychopathy is associated with an instrumental re-learning deficit.

Thus there is preliminary evidence for possible orbitofrontal cortex dysfunction in psychopathy. It is therefore interesting to note that recent structural imaging studies have found evidence of abnormalities of the prefrontal cortex in individuals who show violent behaviour (e.g., Volkow & Tandredi, 1987; Raine, Buchsbaum & LaCasse, 1997; Raine, Meloy, Bihrle et al., 1998; Critchley, Simmons, Daly et al., 2000; Raine, Lencz, Bihrle, LaCasse & Colletti, 2000). However, there are two important caveats to these data. First, these studies do not specifically implicate abnormalities of the orbitofrontal region of the prefrontal cortex. Second, it is important to note that the violent subjects in these studies were not, at least predominantly, psychopathic individuals. Thus these findings do not provide evidence that psychopathy is associated with orbitofrontal cortex dysfunction.

6.7: Summary and Experimental Aims

There is considerable indirect evidence that developmental psychopathy is associated with amygdala dysfunction (Blair & Frith, 2000). There is also some preliminary evidence to suggest that developmental psychopathy is associated with orbitofrontal cortex damage (e.g., LaPierre et al., 1995; Mitchell & Blair, unpub.). It is therefore predicted that impairments in instrumental learning and/or re-learning will be seen in this population. These two predictions were tested using an instrumental learning and re-learning task, the Four Token Snake task.

6.8: Experimental Investigation

6.8.1: Subjects

The participants were 18 inmates from Holloway or Grendon prisons. In accordance with the literature and the guidelines of the PCL-R (e.g., Hare, 1991), the psychopathic group was made up of individuals scoring 30 or above on the checklist, while the non-psychopathic group was made up of individuals scoring less than 20. PCL-R scores were assessed by an experienced rater using file notes and interviews. Table 6.2 shows the mean, standard deviations, and ranges of age, PCL-R scores, and Ravens matrices scores of the psychopathic individuals and the control subjects. An independent samples t-test revealed that there were no significant differences between the two groups in age, t(16) = 1.7, p = ns. There were also no significant differences between the two groups in Ravens matrices score, t(16) = 0.8, p = ns.

Group	**Age** **Mean (s.d.)** **Range**	**PCL-R Total score** **Mean (s.d.)** **Range**	**Ravens matrices** **Mean (s.d.)** **Range**
Psychopathic individuals (n = 9)	30 (6) 24-44	34 (2) 31-38	8 (2) 6 –11
Controls (n = 9)	37 (11) 24-54	10 (3) 4-14	9 (3) 4-12

Table 6.2: Mean, standard deviation, and range of age, PCL-R Total score and Ravens matrices score for psychopathic and control groups.

6.8.2: Procedure

The subjects were tested in one of the interview rooms attached to whichever ward the subject was housed on. All subjects gave informed consent. Subjects were given the Four-Token Snake task. Experimental details of this task have been given previously in Chapter 4 (section 4.6.2).

6.8.3: Data analysis

Scoring of token choices was as in Chapter 5. A trial was scored as "correct" if the subject chose the token with the largest value. A trial was scored as "incorrect" if the subject chose the token with the least value. If the tokens were of equal value, the trial was not scored. In the two reversal phases, phases 2 and 3, data points were excluded if the subject had not had the opportunity to learn the points value of both tokens in that phase.

6.9: Results

The performance of the psychopathic individuals and the controls are shown in figure 6.1, and table 6.3 presents the means, standard deviations, and ranges of percentage correct token choices, in all three phases of the experiment.

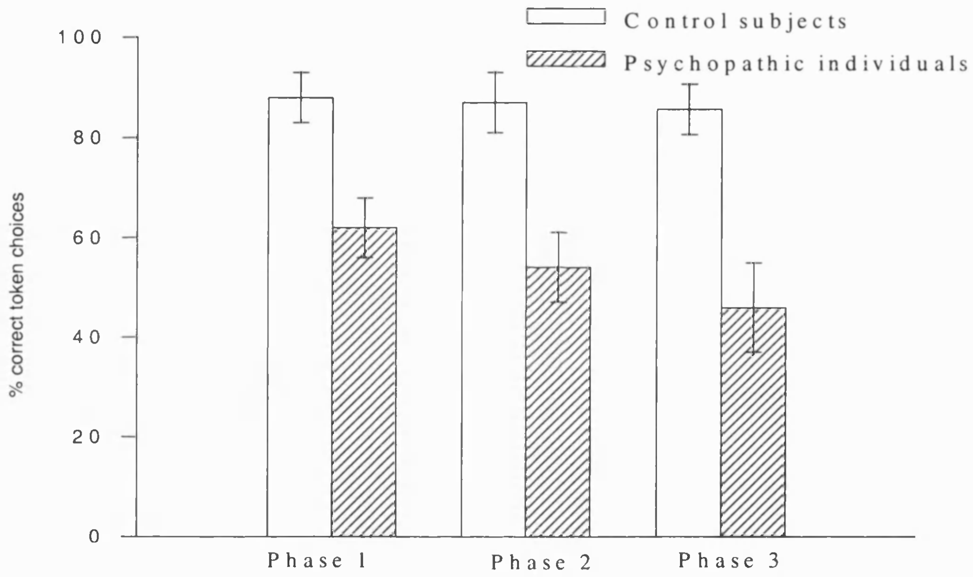

Figure 6.1: Performance of psychopathic individuals and controls in phase 1 (expectation acquisition), phase 2 (first reversal) and phase 3 (second reversal) of the Four-Token Snake task.

The performance of the two groups was compared using a split plot ANOVA with within-subjects factor Phase (phase 1, 2, 3) and between-subjects factor Group (Psychopathic vs. Control). The analysis revealed the predicted main effect of Group, F(1,16) = 19.4, p < .0001. Psychopathic individuals performed worse than the controls in all three phases of the task. The main effect of Phase was not significant, F(2, 32) = 2.0, p = .ns. There was no significant Group by Phase interaction, F(2, 32) = 1.1, p = ns.

Phase	Psychopaths		Controls	
	Mean (s.d.)	Range	Mean (s.d.)	Range
Phase 1 (expectation acquisition)	62 (17)	42-100	88(14)	67-100
Phase 2 (reversal 1)	54 (21)	18-88	87 (18)	46-100
Phase 3 (reversal 2)	46 (27)	0-83	86 (15)	67-100

Table 6.3: Mean, standard deviation, and range of percent correct token choices for psychopathic individuals and controls in phases 1, 2, and 3 of the Four Token Snake task.

A one-sample t-test revealed that the performance of the psychopathic individuals in phase 1 was not significantly better than chance, t(8) = 2.1, p = ns. Instrumental learning in phase 2 (reversal 1) was also assessed. Performance in phase 2 on token combinations in which neither token had changed value (the no valence change token combination) was investigated. The psychopathic individuals chose the correct token with a mean frequency of 76% (s.d. = 25), and a one-sample t-test revealed that this performance was significantly better than chance, t(8) = 3.1, p < .05. Thus overall the pattern of data suggest that instrumental learning was not entirely absent.

Since the psychopathic group performed above chance on the no valence change token combination in phase 2, it was possible to explore their instrumental re-learning performance. Both of the no valence change tokens in phase 2 changed valence in phase 3. Therefore performance for this token combination phase 3 reflects the psychopathic groups' ability to re-learn instrumental responses. The psychopathic subjects chose the correct token with a mean frequency of 44% (s.d. = 50). A one-sample t-test revealed that this performance was not significantly better than chance, t(8) = .36, p = ns. Controls chose the correct token with a mean frequency of 86% (s.d. = 38). An independent samples t-test revealed that the performance of the control subjects was significantly better than that of the psychopathic individuals, t(13) = 1.8, p < .05; one-tailed.

6.10: Discussion

This experiment revealed an impairment in instrumental learning in a group of psychopathic individuals. Although the psychopathic individuals performed significantly worse than the controls throughout the experiment, there was evidence that some instrumental learning took place. This finding was predicted by the hypothesis that psychopathy is associated with amygdala dysfunction (Blair & Frith, 2000). Since some instrumental learning did occur in the psychopathic individuals, it was also possible to assess instrumental re-learning in this group. An impairment was also observed in instrumental re-learning. This finding was predicted by the hypothesis that psychopathy is associated with orbitofrontal cortex dysfunction (LaPierre et al., 1995).

6.10.1: Implications for neurological dysfunction in developmental psychopathy

The finding of an instrumental learning impairment in a group of psychopathic individuals is consistent with the hypothesis that developmental psychopathy is associated with amygdala dysfunction (Blair et al., 1999; Blair & Frith, 2000). Animal research discussed in Chapter 4 indicates an involvement of the (basolateral) amygdala in instrumental learning (e.g., Killcross et al., 1997). Moreover, BM, a patient with early left amygdala damage, performed at chance on the same instrumental learning task used in this experiment (see Chapter 4). Thus the instrumental learning impairment observed in the psychopathic individuals is consistent with dysfunction of the basolateral amygdala. It should be noted that brain regions additional to the basolateral amygdala are also thought to be involved in instrumental learning, for example, the ventral striatum (e.g., Cador, Robbins & Everitt, 1989; Everitt, Cador & Robbins, 1989; Everitt et al., 1992). While it would be possible to attribute the impairment seen in the psychopathic individuals to dysfunction of the ventral striatum, this hypothesis could not explain the impaired fear conditioning, reduced potentiated startle reflex, and selective emotion expression processing impairments seen in psychopathic individuals, as discussed in section 6.3.

It is suggested here that, as modelled in the "impaired" Valence Change Reset model in Chapter 4, psychopathic individuals are slow in learning associations between instrumental responses and their outcomes. In the current task, critical trials always involved the combination of a positive and a negative token together. Thus the psychopathic individual's impairment could have been due to a failure to learn which tokens predicted reward, or which tokens predicted punishment, or both. Further research is required to investigate whether appetitive and aversive instrumental learning are both impaired in psychopathy. As can be seen in table 6.3, there was a range of performance in the psychopathic group, in all three phases of the experiment. It would be possible to use the Valence Change Reset model to simulate the performance of individual psychopaths, by varying the extent to which the learning rate was reduced relative to the "normal" level. It would then be possible to use the model to predict the performance of individuals on other instrumental learning tasks.

An impairment in instrumental re-learning was also observed in the psychopathic individuals. Thus even though the psychopathic individuals did not learn response outcome expectations as well as control subjects (as evidenced by their poor instrumental learning), they were impaired in re-learning even these presumably weaker expectations of reinforcement. It is suggested here that the psychopathic individuals' impairment might reflect damage to an instrumental re-learning system that rapidly resets response outcome expectations following valence changes. The psychopathic individuals' instrumental re-learning deficit is consistent with previous studies that have found that psychopathic individuals are impaired in instrumental re-learning tasks (LaPierre et al., 1995; Mitchell & Blair, unpub.). Instrumental re-learning ability is sensitive to orbitofrontal cortex damage (e.g., Rolls et al., 1994; Dias et al., 1996; Iversen & Mishkin, 1970). Thus the current findings, in line with the research by LaPierre et al. (1995) and Mitchell & Blair (unpub.), suggest the possibility that developmental psychopathy is associated with dysfunction of the orbitofrontal cortex or of the connections between the amygdala and the orbitofrontal cortex (Amaral, Price, Pitkanen & Carmichael, 1992).

Bechara and colleagues have suggested that psychopathy might arise from impaired somatic marker generation (Damasio, 1994; Anderson et al., 1999). It is therefore interesting to note their statement that their ventromedial frontal patients, also hypothesised to be impaired in somatic marker generation, are not impaired on tasks involving response inhibition (Bechara et al., 2000). In contrast with these patients, the evidence currently suggests that psychopathic individuals are impaired in response inhibition, but unimpaired in somatic marker generation (Patrick et al., 1993; LaPierre et al., 1995; Schmitt et al., 1999). In addition, as noted in the introduction, the syndrome that results from early ventromedial frontal lobe damage does not result in all the behavioural components of developmental psychopathy.

Such patients show some behaviours characteristic of psychopathy such as deceit, lack of remorse, and a promiscuous and anti-social life-style (Anderson et al., 1999). However, they differ to developmental psychopaths in showing reactive rather than instrumental aggression.

Thus currently the evidence does not suggest that early orbitofrontal cortex damage alone can cause the psychopathic syndrome. Damage to the amygdala, for example, may also be necessary. Alternatively, orbitofrontal cortex dysfunction may arise later in life. With regard to this speculation, it is interesting to note that performance on the reversal component of the Intra-dimensional-Extra-dimensional Shift task was normal in a population of children with psychopathic tendencies (Colledge, in prep). It may then be that orbitofrontal cortex damage accrues over the life-time of the developmental psychopath. This damage may result from, perhaps, a lack of input from the amygdala, or possibly the drug-taking life-style of the typical developmental psychopath. In line with this latter speculation, the ventromedial prefrontal cortex appears to be vulnerable to amphetamine abuse (Rogers, Everitt, Baldacchino et al., 1999).

6.10.2: BIS deficit and response set modulation deficit accounts of findings

The current findings are also consistent with the Behavioural Inhibition System (BIS) deficit view of psychopathy (Fowles, 1980; Gray, 1982; Quay, 1993). According to this account, psychopathic individuals are impaired in their ability to learn to inhibit responses towards stimuli associated with punishment. It is interesting to note that the BIS deficit position proposes that psychopaths have an unimpaired Behavioural Approach System (BAS). It might then be expected that in the current task, the psychopathic individuals would have learnt to approach the positive tokens, even if they did not learn to avoid the negative tokens. The instrumental learning seen in the psychopathic individuals, albeit slow, might perhaps reflect the intact functioning of the BAS system.

The BIS deficit hypothesis can also explain the instrumental re-learning impairment observed in the psychopathic individuals. The BIS is hypothesised to be triggered by unexpected reinforcements, resulting in behavioural inhibition and the replacement of non-rewarded responses (e.g., Gray, 1982). The BIS deficit account can thus provide the most parsimonious account of the data in this chapter, as it can explain the psychopathic individuals' impaired performance in all phases of the task. However, it should be noted that in the previous chapter, a dissociation between instrumental learning and re-learning was seen in patients in CM and DJ. These patients were intact in instrumental learning but impaired in instrumental re-learning. Thus if a BIS deficit is hypothesised to result in both instrumental learning and re-learning impairments in the Four Token Snake task, then a BIS deficit account cannot easily explain the performance of CM and DJ.

Both the instrumental learning and re-learning impairments seen in the psychopathic individuals can also be explained in terms of a response set modulation deficit. (Newman et al., 1997) state that,

> *" ... the response modulation hypothesis predicts that the deficient avoidance learning of psychopathic individuals will be relatively specific to conditions requiring psychopaths to suspend a dominant response set to process negative feedback ..." Newman et al., 1997: p564.*

Thus with regard to the instrumental learning phase of the experiment, it may be that the psychopathic individuals failed to adequately process the negative reinforcement received following choice of negatively valenced tokens. With regard to the instrumental re-learning phase of the experiment, it may be that having developed a dominant response set to approach positively valenced tokens, the psychopathic individuals are unable to suspend this response set following negative reinforcement.

6.11: Summary

The present study provides evidence that psychopathic individuals are impaired in instrumental learning. This was predicted by the hypothesis that developmental psychopathy is associated with amygdala dysfunction (Blair et al., 1999; Blair & Frith, 2000). Evidence of an instrumental re-learning impairment was also found. This is consistent with suggestions that the orbitofrontal cortex may be dysfunctional in psychopathy (e.g., LaPierre et al., 1995; Mitchell & Blair, unpub.). In particular, the current findings suggest that the function of the orbitofrontal cortex that is impaired in psychopathy is that of reversing instrumental responses following unexpected reinforcements. The current findings were also consistent with

Behavioural Inhibition System deficit, and Response Set Modulation deficit accounts of psychopathy (e.g., Fowles, 1980; Gray, 1982; Newman et al., 1990; Newman et al., 1997; Wallace et al., 1999).

Chapter 7

Investigating the behavioural similarities between the effects of early amygdala damage and developmental psychopathy

7.1: Introduction

In Chapter 6, a group of psychopathic individuals were found to be severely impaired on a test of instrumental learning. A strong prediction from the amygdala dysfunction hypothesis of psychopathy is that BM, although not diagnosed as a psychopath, will nonetheless show behavioural impairments similar to those seen in developmental psychopaths, due to his early amygdala damage. Psychopathic individuals and children with psychopathic tendencies are selectively impaired in the processing of emotional expressions (Blair, Jones, Clark & Smith, 1997; Blair, Colledge, Mitchell & Murray, 2000; Mitchell, Colledge & Blair, in prep.). BM's performance on emotion expression processing tasks was therefore assessed. His performance on these tasks was compared with the findings of previous research with psychopathic individuals and patients with amygdala damage acquired early and late in life. In addition, BM was assessed for behavioural signs of the psychopathic syndrome, using the Psychopathy Checklist-Revised (Hare, 1991).

7.2: Emotion expression processing in psychopathic individuals and the possible role of the amygdala

It was noted in the previous chapter that one indirect line of evidence that developmental psychopathy is associated with amygdala dysfunction is that both developmental psychopaths and amygdala patients are impaired in processing emotional expressions of fear and sadness. Recently, Mitchell et al. (in prep.) have found that psychopathic individuals are selectively impaired in the recognition of fear. In children with psychopathic tendencies, both fear and sadness recognition are impaired (Blair et al., 2000). Moreover, psychopathic individuals are hyporesponsive autonomically to fearful and sad expressions, but show normal autonomic responsiveness to angry expressions (Blair et al., 1997). Neuropsychological data implicate the amygdala in the processing of fear, sadness and anger. At least two of these – fear and sadness – appear to be abnormally processed in developmental psychopathy.

7.2.1: Neuropsychological and functional imaging evidence for the role of the amygdala in the processing of fear and sadness

A number of single-case studies of amygdala patients have demonstrated impairments in the processing of facial and auditory expressions of emotion. Recognition has been assessed in both modalities, and judgements of intensity in the visual modality. These findings are summarised in table 7.1 (for review, see Fine & Blair, 2000). While an impairment in each of the six universal expressions has been observed at least once, the most frequently observed impairment is that of fear. Ten of the thirteen patients tested show a recognition impairment and/or significantly lowered intensity rating for this emotion (Adolphs, Tranel, Damasio & Damasio, 1994; Adolphs, Tranel, Damasio & Damasio, 1995; Calder, Young, Rowland & Perrett, 1996; Scott, Young, Calder et al., 1997; Broks, Young, Maratos et al., 1998; Sprengelmeyer, Young, Schroeder et al., 1999; Adolphs & Tranel, 1999; Adolphs, Tranel, Hamann et al., 1999). Sadness and anger are the next most frequently impaired with a deficit seen in six of 13 patients (Scott et al., 1997; Broks et al., 1998; Sprengelmeyer et al., 1999; Adolphs & Tranel, 1999; Adolphs et al., 1999). The reasons for the heterogeneity of impairments are not clear, although selectivity of lesions is probably a factor (e.g., Anderson & Phelps, 1998; Hamann, Stefanacci, Squire et al., 1996). A broader impairment in emotion recognition might reflect neural damage to surrounding brain regions, or other brain regions implicated in emotion expression recognition such as the anterior insula (e.g., Phillips, Young, Senior et al., 1997). Figure 7.1 shows how the frequency of impairment for different emotion expressions varies as the total number of emotion categories impaired is increased from most selective to least selective. Figure 7.1 qualitatively suggest again that, following fear, the processing of expressions of sadness and anger are the most vulnerable to amygdala damage. Functional imaging data implicate the amygdala only in the processing of fearful faces (e.g., Morris, Frith, Perrett et al., 1996; Brieter, Etcoff, Whalen et al., 1996; Whalen, Shin, McInerney & Rauch, 1998) and sad faces (Blair, Morris, Frith, Perrett & Dolan, 1999). In contrast, angry faces do not have the same effect on amygdala activity (Whalen et al., 1998; Blair et al., 1999).

Patient	FEAR	SAD	ANGER	SUR	DISG	HAPPY	No. categories impaired	Reference
DBB							0	Adolphs et al, 1999
DR (1)	FR		FR		FR			Calder et al, 1996
DR (2)	PV	P	PV				5	Scott et al, 1997
DR (3)	FI	FI	FI	FI	FI			Adolphs et al, 1999
EP (1)								Hamann et al, 1996
EP (2)			FI				1	Adolphs et al, 1999
GT (1)							0	Hamann et al, 1996
GT (2)								Adolphs et al, 1999
JC	FR	FR	FR			FR	4	Broks et al, 1998
JM	FI	FI	FI		FI		4	Adolphs et al, 1999
NM	FR, V	FR					2	Sprengelmeyer et al, 1999
RB	FR						1	Broks et al, 1998
RH (1)	P	P					3	Adolphs & Tranel, 1999
RH (2)			FI					Adolphs et al, 1999
SE (1)	FR						2	Calder et al, 1996
SE (2)	FR							Broks et al, 1998
SE (3)				FI				Adolphs et al, 1999
SM (1)	FR		FI	FI				Adolphs et al, 1994
SM (2)	FI, FR		FI	FI			3	Adolphs et al, 1995
SM (3)	FI			FI				Adolphs et al, 1999
SP (1)					P		3	Anderson & Phelps, 1998
SP (2)	FI	FI			FI			Adolphs et al, 1999
YW	FR						1	Broks et al, 1998
No. impaired	10	6	6	3	3	1		

FI = impairment in judgement of intensity of facial emotion expression; FR = impairment in recognition of facial emotion expression; P = impairment in emotional prosody recognition; V = impairment in emotional vocalization recognition.

Table 7.1: Statistically significant impairments on facial and auditory emotion processing tasks in 13 bilateral amygdala patients.

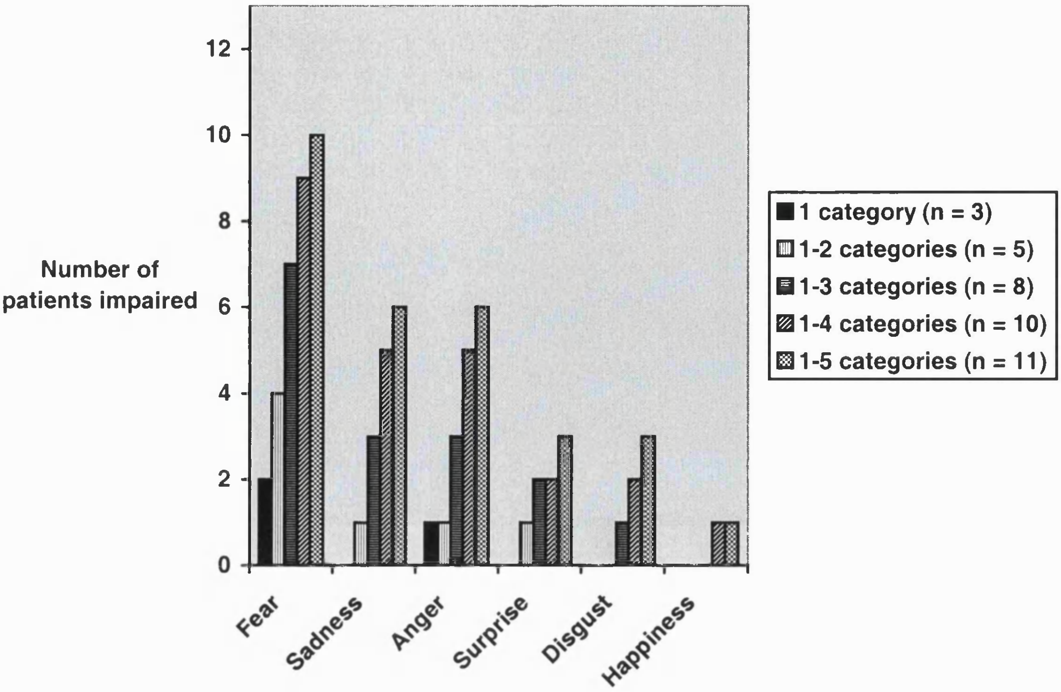

Figure 7.1: Number of patients impaired on each category of emotion depending on the selectivity of impairment in the sample.

7.3: Summary and aim of Experiments 1-3

Psychopathic individuals and children with psychopathic tendencies are impaired in the processing of fear and sadness (Blair et al., 1997; Blair et al., 2000; Mitchell et al., in prep.). Neuropsychological and functional imaging data also implicate the amygdala in the processing of these two emotional expressions (Adolphs et al., 1994; Adolphs et al., 1995; Calder et al., 1996; Scott et al., 1997; Broks et al., 1998; Whalen et al., 1998; Blair et al., 1999; Adolphs & Tranel, 1999; Adolphs et al., 1999). The aim of Experiments 1-3 was to investigate whether BM would be selectively impaired in the recognition of fear and sadness.

7.4: Experiment 1- Hexagon emotion recognition task

This task, based on the paradigm described by Calder et al, (1996), assesses recognition of six emotion expressions: surprise, happiness, anger, disgust, sadness and fearfulness. BM's performance was compared with that of five prison inmates with no psychiatric disorders.

The stimuli were continuous tone images in which two expressions were morphed together. The expressions were morphed from one to the next over a series of five stages. This was achieved by taking the two prototypes of each emotion and stretching them across so that "all the points representing the same features were aligned across images" (Calder et al., 1996). The emotion blends were, for example, 90% anger with 10% happiness, 70% anger with 30% happiness, 50% anger with 50% happiness, 30% anger with 70% happiness, 10% anger with 90% happiness. The expressions morphed were anger to happiness, happiness to surprise, surprise to fearfulness, fearfulness to sadness, sadness to disgust, and disgust back to anger. All the stimuli involved same face. There were 30 faces in total.

Each face was presented to subjects on a computer screen. There were six blocks of stimuli. In each block all 30 stimuli were presented in a randomised order. The first block was counted as practice trials, and the date for these trials was not recorded. Each stimulus was presented for 3 seconds and there was a 4-6 second interval between each stimulus during which the screen was blank. Subjects were presented with a list of 6 response options (surprise, happiness, anger, disgust, sadness and fearfulness) and were instructed to name the expression being displayed.

Table 7.2 shows the performance of BM on expression recognition compared with that of the control subjects. It can be seen that, contrary to prediction, BM's recognition of all emotions, including fear and sadness, were unimpaired.

7.5: Comment

BM's performance on a simple verbal labelling test of emotion recognition was not impaired for any of the six universal emotions. Interestingly, psychopathic individuals have also been found to perform normally on this emotion recognition task (Blair, unpublished data), yet they are impaired on a more sensitive task in which the threshold for recognition of an emotional expression is assessed (Mitchell et al., in prep.). The aim of Experiments 2 and 3 therefore was to assess BM's emotion recognition using two additional tasks. The first task (Morph emotion recognition task) assesses the ability to recognise emotional expressions with gaze straight ahead and averted. The second task (Threshold task) assesses sensitivity to the presence of an emotional expression. The Threshold task was used with psychopathic individuals and children with psychopathic tendencies, studied by Mitchell et al. (in prep.) and Blair et al. (2000), respectively. It was therefore of particular interest to ascertain whether BM would also be impaired on this task.

7.6: Experiment 2 – The Morph emotion recognition task

This task was designed by Murray & Perrett (*pers. comm.*). On each trial the participant is presented with a neutral face which morphs into an emotion expression: happy, sad, fear, disgust, surprise or anger. The subject must indicate which emotion he thinks is being expressed. For each emotion, on half of the trials the gaze of the model was averted to one side. There were six practise trials not included in the scoring. These were followed by 48 trials in 2 blocks, comprising eight trials of each emotion, presented in a random order. Thus the subject attained a score out of eight for each emotion. BM's performance was compared with unpublished norms (Perrett, *pers.com.*) Table 7.2 shows the mean recognition scores for BM and the comparison subjects. In line with prediction, BM's performance for sad faces was more than two standard deviations below the control mean. However, once again his performance on fearful faces was within the normal range.

7.7: Comment

In line with prediction, and like children with psychopathic tendencies, BM showed an impairment in the recognition of sadness (Blair et al., 2000). However, BM did not show an impairment in fear recognition. This is the most frequently observed emotion recognition impairment in patients with amygdala damage (see table 7.1). In addition, a fear recognition impairment has been seen in psychopathic individuals and children with psychopathic tendencies, using the Threshold task (Blair et al., 2000; Mitchell et al, in prep.). The aim of Experiment 3 was to assess BM's emotion expression recognition using this task.

Task (section)	**BM**	**CONTROLS**	
		Mean	s.d.
*Hexagon emotion recognition task (7.4)	Max = 20		
Sad	15	18.2	2.2
Fear	15	12.6	3.4
Happy	18	19.2	0.8
Anger	20	15.2	1.6
Surprise	20	18.2	1.3
Disgust	20	16.6	1.7
**Morph emotion recognition task (7.6)	Max = 8		
Sad	2	5.9	1.7
Fear	6	5.0	1.9
Happy	8	8	0
Anger	6	5.3	1.8
Surprise	6	5.2	1.5
Disgust	5	6.9	0.9
**Threshold task (7.8)			
Sad	3 errors	14.8	4.7
Fear	2 errors: 21[†]	14.2	3.2
Happy	6.3	9.0	3.3
Anger	13.3	13.2	5.4
Surprise	9.7	13.0	1.6
Disgust	13	13.8	4.5

[†] I.e., BM required 21 steps to identify the expression on the one trial in which he correctly identified a fearful face.

* Control data from non-psychiatric inmates

** Control data from unpublished norms.

Table 7.2: Performance of BM and control subjects on three emotion expression recognition tasks.

7.8: Experiment 3 - Threshold task

The Threshold task ascertains the threshold for the recognition of the presence of an emotional facial expression, as well as recognition of that emotion. This task was designed by Murray et al. (*pers.com.*). On each trial, the subject was presented with a neutral face which gradually morphed, under the experimenter's control, into an emotion expression; either happy, sad, fear, disgust, surprise or anger. The subject was asked to state what emotion s/he thinks that the face is pulling, as early on in the morphing process as possible. The face is morphed to its full emotion expression on each trial, and the participant is able to change his or her response if desired. There were six practice trials, one of each emotion, which were not included in the scoring.

These were followed by 18 trials, comprising three trials of each emotion, presented in a random order.

Each expression morphed from a neutral expression to a full emotion expression in 21 steps. The test was scored by calculating the mean number of steps required to guess the emotion. Thus, the best possible score was 1 and the worst possible score was 21. BM's performance was compared with unpublished norms (Perrett, *pers.com.*). Table 7.2 shows the mean number of steps for recognition for BM and the control subjects. BM incorrectly identified all three sad faces as fearful faces. He also incorrectly identified two of the fearful faces as surprised faces. This placed his performance for these two emotion expressions well below the performance of the comparison group. BM's performance on the remaining four emotion expressions was within the normal range.

7.9: Comment

In line with prediction, BM was found to be selectively impaired in the processing of facial expressions of fear and sadness. From the large number of errors that BM made for these two emotions (5/6 incorrect responses) it can be inferred that BM's impairment was not merely one of a heightened threshold of sensitivity to the presence of the emotion, but in reliably recognising expressions of fear and sadness. His performance on the remaining four universal emotion expressions was intact, as it was in the previous two tasks.

7.10 Discussion of Experiments 1-3

BM showed a selective impairment in sadness recognition in the Morph emotion recognition task, and impairments in both fear and sadness recognition in the Threshold task. These findings are in line with the prediction that early amygdala damage will be associated with selective impairment in the processing of fear and sadness expressions. It should be noted that while BM's performance on the Hexagon emotion recognition task was unimpaired, this could be explained by poorer task sensitivity. In two out of the three tasks used with BM, a selective emotion recognition impairment was seen. This highlights the importance of task-related factors which make it necessary to not rely on a single task, but to also use more sensitive tasks such as the Threshold task.

Blair & Frith (2000) have argued that fearful and sad facial expressions act as punishing unconditioned stimuli. It is argued that in healthy individuals, the pairing of actions that harm others with the aversive stimulus of the other's displayed fear/pain result in these actions, through classical conditioning, becoming perceived as aversive. This process is thought to underlie normal socialization (Blair, 1995), and social referencing, whereby the child learns from the mother's expression of which new objects they should be frightened (cf., Mineka & Cook, 1993). The early amygdala dysfunction hypothesis of developmental psychopathy stresses the known role of the amygdala in aversive conditioning (see LeDoux, 1998). It is suggested that early amygdala dysfunction attenuates the normal pairing of distress cues with actions that cause harm, contributing to the development of psychopathy (Blair et al., 1999; Blair & Frith, 2000). This could be due to weak representations of distress stimuli as well as an impairment in classical conditioning. Blair (1995) has argued that the inability to form associations between actions that cause harm with distress impairs the development of empathy and other moral emotions such as guilt and remorse. As demonstrated here, BM is impaired in the processing of fearful and sad faces. In addition, while fear conditioning has not been assessed, in Chapter 4 it was shown that BM was impaired in instrumental learning, which involves developing expectations of the outcome of responses. Blair's model therefore suggests that the impact of the cognitive deficits underlying BM's behavioural abnormalities might also have impaired BM's ability to show empathy, remorse, and guilt. These are characteristics features of psychopathy (e.g., Cleckley, 1950; Hare, 1991). The aim of Experiment 4 was to investigate whether BM demonstrates this aspect of the psychopathic syndrome.

7.11: Experiment 4 - Hare Psychopathy Checklist-Revised assessment

The aim of Experiment 4 was to assess BM for behavioural signs of psychopathy, in particular, lack of empathy, guilt and remorse. As described in Chapter 6 (section 6.2), the Psychopathy Checklist-Revised (PCL-R) is composed of 20 items (Hare, 1991). Factor 1 assesses "callous and unemotional behaviour" and Factor 2 assesses "impulsivity and conduct problems". Interestingly, Cooke & Michie (*pers. comm.*) have recently suggested that items from Factor 1 that index lack of empathy, remorse, guilt and shallow affect, form a subfactor which they term "deficient affective experience". The remaining items of Factor 1 are proposed to reflect "deceitful interpersonal conduct". BM was therefore scored for Factor 1 and Factor 2 of the PCL-R, and was also scored on the two subfactors proposed by Cooke & Michie. To score a subject on the PCL-R, the rater assesses how well the subject fulfils the characteristics specified in the item descriptors. For each item, e.g., "Pathological lying", the subject can score 0 ("absent)", 1 ("maybe/in some respects"), or 2 ("present"). This reflects the degree to which the item applies to the subject. The test manual specifies in detail the necessary criteria for a score of 0, 1, or 2 on each item. For BM, this information was collected from both file information and from interview, by an experienced rater.

BM scored 6/16 on Factor 1 ("callous/unemotional") and 3/18 ("impulsivity/conduct problems") on Factor 2, with a total score of 11/40. This was well below the threshold for diagnosis of psychopathy (threshold score = 30). However, further analysis of his score revealed that all of the points that BM scored on Hare's Factor 1 could be attributed to Cooke & Michie's "deficient affective experience" factor. BM's score on this factor was therefore high (6/8). BM's assessment for behavioural signs of psychopathy therefore revealed a lack of empathy, remorse and guilt, and shallow affect.

7.12: General Discussion

This study investigated the effect of early amygdala damage on facial emotion expression processing and the development of empathy. BM, a patient with early left amygdala damage, was given three tests assessing emotion expression recognition. His recognition of the facial expression of sadness was very impaired and there was also evidence of impairment in the recognition of fear. In contrast, BM showed normal recognition of the other four universal expressions. BM did not fulfil criteria for developmental psychopathy, but he scored highly on Cooke & Michie's "deficient affective experience" psychopathy factor, which reflects a severe deficiency in empathy.

7.12.1: Implications for the role of the amygdala in developmental psychopathy

BM has been shown to share three behavioural similarities with psychopathic individuals. First, both BM and the psychopathic individuals were impaired on an instrumental learning task (Chapters 4 and 6). Second, BM showed an impairment in the recognition of sad and fearful facial expressions. An impairment in the recognition of fear has been seen in psychopathic individuals (Mitchell et al., in prep.) and they fail to show normal autonomic responses to both sadness and fear (Blair et al., 1997). Children with psychopathic tendencies are impaired in the recognition of fear and sadness (Blair et al., 2000). Third, BM suffered from an absence of empathic feelings such as guilt and remorse. This is one of the critical features of psychopathy (Cleckley, 1950; Hare, 1991). These three findings are consistent with the hypothesis that early amygdala dysfunction may contribute to the development of the psychopathic syndrome. In particular, the data suggest that amygdala dysfunction may contribute to Hare's Factor 1 ("callous / unemotional behaviour"), or perhaps more specifically to Cooke & Michie's (*pers. comm.*) factor of "deficient affective experience". Interestingly, Factor 1 is thought to have a strong biological component (Hare, 1991).

The association between impaired processing of sadness and fear and a lack of empathy in both BM and developmental psychopaths (Blair et al., 1997; Mitchell et al., 2000) is suggestive that perhaps the processing of distress cues is important for normal moral development (Blair, 1995) and that this process relies upon the amygdala (Blair et al., 1999; Blair & Frith, 2000). One question that requires further investigation is the relative importance of the amygdala in development and in adulthood. Several abnormalities in social behaviour have been observed following amygdala damage in humans. Although it can often only be estimated when amygdala damage occurred, social deficits appear to be more severe the earlier the onset. These deficits range from mild to severe disturbances of social behaviour (Jacobsen, 1986; Tranel & Hyman, 1990; Fudge, Powers, Haber & Caine, 1997; Adolphs, Tranel & Damasio, 1998; Broks et al., 1998) to violent behaviour (Hayman, Rexer, Pavol, Strite & Meyers, 1998). Most interesting in the context of the current findings is a case study presented by Martinius (1983). Martinius (1983) described a 14 year old boy, RN, who killed another child. RN suffered a lesion that interrupted most of the fibre connections between the right nucleus amygdalae and the right middle and posterior temporal cortex. The cause of the damage was unknown, but hypoxic birth trauma was suggested as a possibility. Thus RN provides a second example of fatally violent behaviour associated with possible pre-natal damage to the amygdala.

While BM did show similarities to psychopathic individuals, he did not show the full behavioural syndrome of psychopathy. It is possible that psychopathy is associated with damage to brain regions other than the amygdala that were intact in BM, for example, the orbitofrontal cortex (LaPierre, Braun & Hodgins, 1995). As will be shown in the next chapter, BM showed no impairment in reversals in the Intra-dimensional-Extra-dimensional Shift task, unlike psychopathic individuals (Mitchell et al, unpub). This task is sensitive to orbitofrontal cortex damage (Dias, Robbins & Roberts, 1996). In addition, it should be noted that BM was diagnosed with Asperger's syndrome. Individuals with Asperger's syndrome show difficulties in social interactions and in the understanding of mental states (e.g., Happé, 1994; Frith, Happé & Siddons, 1994). It is therefore possible that BM did not show behaviours such as pathological lying and deception, and conning/manipulation because he did not have sufficient skill in the manipulation of mental states. BM's mental state ability is explored in the next chapter. Additionally, BM's difficulty in social interactions, described in Chapter 4 (section 4.4.1) would presumably make it impossible for him to cope with the high social demands of behavioural items such as promiscuity, frequent marital relationships, and a parasitic life-style.

7.12.2: Implications for the role of the amygdala in emotion expression processing

A number of researchers have highlighted the importance of the amygdala in processing fearful and angry faces, and used these data to support the hypothesis that the amygdala is involved in processing threat-related stimuli (e.g., Adolphs et al., 1995). However, BM's selective impairment in the recognition of fear and sadness, but not anger, does not support the "threat system" account of the amygdala. BM is the seventh patient to be impaired in sadness expression recognition, compared with six patients impaired in anger recognition. It remains unknown what cognitive deficit underlies amygdala patients' poor recognition of, and insensitivity to, certain emotional expressions. It may be that one response of the amygdala to emotional cues is to facilitate processing in other brain regions involved in expression recognition (e.g., Whalen, 1998).

7.13: Summary

Emotion expression recognition, and psychopathic tendencies were assessed in a forensic patient with early or congenital amygdala damage, BM. There were two main findings. First, BM was selectively impaired in the recognition of facial expressions of sadness and fear. This supports the hypothesis that the amygdala is concerned with processing distress-related stimuli (e.g., Blair et al., 1999). Second, BM showed an absence of empathic feeling characteristic of psychopathy. This was consistent with the hypothesis that amygdala dysfunction impairs the development of empathy and other moral emotions through a deficit in the processing of distress cues and/or the pairing of these cues with actions that cause harm to others (Blair, 1995; Blair et al., 1999; Blair & Frith, 2000).

Chapter 8

Dissociation between Theory of Mind and executive functions: A case study[5]

8.1: Introduction

Previous chapters have investigated and discussed the role of the amygdala in emotional learning, emotion expression recognition, and emotional responsiveness to others' distress. The detrimental impact of amygdala damage on emotional and social behaviour has led Brothers (e.g., Brothers, 1997) to suggest that the amygdala acts as a "social editor" that increases attention to all socially important stimuli. Brothers (1997) has proposed that a dysfunctional social editor results in an impairment in the ability to represent mental states. Baron-Cohen and colleagues have also argued that the amygdala is important for the processing of mental state information (Baron-Cohen, 1995; Baron-Cohen, Wheelwright, Bullmore, et al., 1999; Baron-Cohen, Ring, Bullmore et al., 2000). These positions predict that BM, a patient with early left amygdala damage, will be impaired in mental state processing. The first aim of this chapter was to test this prediction.

The second aim of the chapter was to determine BM's executive functioning. In the literature, there have been frequent claims that Theory of Mind is mediated by general executive functioning (e.g., Frye, Zelazo, Brooks & Samuels, 1996). It was therefore of interest to know whether any mental state processing impairment in BM was also associated with executive dysfunction, as predicted by such accounts. Such data are important with regard to models concerning the role of the amygdala in the development of Theory of Mind and the degree of dissociation between Theory of Mind and executive functioning.

8.2: Theory of Mind

Theory of Mind refers to the ability to attribute mental states to self and others, and to predict and understand other people's behaviour on the basis of their mental states (Premack & Woodruff, 1978). Operationally, individuals are credited with a Theory of Mind if they pass tasks designed to test their understanding that an individual may

[5] The data presented in this chapter have been accepted for publication in *Brain*.

hold a false belief. For example, in the classic false belief test (Wimmer & Perner, 1983), the subject is introduced to two dolls, Sally and Ann. Ann moves Sally's marble from the basket, where Sally placed it, to another hiding place while Sally is out of the room. The child is asked where Sally will look for her marble when she returns. Normally developing children of approximately 4 years correctly attribute a false belief to Sally, and predict that she will search in the original location, i.e., where Sally thinks her marble is (Wimmer & Perner, 1983). Severe impairments in theory of mind have been reported in individuals with autism, Asperger's syndrome and paranoid delusional schizophrenia (e.g., Frith, 1989; Happé, 1994; Frith & Corcoran, 1996; Corcoran, Cahill & Frith, 1997; Baron-Cohen et al., 1999).

8.3: The anatomical basis of theory of mind

Several attempts have been made to delineate the brain regions implicated in Theory of Mind. For example, Baron-Cohen (1995) has suggested a neural circuit that includes the amygdala, superior temporal sulcus and orbitofrontal cortex. In line with this, Baron-Cohen et al. (1999) used fMRI to measure brain activity during a task requiring the subject to infer the mental state of an individual from the expression of their eyes. Areas significantly activated by the task included the left amygdala. Interestingly, amygdala activation during this task was not seen in individuals with Asperger's syndrome, who were impaired on this task relative to controls. In an earlier study, Baron-Cohen and colleagues found activation in orbitofrontal cortex in a SPECT study during a task in which subjects had to decide which aurally presented words "described the mind or things the mind can do" (Baron-Cohen, Ring, Moriarty et al., 1994).

An alternative view of the neural circuitry for Theory of Mind has been put forward by Frith & Frith (1999). They have argued that this circuitry comprises superior temporal sulcus, inferior frontal regions, and medial prefrontal cortex. In line with this, a number of neuroimaging studies of mental state processing have observed activity in medial prefrontal cortex and the region of the temporo-parietal junction (Fletcher, Happe, Frith et al., 1995; Gallagher, Happé, Brunswick et al., 2000; Goel, Grafman, Sadato & Hallett, 1995; Castelli, Happé, Frith & Frith, 2000).

Potentially, the study of individuals with autism and Asperger's syndrome should aid in the identification of the neural substrate for Theory of Mind. Individuals with autism and Asperger's syndrome consistently fail Theory of Mind tasks (for reviews, see Baron-Cohen, 1995; Happé & Frith, 1996). This would suggest that any brain abnormality consistently observed in autistic individuals might be implicated in Theory of Mind. One of a number of brain regions in which there are consistent reports of abnormality in individuals with autism is the amygdala (see Baron-Cohen et al., 2000). Thus, autopsies of autistic individuals point to an abnormal increase in the packing density of grey matter in the amygdala (for brief review, see Courchesne, 1997). In addition, a structural MRI study revealed increased volume in the left amygdala and surrounding temporal areas in an Asperger's syndrome group (Abell, Krams, Ashburner et al., 1999). Moreover, in a recent proton MR spectroscopy study, Otsuka, Harada, Mori, Hisaoka & Nishitani (1999) found reduced N-acetyl aspartate concentrations in the amygdala and hippocampal regions of a group of autistic children. They suggest that this may reflect the presence of neuronal dysfunction or immature neurons.

Thus the amygdala, in particular, left amygdala, may be part of the neural circuitry involved in the processing of mental states (Baron-Cohen et al., 1999). Alternatively, the amygdala and/or its connections to regions such as the superior temporal sulcus and medial prefrontal cortex (see Amaral, Price, Pitkanen & Carmichael, 1992), may be critical for the development of Theory of Mind. If this is the case, then early damage to the amygdala and/or its connections should result in deficits in mental state processing.

8.4: Theory of Mind and executive functioning

The finding that Theory of Mind is relatively selectively impaired in autistic individuals has led some to suggest that Theory of Mind ability is domain-specific, with a dedicated neural system (e.g., Frith, Morton & Leslie, 1991; Leslie & Roth, 1993; Baron-Cohen, 1995; Frith & Frith, 1999). In contrast, others have argued that mental state information is processed by domain-general cognitive functions, namely executive functions (e.g., Frye, Zelazo & Palfai, 1995; Frye et al, 1996). Executive functions refer to the processes that underlie flexible goal-directed behaviour, e.g., inhibiting dominant responses, creating and maintaining goal-related behaviours, and temporally sequencing behaviour (Burgess, Alderman, Evans, Emslie & Wilson, 1998). Impairment of executive functions is associated with damage to prefrontal areas (e.g., Luria, 1966; Fuster, 1980; Duncan, 1986; Shallice, 1988). Neuropsychological, functional imaging, and animal lesion evidence suggest that different aspects of executive functions are dissociable, and mediated by distinct neural systems subserved by different regions of the prefrontal cortex (e.g., Luria, 1966; Fuster, 1980; Shallice & Burgess, 1996; Damasio, 1996; Robbins, 1996).

There are three positions regarding the relationship between Theory of Mind and executive functions. First, it has been argued that the development of executive functions allows the child's Theory of Mind to develop, or show its full potential on Theory of Mind tasks (e.g., Ozonoff, Rogers & Pennington, 1991; Russell, 1995; Russell, 1996; Ozonoff, 1997; Russell, 1997). Secondly, it has been argued that there are no specific systems for processing mental states and that performance on Theory of Mind tasks can be reduced to executive function ability. For example, Frye and colleagues (Frye et al., 1995; 1996) have suggested that Theory of Mind is merely one facet of the ability to act according to embedded rules. Embedded rules are of the form, "if x, if y, then z". They argue that many executive function tasks can be understood in terms of such rules. A third position is that the capacity to represent mental states is necessary for the development of executive functioning (Carruthers, 1996; Perner, 1998; Perner & Lang, 2000). Thus, Perner (1998) argues that planning skills require representing one's own intentions, and that other executive functions, such as inhibitory control and set shifting, require a representation of one's knowledge that the habitual act is maladaptive.

Two lines of evidence have been used to suggest that executive functions mediate Theory of Mind performance. First, recent studies have found that Theory of Mind and executive function abilities are correlated in pre-school children (Frye et al., 1995; Hughes, 1998a). Moreover, executive function performance predicts later Theory of Mind performance, but not vice versa (Hughes, 1998b). Recent research has begun to relate success and failure on particular executive function tasks to performance on Theory of Mind tests in normal children. In normally developing pre-school children, correlations have been found between tests of inhibitory control and attentional flexibility, and a test of deceit (Hughes, 1998a). Secondly, individuals with autism have been found to perform poorly on tests of executive functioning as well as tests of Theory of Mind (Ozonoff et al., 1991; Hughes, Russell & Robbins, 1994). Ozonoff et al. (1991) found a correlation between performance on executive function and Theory of Mind tasks in individuals with autism but not control subjects. Thus, it has been suggested that the difficulty that autistic individuals have on Theory of Mind tests is at least in part attributable to their lack of executive control (e.g., Russell, 1995; 1996; 1997). Consistent with this, children with autism appear to have particular difficulty with inhibitory control and attentional flexibility (e.g., Hughes & Russell, 1993; Hughes et al., 1994; Ozonoff, 1997). These are the two components of executive functioning that have been shown to predict Theory of Mind performance in normal children (Hughes, 1998a).

While the above data are interesting, such studies are not suitable for distinguishing between the different accounts of the developmental interaction between Theory of Mind and executive functioning for two main reasons. First, in the way that most executive function tasks assess the functioning of more than one executive function, Theory of Mind tasks may not be "pure" tests of Theory of Mind but also involve an executive function component (e.g., Leslie & Thaiss, 1992). Thus, it is to be expected that there will be correlations, or at least a lack of dissociation, between tests of executive function and Theory of Mind performance in populations who do not perform at ceiling on executive function tests, such as individuals with autism and pre-school children. However it should be noted that Perner & Lang (2000), in their review of the literature, consider that the association between Theory of Mind and executive function performance is found, even when Theory of Mind explanation tasks that putatively have a low executive function component are used. Second, it may be that the regions of the brain that mediate Theory of Mind and executive functions are anatomically proximal. If this were the case, even if they are cognitively separable processes, we would still expect to see the observed association of impairment in individuals with autism, at least at the group level. Indeed, given the importance of prefrontal circuits in executive functions e.g., (Shallice & Burgess, 1996), and the proposed role of medial frontal areas in Theory of Mind processing (e.g., Fletcher et al., 1995), this account of the data is not implausible.

This chapter reports the forensic patient, BM, first described in Chapter 4, who had a congenital or early lesion of the left amygdala. The first aim was to investigate to what extent BM showed impairment in Theory of Mind. The second aim was to determine the degree to which any Theory of Mind impairment was independent of executive functioning.

8.5: Case Report

A case report of BM has been provided previously in Chapter 4 (section 4.4).

8.6: Experimental Investigation

The following experimental investigation was carried out over a 20 month period. Substantial assessments of BM's mental state processing and executive functioning were conducted. The first aim of this investigation was to determine whether, given the suggestions of a role for the amygdala in Theory of Mind (e.g., Baron-Cohen et al., 1999; Baron-Cohen et al., 2000), BM had an impairment in mentalising.

8.7: Control subjects

BM's performance was compared with those of thirteen healthy males, matched for educational level, with mean age 30 years (s.d. = 4) and mean WAIS-R subtests scores of 10.7 (s.d. = 1.3). While not every control subject performed every task, five subjects performed at least two of the Theory of Mind tasks and seven executive functions tasks, and eight subjects performed at least three Theory of Mind tasks and two executive functions tasks. On seven of the 16 executive functions tasks, standardised data were used as the comparison. All control subjects gave informed consent.

8.8: Theory of Mind Assessment:

Ten Theory of Mind tasks were administered to BM. There were five tests assessing understanding of false belief, two tests assessing understanding of the mental states implied in cartoons, and three tests assessing understanding of intended meaning in non-literal utterances. The control subjects were given these tasks also, with the exception of the False Belief tasks since these are passed by normally developing children of 4-8 years.

8.8.1: Tasks 1-5: False Belief Tests

In False Belief tests, the participant must predict a story character's action on the basis of the character's mistaken belief about the situation. These tests can either be first order ("Anne thinks that ... ") or second order ("Mary thinks that John thinks that ... "). Two first order tests and three second order tests were given to BM. Both the first order tests ["Smarties"; (Perner, Frith, Leslie & Leekam, 1989), and "Sally-Anne"; (Baron-Cohen, Leslie & Frith, 1985)] and the second order tests ["Chocolates"; (Roth & Leslie, 1991), "Ice Cream Van"; (Perner & Wimmer, 1985), and "Coat Shopping"; (Bowler, 1992)] include control questions that assess story comprehension and memory for what happened in the story. BM's performance on these tasks was exceedingly poor (2/5). He passed the two first order False Belief tests, but failed the three False Belief tests that required second order mental state representation (see table 8.1). In contrast, BM answered all of the control questions correctly. His failure on the majority of the tasks is striking given that the tests are usually passed by normally developing children between the ages of four and eight years.

8.8.2: Tasks 6-7: Joke Comprehension tests

BM was given 20 cartoons (Joke Comprehension Test Set 1; Corcoran et al., 1997). There were ten 'mental state' cartoons, and ten 'physical state' cartoons. To understand the 'mental state' cartoons required an appreciation of the mental states of the characters. A score of one is given for each cartoon that is appropriately explained using a mental state term. The physical state cartoons could be understood without reference to mental states using physical and semantic analysis. A score of one is given for each cartoon that is appropriately explained by reference to the physical situation. As shown in table 8.1 BM was at floor for the mental state cartoons (1/10), but in the normal range for the physical state cartoons (9/10). This test was extended and replicated with a second set of mental state and physical state cartoons (Joke Comprehension Test Set 2). Again, BM's performance on the mental state cartoons was below the normal range (6/21), but in the normal range for the physical state cartoons (17/22).

	BM	Controls		
		Mean	s.d.	Range
* False Belief TOM tests (Tasks 1-5)	2/5			
Joke Comprehension Test Set 1 (Task 6)				
Mental state jokes	1/10	5.8	1.7	4-8
Physical state jokes	9/10	8.3	1.9	6-10
Joke Comprehension Test Set 2 (Task 7)				
Mental state jokes	6/21	10.7	3.1	7-15
Physical state jokes	17/22	18.6	4.7	6-23
Advanced TOM Test Set 1 (Task 8)				
Test question score	17/24	22.6	1.7	19-24
Correct mental state use	13/17	17	3.2	15-24
Control physical story comprehension	6/8	6.5	0.9	5-8
Advanced TOM Test Set 2 (Task 9)				
Test question score	20/24	23.3	0.8	22-24
Correct mental state use	16/20	21.0	1.7	18-24
Non-literal speech comprehension (Task 10)				
Sarcasm	5/24	22.3	2.1	18-24
Metaphor	23/24	23.8	0.4	23-24
* These tasks are passed by normally developing children from ages 4-8 years.				

Table 8.1: Performance of BM and control subjects on Theory of Mind (TOM) tasks

8.8.3: Tasks 8-9: Advanced Theory of Mind Test

The Advanced Theory of Mind Test Set 1 (Happé, 1994) assesses the ability to use mental state understanding to make sense of non-literal utterances (for example, see Appendix). There are 24 mental state stories and 8 physical state control stories. In each of the 24 mental state stories, a protagonist says something that isn't literally true for a variety of different motivations, e.g., tact or sarcasm. The subject must offer an explanation of why the protagonist said what s/he did.

Three scores are generated from the subject's performance on the mental state stories. The first, termed Total Score, indicates the subject's ability to comprehend the situation. The other two scores refer to the justifications the subject uses when interpreting the behaviour of the story characters, in particular, whether the subject refers to the character's mental states of physical information. An example justification involving mental states for the example story is "Because Jim knows that Simon always lies and so he should look in the other locations". An example justification involving physical information for the same story is, "Because it will be in the opposite place to wherever Simon says". As shown in table 8.1, BM's performance was below the range of the comparison group for both Total Score (17/24) and number of Mental State justifications (13/17). For the physical state control stories, only a Total Score, indicating comprehension of the situation, is given to subjects' responses. BM was in the normal range for the control physical state stories (6/8).

The experiment was replicated and extended with a different set of mental state stories that were structurally identical to Test Set 1 but with superficial details changed (Advanced Theory of Mind Test Set 2). Again BM was below the normal range of the comparison group for both Total Score (20/24) and number of mental state justifications (16/20).

8.8.4: Task 10: Non-literal speech comprehension

The comprehension of sarcasm requires mental state understanding. In sarcasm, the thoughts of the speaker must be taken into account in order to reject the incorrect literal interpretation. For example, the listener can only reject the literal interpretation of: "You're looking smart tonight, Frank" if the hearer knows that the speaker thinks that Frank looks scruffy. However, if the listener does not take the speaker's thoughts into account, the literal meaning of the utterance will not be rejected. Individuals with autism have been shown to find sarcasm particularly difficult to understand (Happé, 1993). Metaphor comprehension was also assessed. Metaphor, like sarcasm, involves understanding that the literal meaning is not the intended one, and abstracting implicit meaning.

BM was given 24 stories involving a conversation in which both sarcasm and metaphor were used. After each sarcastic and metaphorical utterance, BM was asked, "What did so-and-so mean by this?" (for an example, see Appendix). BM was markedly impaired on comprehension of sarcasm (5/24). For all incorrect answers, BM gave the literal meaning as the intended one. In contrast, BM was normal on the metaphor task (23/24), demonstrating an intact ability to understand non-literal language and to abstract implicit meanings from utterances (see table 8.1). BM may have performed normally on the metaphor task because unlike sarcasm comprehension, in understanding metaphor it is not necessary to take into account the thoughts of the speaker in order to reject the nonsensical literal meaning. The metaphor itself can suggest the intended meaning is, e.g., "You're a little computer" implies skill at maths. Individuals with autism have been found to show impairments in metaphor comprehension (Happé, 1993). However, this may reflect their difficulties in rejecting literal meanings rather than their difficulties in the representation of mental states.

8.9: Comment

The above results clearly indicate that BM has a significant Theory of Mind impairment (see table 8.1). However, his performance on all of the control tasks was normal. Thus, his Theory of Mind impairment cannot easily be accounted for in terms of difficulty in comprehension, abstraction, or memory since the control tasks also required these abilities. Moreover, since many of the Theory of Mind tests involved the use of stories, it is worth noting BM's good performance on the WAIS-R comprehension subtest and the NART.

In the literature, there have been frequent claims that Theory of Mind is mediated by general executive functioning (e.g., Frye et al., 1996; Russell, 1997). It was therefore of interest to determine whether BM's impairment in mentalising could be accounted for in terms of a deficit in executive functioning.

8.10: Executive Functions Assessment

These tests were grouped into three categories: 'Inhibition' (the ability to suppress a habitual response); 'Intentionality' (the creation and maintenance of goal-related behaviours); and 'Executive Memory' (temporal sequencing). This grouping was based on the results of a factor analysis in which these categories emerged as the three cognitive components to executive function (Burgess et al., 1998). Where possible, tests were grouped according to how strongly they loaded onto the three factors in the factor analysis. Tests that were not used in the factor analysis study were grouped according to their similarity to tests that were used. It should be noted that many of the tasks have been conceptualised in a variety of ways (i.e., trail-making has been conceptualised as reflecting set-shifting task in addition to inhibition). Indeed, many of the tests are likely to index multiple executive functions. However, in the absence of detailed information regarding the functions that each of the tasks index, an approach validated empirically was chosen (Burgess et al., 1998).

8.10.1: Tasks 11-16: Inhibition Tests

BM was given six standardised executive function tests of inhibition. Although the superficial features of the tasks are very different, each is thought to require the participant to inhibit a prepotent response. The tests were: Trail-making Part B (Army individual test battery, 1944); Stroop (Stroop, 1935); Hayling Sentence Completion (Burgess & Shallice, 1996a) Verbal Fluency (Miller, 1984); Cognitive Estimates (Shallice & Evans, 1978); and Temporal Judgements (Wilson, Alderman, Burgess, Emslie & Evans, 1996). BM performed in the normal range or above on all six tests of Inhibition (see table 8.2).

	BM	Controls		
		Mean	s.d.	Range
Inhibition tests (Tasks 11-16)				
* Trail-making Part B (secs to complete)	>75%ile			
* Stroop	100%ile			
Hayling Sentence Completion (scaled score)	19	16.2	2.5	12-18
Verbal Fluency	36	48.6	25.8	16-81
Cognitive Estimates (errors)	0	4.8	4.2	0-11
* Temporal Judgements	3/4	2.2	0.9	
Intentionality tests (Tasks 17-21)				
Modified Six Elements Task	4/4	3.6	0.5	3-4
* Zoo Map	3/4	2.4	2.0	
* Key Search	4/4	2.6	1.3	
* Action Program	4/4	3.8	0.5	
Tower of London (score system 1):	25	26.6	4.0	21-31
Executive Memory tests (Tasks 22-26)				
Rule Shift	3/4	3.4	0.9	2-4
Modified Wisconsin Card Sort Task:				
Shifts	7	4.6	2.5	1-7
Perseverative errors	0	2.8	4.8	0-11
Intra-dimensional/Extra-dimensional Shift:				
Intra-dimensional errors	0	0.1	2.0	0-3
Extra-dimensional errors	3	22.1	22.1	1-58
Reversal errors	0	2.3	1.8	0.5
* Brixton Spatial Anticipation (errors)	14	16.0	5.7	
Non-spatial Conditional Learning (errors)	16	21.0	12.0	7-34

* Performance compared with published data

Table 8.2: Performance of BM and control subjects on executive functions tasks

8.10.2: Tasks 17-21: Intentionality Tests

It has been argued that 'Intentionality', the ability to create and monitor goal-related behaviour, is a necessary precursor to self-awareness and the development of concepts of mental states (Russell, 1996). An impairment in this executive capacity might therefore be predicted in BM. 'Intentionality' tests require the subject to create and maintain a plan in order to achieve a goal, in the absence of any external stimuli cueing the appropriate responses. Such tasks also involve embedded rule use, which Frye et al. (1995) have argued encompasses Theory of Mind. BM was given five standardised tests assessing this ability: Modified Six Elements Task; Zoo Map; Key Search; Action Program (Wilson et al., 1996); and Tower of London (Shallice, 1982). BM's performed in the normal range or above on all five tests of intentionality (see table 8.2).

8.10.3: Tasks 22-26: Executive Memory

'Executive Memory' tests require the participant either to shift attention away from a given cue, transfer attention to another cue, or both. As noted previously, individuals with autism have been shown to be impaired on such tasks (Hughes & Russell, 1993) (Hughes et al., 1994). BM was given five tests that reflect Executive Memory processes. In all tests but the last, the participant must shift set from a dominant response according to an arbitrary rule. The tests were: Rule Shift (Wilson et al., 1996); Modified Wisconsin Card Sort Task (Nelson, 1976); Intra-dimensional/Extra-dimensional Shift (Hughes et al., 1994); Brixton Spatial Anticipation (Burgess & Shallice, 1996b); and Non-spatial Conditional Learning (Petrides, 1990). BM's performed in the normal range or above on all five tests of executive memory (see table 8.2).

8.11: Comment

BM clearly showed normal performance on all aspects of executive functioning. The 16 tests he was given included those that have been frequently associated with poor performance on Theory of Mind tasks, that is, those involving inhibition, embedded rule use, and the execution of an arbitrary response in competition with a dominant response. Thus BM's poor Theory of Mind performance cannot be accounted for in terms of executive dysfunction. Moreover, it is interesting to note that one of the control individuals presented with impaired performance on the Hayling Sentence

Completion, Verbal Fluency, Cognitive Estimates, Rule Shift, and the Modified Wisconsin Card Sort tasks. However, this subject showed no impairment on any of the Theory of Mind tasks.

8.12: Discussion

BM, who had a unilateral left amygdala lesion of longstanding or congenital origin, showed profound difficulty in representing the mental states of others. BM performed consistently poorly on ten mental state processing tasks, assessing false belief understanding (Tasks 1-5), mental state understanding in the comprehension of cartoons (Tasks 6 & 7), and understanding of intended meaning in non-literal utterances (Tasks 8-10). The degree to which BM's Theory of Mind impairment was independent of executive functioning was investigated. BM was given 16 executive function tests assessing his ability to inhibit dominant responses, create and maintain goal-related behaviours, and temporally sequence behaviour (Tasks 11-26). The battery included executive function tests that previous research has associated with Theory of Mind development. BM performed in the normal range or above on all the executive function tests. These findings show that the neurocognitive system mediating Theory of Mind is developmentally separable from the neurocognitive systems mediating executive functions, and that executive functions can develop and function on-line, independently of Theory of Mind.

8.12.1: Implications for the anatomy of Theory of Mind

There have been recent claims that the amygdala may be involved in the mediation of Theory of Mind (Baron-Cohen, 1995; Baron-Cohen et al., 1999; Baron-Cohen et al., 2000). However, there have been no investigations of Theory of Mind performance in individuals with amygdala lesions. BM presented with a lesion in the basal nuclei of the left amygdala that was consistent with a dysembryonablastic neuroepithelial tumour of longstanding or congenital origin. In line with suggestions that the amygdala may be one brain region involved in the mediation or development of Theory of Mind, BM presented with profound impairment in mentalizing ability.

Several hypotheses can be developed concerning the role of the amygdala in Theory of Mind functioning. These could be tested in future neuropsychological case studies. First, as suggested by Baron-Cohen and colleagues (Baron-Cohen, 1995;

Baron-Cohen et al., 1999; Baron-Cohen et al., 2000), the amygdala may mediate Theory of Mind functioning. In line with this position, Baron-Cohen et al. (1999), using fMRI, reported left amygdala activation during a task requiring the subject to infer the mental state of an individual from the expression of their eyes. Moreover, individuals with autism and Asperger's syndrome who show profound Theory of Mind impairment present with structural abnormalities involving the amygdala, particularly the left (Abell et al., 1999; Courchesne, 1997; Otsuka et al., 1999). In addition, individuals with paranoid delusional schizophrenia who also show Theory of Mind impairment also present with structural abnormalities involving the amygdala (see, for a review, Lawrie & Abukmeil, 1998). This position predicts that other patients with amygdala lesions, whether these occur early in development or in adulthood, should present with Theory of Mind impairment. Secondly, appropriate amygdala functioning may be a prerequisite for the development of Theory of Mind even if it is not, in itself, involved in mediating the representation of mental states. The amygdala certainly has extensive interconnections with regions of medial prefrontal cortex and the superior temporal sulcus (e.g., Amaral et al., 1992). Both these areas have been implicated in the circuitry that mediates Theory of Mind (Fletcher et al., 1995; Goel et al., 1995; Gallagher et al., 2000). This position predicts that patients whose amygdala lesions were acquired very early in life should show impairment in Theory of Mind but patients whose lesions were acquired in adulthood should not. Thirdly, it is possible that BM's amygdala lesion plays no role in his Theory of Mind impairment. For example, BM's impairment could be due to undetected damage elsewhere in the system.

8.12.2: Implications for the relationship between Theory of Mind and executive functioning

There has been considerable debate concerning the association between Theory of Mind and executive functioning. Indeed, many have argued that there is no specific neuro-cognitive system which mediates Theory of Mind but rather that performance on Theory of Mind tasks is mediated, at least in part, by executive functioning (e.g., Ozonoff et al., 1991; Frye et al., 1995; Russell, 1995; 1996; Ozonoff, 1997). There are two main forms of this argument. First, it has been argued that the development of executive functions allows the child's Theory of Mind to develop, or show its full potential (e.g., Ozonoff et al., 1991; Russell, 1995; 1996; 1997; Ozonoff, 1997).

Secondly, it has been argued that there are no specific systems for processing mental states and that performance on Theory of Mind tasks can be reduced to executive function ability (e.g., Frye et al., 1996).

BM presented with a profound impairment in Theory of Mind in the complete absence of any impairment in executive functioning. These data clearly indicate, contrary to some suggestions, that performance on Theory of Mind tasks cannot be reduced to executive function ability. Indeed, it is important to note that BM showed no impairment on the executive function tasks that are typically found to be impaired in individuals with autism, for example Wisconsin Card Sorting Task and the Tower of London (see Pennington & Ozonoff, 1996). Moreover, it is important to note that BM showed no impairment on the executive function tasks that neuroimaging and lesion work has indicated recruit areas of medial frontal cortex. These tasks include the Stroop task and Conditional learning (e.g., Bench, Frith, Grasby, et al., 1993; Carter, Mintun, Nichols & Cohen, 1997; Petrides, 1990). Very proximal areas of medial frontal cortex have been shown to be recruited during Theory of Mind processing (Fletcher et al., 1995; Goel et al., 1995; Gallagher et al., 2000). Thus, it is clear that an impairment, other than in executive functioning caused BM's impairment on the Theory of Mind tasks. Given the considerable variety of tasks used, addressing different modalities and with a range of task demands, the most parsimonious explanation is to assume that his impairment was due to an impairment in the ability to represent mental states.

While BM's difficulty on Theory of Mind tasks was clearly due to a specific problem with the representation of mental states, this may not be typically the case in autism, where executive dysfunction has been widely reported (Ozonoff et al., 1991). Indeed, it could still be argued that executive functions are necessary (if not sufficient) for successful performance on Theory of Mind tasks (cf., Ozonoff, 1997; Russell, 1995). This position has to predict that individuals with executive function impairment should show failure on Theory of Mind tasks. However, with regard to the executive functions of inhibitory control and attentional set-shifting, this prediction does not appear to hold. Recently, a patient, JS, was reported with 'acquired sociopathy' following frontal lobe damage. JS failed two of four tests of inhibitory control and one of two tests of attentional set-shifting. However, he performed normally on the Advanced Theory of Mind test (Blair & Cipolotti, 2000). Similarly, a second orbitofrontal cortex patient with executive dysfunction has recently been found to be unimpaired in Theory of Mind (Bach, Happé, Fleminger & Powell, 2000). The frequent occurrence of executive dysfunction in autistic and Asperger's syndrome individuals may be because many patients with these disorders have suffered damage to many neuro-cognitive systems that rely on prefrontal cortex. Similarly, the findings of correlations in normal developing children between Theory of Mind and executive functioning (e.g., Hughes, 1998a; Hughes, 1998b), may either reflect proximal systems or similar developmental time courses between Theory of Mind and specific executive functions.

An alternative position on the relationship between Theory of Mind and executive functioning has been developed by Carruthers (1996) and, more formally, by Perner and colleagues (Perner, 1998; Perner, Stummer & Lang, 1999; Perner & Lang, 2000). These authors suggest that the capacity to represent mental states is necessary in order to develop executive functions. Indeed, Perner has argued that,

> *"Since executive functions are characterised by formulation of higher-order intentions and representations, they need the conceptual repertoire for expressing these higher-order states, i.e., a Theory of Mind. So one would expect people with a deficient Theory of Mind to have executive function problems."* Perner, 1998: p277-278).

More specifically, Perner argues that meta-representational abilities are essential in order to overcome dominant responses or old strategies, as in tests of inhibition and attentional set-shifting (Perner, 1998; Perner & Lang, 2000). BM passed only two out of five simple false belief tests. From this, Perner's position would predict impairment in the inhibitory and attentional set-shifting components of executive functions. However BM's performance on all executive functions was normal. This suggests that executive functions do not require the same representational abilities as those involved in mental state processing.

The performance of BM thus clearly supports the position that Theory of Mind ability is domain-specific, with a dedicated neural system (e.g., Frith et al., 1991; Leslie & Roth, 1993; Baron-Cohen, 1995; Frith & Frith, 1999). BM presented with a very severe impairment in Theory of Mind but no impairment in executive functioning.

8.13: The amygdala, Theory of Mind, and developmental psychopathy

It has been suggested here that BM's Theory of Mind impairment may have arisen from his early amygdala function. In previous chapters, it has been suggested that developmental psychopathy is associated with early amygdala dysfunction. In line with this suggestion, BM was found to show three important similarities with developmental psychopaths: an instrumental learning impairment; a selective impairment in emotion expression recognition; and a lack of empathy and other moral emotions. It might then be expected that psychopathic individuals will also be impaired in Theory of Mind. However, psychopaths show normal Theory of Mind performance (Blair, Sellars, Strickland et al, 1996). It may be that the amygdala dysfunction in developmental psychopathy is less severe or more selective than that seen in BM. Certainly, in contrast with BM, there is no known evidence of gross structural abnormalities of the amygdala in developmental psychopathy. Alternatively, the critical difference between BM and typical developmental psychopaths may lie in the connections between the amygdala and other areas concerned with mental state processing. Current imaging techniques cannot detect abnormalities in the connections between regions. It may then be that BM also has abnormal connections with regions important for the processing of mental states, whereas in developmental psychopathy these connections are intact. It is also possible that BM's amygdala damage was not causal in his Theory of Mind impairment.

8.14: Summary

There have been recent suggestions that the amygdala may be involved in the development or mediation of Theory of mind (e.g., Baron-Cohen et al., 1999). This chapter reported a series of experimental investigations to determine BM's cognitive functioning. In line with his diagnoses of Asperger's syndrome and schizophrenia, BM was found to be severely impaired in his ability to represent mental states. Following this, a second series of studies was conducted to determine BM's executive functioning. In the literature, there have been frequent claims that Theory of Mind is mediated by general executive functioning (e.g., Frye et al., 1996;

Ozonoff, 1997). BM showed no indication of executive function impairment, passing 16 tests assessing his ability to inhibit dominant responses, create and maintain goal-related behaviours, and temporally sequence behaviour. The findings are discussed with reference to models regarding the role of the amygdala in the development of Theory of Mind and the degree of dissociation between Theory of Mind and executive functioning. It was concluded that Theory of Mind is not simply a function of more general executive functions, and that executive functions can develop and function on-line, independently of theory of mind. Moreover, it was concluded that the amygdala may play some role in the development of the circuitry mediating Theory of Mind.

Chapter 9

Summary, Conclusions, and Future Directions

9.1: Introduction

The aim of this chapter is to review the conclusions that can be drawn from the empirical and theoretical work presented in this thesis. These conclusions concern the cognitive processes underlying the guidance of behaviour by expectations and behavioural change following expectation violations. The differential roles of the amygdala and orbitofrontal cortex in these processes are also discussed, and the impact of damage to these brain regions on emotional behaviour. Later sections of this chapter will detail potential future directions for this research. This comprises the development and further testing of the Valence Change Reset model presented in Chapter 3, and further explorations of the role of the amygdala and orbitofrontal cortex in social and emotional cognition.

9.2: Summary and Conclusions

This thesis began with a discussion of the representation of response outcome expectations. A number of researchers have proposed models of the physiological and behavioural effects of violations of these expectations. These models were discussed in Chapter 1. Physiologically, expectation violations have been proposed to trigger orienting responses and autonomic arousal (e.g., Grossberg, 1982; Mandler, 1984; Amsel, 1992). Behaviourally, expectation violations have been proposed to result in behavioural inhibition (e.g., Gray, 1982; Mandler, 1984) or rapid behavioural change (e.g., Grossberg, 1982; Rolls, 1990). Chapter 2 thus tested the predictions that reinforcement expectation violations would trigger arousal and rapid behavioural change. The results of the first three experiments were not predicted by any existing model. Subjects were clearly using information about the magnitude of expected reinforcement to guide their instrumental responses. However, unexpected changes in magnitude did not trigger autonomic arousal (indexed by skin conductance response), and nor did they trigger rapid behavioural change. In contrast, changes in the expected valence of reinforcement (either reward or punishment) did produce arousal and trigger rapid behavioural change. These findings suggested that existence of two different representational systems: one system that represents both magnitude and valence and guides instrumental responding; and a second system that represents only valence, and triggers rapid behavioural change. That the two systems are separable was suggested from the evidence that they represented reinforcement expectations differently. A fourth experiment suggested the possible existence of a third representational system. This system appeared to represent the motivational significance of stimuli, that is, whether they should be approached or avoided. Violations of these expectations were also associated with arousal responses and rapid behavioural change.

In Chapter 3, hypotheses generated from the experiments in Chapter 2 were used to develop a computational model, the Valence Change Reset model. This model was based on computational models of classical conditioning developed by Rescorla & Wagner (1972), Mackintosh (1975) and Pearce & Hall (1980). However, the Valence Change Reset model was unique in that it implemented the hypothesis that an instrumental learning system that represents magnitude and valence interacts with an instrumental re-learning system that only represents valence (see figure 3.1). When the instrumental re-learning system was activated by a mismatch between expected and actual reinforcement, it reset response outcome expectations in the instrumental learning system. The Valence Change Reset model successfully simulated the human behavioural data from the first three experiments in Chapter 2. This demonstrated that the two hypothesised systems could account for the behavioural and psychophysiological findings of Chapter 2.

The Valence Change Reset model predicted that the two systems – the Instrumental Learning system and the Instrumental Re-learning system – could be independently damaged. Chapter 4 outlined evidence that the (basolateral) amygdala is involved in instrumental learning. In line with this, an instrumental learning impairment was demonstrated in BM, a patient with early left amygdala damage. Chapter 5 reviewed the evidence that the orbitofrontal cortex is involved in instrumental re-learning. It was predicted that orbitofrontal cortex damage would allow intact instrumental learning, but impair instrumental re-learning. This prediction was supported by the performance of two patients with orbitofrontal cortex damage. Their pattern of performance on the task – intact instrumental learning together with impaired instrumental re-learning – supported the hypothesis that instrumental learning and re-learning involve separable mechanisms.

Following these findings from the amygdala and orbitofrontal cortex patients, it was decided to investigate instrumental learning and re-learning in a group of psychopathic individuals. Both amygdala and orbitofrontal cortex dysfunction have been associated with psychopathy (e.g., LaPierre, Braun & Hodgins, 1995; Anderson, Bechara, Damasio, Tranel & Damasio, 1999; Blair & Frith, 2000). Chapter 6 tested the predictions that developmental psychopaths would be impaired in instrumental learning and/or re-learning. Both predictions were supported.

The hypothesis that developmental psychopathy is associated with early amygdala dysfunction was explored further in Chapter 7. In this chapter, BM's emotion expression recognition and empathy were assessed. Previous research has found that psychopathic individuals are impaired in fear recognition, and that children with psychopathic tendencies are impaired in fear and sadness recognition (Blair, Colledge, Mitchell & Murray, 2000; Mitchell, Colledge & Blair, in prep.). It was therefore predicted that BM, who suffered from early amygdala damage, would be impaired in the recognition of fear and sadness. This prediction was supported. Chapter 7 then discussed Blair's model of moral development, in which representations of situations that cause distress come to predict distress in others, through a process of classical conditioning (Blair, 1995). This process allows the development of the moral emotions such as empathy, guilt and remorse. If this moral development process is mediated at least in part by the amygdala, then this process should be disrupted by early amygdala dysfunction. It was therefore predicted that, like psychopathic individuals, BM would show an absence of the moral emotions. An assessment of BM's behaviour using the Psychopathy Checklist-Revised (Hare, 1991) revealed that, as predicted, BM had a deficiency in empathy, guilt and remorse.

Chapter 8 investigated suggestions that early amygdala damage has a detrimental effect on another aspect of social cognition, that is, Theory of Mind (e.g., Baron-Cohen, 1995; Brothers, 1997; Baron-Cohen, Wheelwright, Bullmore, et al., 1999; Baron-Cohen, Ring, Bullmore et al., 2000). This chapter also investigated the role of executive functions in theory of mind, as it has been suggested that Theory of Mind is mediated by these domain-general cognitive abilities (e.g., Frye, Zelazo & Palfai, 1995; Frye, Zelazo, Brooks & Samuels, 1996). BM showed a severe impairment in Theory of Mind ability but his executive function skills were intact. It was therefore concluded that Theory of Mind is not simply a function of more general executive functions, and that executive functions can develop and function on-line, independently of theory of mind. The findings also supported suggestions that the amygdala plays a role in the development of the circuitry mediating Theory of Mind (e.g., Baron-Cohen, 1995). Interestingly, despite the other behavioural similarities of BM with psychopathic individuals, the latter do not seem to be impaired in Theory of Mind (Blair, Sellars, Strickland, et al, 1996). Thus the relationship between instrumental learning, emotion expression recognition, and Theory of Mind, and the role of the amygdala in these processes, remains to be explored.

9.3: Future directions

The empirical work in this thesis has highlighted a number of research questions in need of investigation.

9.3.1: Further tests of the Valence Change Reset model

The different patterns of performance on an instrumental learning and re-learning task observed in BM and two patients with orbitofrontal damage, CM and DJ, supported the hypothesis that instrumental re-learning involves a mechanism separable to that involved in instrumental learning. Further neuropsychological testing is necessary to confirm or refute the presence of this dissociation. In addition, two crucial predictions of the Valence Change Reset model are generated from the hypothesis that only valence changes are processed in the Instrumental Re-learning system. The first prediction is that individuals who are intact in instrumental learning but impaired in instrumental re-learning (such as CM and DJ) should not differ to controls in their slow behavioural responses to magnitude changes. This is because the Instrumental Learning system, which processes magnitude changes, is hypothesised to be intact in such patients. A second prediction is that very small changes in valence should trigger autonomic arousal increases and behavioural change. In contrast, even very large magnitude changes should not have this effect.

It must be noted that removing the Instrumental Re-learning system in the Valence Change Reset model did not simulate the performance of the orbitofrontal cortex patients in the re-learning phases of the task. It was suggested that this might be because the Valence Change Reset model did not model the development of predominant responses. Certainly, it is currently suggested that one role of the orbitofrontal cortex is in the inhibition of incorrect dominant responses (e.g., Elliott, Dolan & Frith, 2000; Shimamura, 2000). It has also been suggested that patients with orbitofrontal cortex damage are impaired in their ability to learn new stimulus-reinforcement associations (e.g., Rolls, 1990; 1996; 2000). One difficulty in distinguishing between these two accounts is that previous research investigating emotional re-learning in orbitofrontal cortex patients has always confounded changes in the reinforcement value of a stimulus with the need to inhibit a previously rewarded response to it. It is therefore currently unknown whether patients' perseverative behaviour is due to a deficit in stimulus-reinforcement re-learning, and/or response inhibition. Future work could test contrasting predictions from the response inhibition and stimulus-reinforcement re-learning accounts. In particular, the response inhibition hypothesis predicts that patients will be impaired in emotional re-learning when they have to inhibit a previously rewarded response, even when they do not have to learn a new stimulus reinforcement value. In contrast, the stimulus-reinforcement re-learning hypothesis predicts that patients with OFC damage will be impaired in emotional re-learning when stimulus reinforcement values change, even when there is no previously rewarded response to inhibit. The findings from this proposed study would result in a greater understanding of the specific cognitive impairment(s) underlying these patients' instrumental re-learning impairment. It would then be possible to modify the Valence Change Reset model on the basis of these findings.

9.3.2: The amygdala, developmental psychopathy, and instrumental learning

Both BM and a population of psychopathic individuals were shown to have a severe impairment in instrumental learning. However, it was not possible to determine from the task used whether both appetitive and aversive instrumental conditioning were impaired. Animal work indicates that the (basolateral) amygdala is involved in both appetitive and aversive instrumental conditioning (e.g., Cador, Robbins & Everitt, 1989; Everitt, Cador & Robbins, 1989; Everitt & Robbins, 1992; Burns, Robbins &

Everitt, 1993; Killcross, Robbins & Everitt, 1997). Further work needs to explore appetitive and aversive instrumental learning in both amygdala patients and psychopathic individuals, in order to delineate the similarities and differences between the two groups, and to ascertain more precisely the cognitive impairment underlying their instrumental learning deficit.

9.3.3: The role of the amygdala in social cognition

Chapter 8 demonstrated for the first time a Theory of Mind impairment in a patient with early left amygdala damage. However, it cannot be known whether BM's amygdala damage was in fact the causal factor in his Theory of Mind deficit. Thus, this finding needs to be followed-up by investigating whether Theory of Mind impairments are seen in other patients with amygdala damage. In particular, it should be investigated whether Theory of Mind impairment is seen only following early amygdala damage, or also following damage in adulthood. Such data are essential if we are to know whether the amygdala is necessary for the development of Theory of Mind, the on-line processing of mental state information, or both.

9.3.4: The role of the amygdala in emotion expression processing

Chapter 7 demonstrated an impairment in the recognition of fearful and sad faces in BM. This finding was consistent with previous neuropsychological data and functional imaging data (Morris, Frith, Perrett, et al., 1996; Blair, Morris, Frith, Perrett, Dolan, 1999). What remains unknown is what cognitive deficit underlies amygdala patients' difficulty in processing emotional expressions. One possibility is that the amygdala facilitates processing of emotional expressions in other brain regions (Whalen, 1998). In line with this, the amygdala can be activated by fearful and angry faces in the absence of awareness (Morris, Friston, Büchel et al., 1998; Whalen, Rauch, Etcoff et al., 1998), suggesting that this region may be involved in automatic monitoring of such stimuli. This hypothesis predicts a selective deficit in modulating attention to fearful and sad facial expressions.

This hypothesis could be tested by exploiting the established distracting effect of faces in attentional reaction time (RT) tasks (e.g., Young, Ellis, Flude, McWeeny & Hey, 1986; Jenkins, Lavie & Driver, 2000). An adapted version of the task developed by Jenkins et al. could be used to test the prediction that the amygdala enables preferential processing of fearful and sad faces, or fearful and angry faces. If the amygdala facilitates processing of a particular emotional expression, then amygdala patients should show significantly less slowing of reaction time compared with controls when those facial stimuli are used as distracters, compared with a condition in which neutral faces or faces showing other emotions are used as distracters.

9.4 Summary

In this thesis, behavioural and psychophysiological responses to expectation violations generated a computational model of instrumental learning and re-learning. In this model, instrumental learning was mediated by a system that represented both magnitude and valence. In contrast, instrumental re-learning was mediated by a system that only represented valence. It was predicted that these systems were separable, and this prediction was supported by the demonstration of a dissociation between instrumental learning and re-learning. An amygdala patient was severely impaired in instrumental learning, whereas two orbitofrontal cortex patients were only impaired in instrumental re-learning. A population of psychopathic individuals were found to be impaired in both instrumental learning and re-learning, supporting suggestions that the amygdala and orbitofrontal cortex are dysfunctional in this developmental disorder. An investigation of emotional and social cognition impairments in a patient with early amygdala damage was conducted. The patient was found to display an absence of empathic feeling, consistent with suggestions that this feature of developmental psychopathy may result from early amygdala dysfunction. In addition, the patient showed a severe Theory of Mind deficit, highlighting questions regarding the role of the amygdala in the development of mentalizing ability. This chapter has identified future research directions to investigate questions raised by the findings of this thesis.

Reference List

Abell F, Krams M, Ashburner J, Passingham R, Friston K, Frackowiak R et al. The neuroanatomy of autism: A voxel based whole brain analysis of structural scans in high functioning individuals. Neuroreport 1999; 10:1647-1651.

Adolphs R, Tranel D, Damasio H, Damasio AR. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature 1994; 372:669-672.

Adolphs R, Tranel D, Damasio H, Damasio AR. Fear and the human amygdala. Journal of Neuroscience 1995; 15:5879-5891.

Adolphs R, Damasio H, Tranel D, Damasio AR. Cortical systems for the recognition of emotion in facial expressions. Journal of Neuroscience 1996; 16(7678):7687.

Adolphs R, Tranel D, Damasio AR. The human amygdala in social judgement. Nature 1998; 393:470-474.

Adolphs R, Tranel D, Hamann S, Young AW, Calder AJ, Phelps EA et al. Recognition of facial emotion in nine individuals with bilateral amygdala damage. Neuropsychologia 1999; 37:1111-1117.

Adolphs R, Tranel D. Intact recognition of emotional prosody following amygdala damage. Neuropsychologia 1999; 37:1285-1292.

Amaral DG, Price JL, Pitkanen A, Carmichael ST. Anatomical organization of the primate amygdaloid complex. In: Aggleton JP, editor. The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction. New York: Wiley-Liss, Inc., 1992: 1-66.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (4th Edition) (DSM - IV-R). Washington, DC: American Psychiatric Association, 1994.

Amsel A. Frustration theory. Cambridge: Cambridge University Press, 1992.

Anderson A, Phelps E. Intact recognition of vocal expressions of fear following bilateral lesions of the human amygdala. Neuroreport 1998; 9:3607-3616.

Anderson S, Bechara A, Damasio H, Tranel D, Damasio A. Impairment of social and moral behavior related to early damage in human prefrontal cortex. Nature: Neuroscience 1999; 2(11):1032-1037.

Angrilli A, Mauri A, Palomba D, Flor H, Birhaumer N, Sartori G et al. Startle reflex and emotion modulation impairment after a right amygdala lesion. Brain 1996; 119:1991-2000.

Aniskiewica AS. Autonomic components of vicarious conditioning and psychopathy. Journal of Clinical Psychology 1979; 35:60-67.

Army individual test battery. Manual and directions for scoring. Washington (DC): War Department, Adjutant General's Office, 1944.

Arnett P, Howland E, Smith S, Newman J. Autonomic responsivity during passive avoidance in incarcerated psychopaths. Personality & Individual Differences 1993; 14(1):173-184.

Bach L, Happé F, Fleminger S, Powell J. Theory of mind: Independence of executive function and the role of the frontal cortex in acquired brain injury. Cognitive Neuropsychiatry 2000; 5(3):175-192.

Balleine B, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 1998; 37, 407-419.

Baron-Cohen S. Mindblindedness: An essay on autism and theory of mind. Cambridge, Massachusetts: MIT Press, 1995.

Baron-Cohen S, Leslie AM, Frith U. Does the autistic have a "Theory of Mind"? Cognition 1985; 21:37-46.

Baron-Cohen S, Ring H, Moriarty J, Schmits B, Costa D, Ell P. Recognition of mental state terms: Clinical findings in children with autism and a functional neuroimaging study of normal adults. British Journal of Psychiatry 1994; 165:640-649.

Baron-Cohen S, Wheelwright S, Bullmore ET, Brammer M, Simmons A, Williams SC. Social intelligence in the normal and autistic brain: an fMRI study. European Journal of Neuroscience 1999; 11(6):1891-1898.

Baron-Cohen S, Ring H, Bullmore ET, Wheelwright S, Ashwin C, Williams SCR. The amygdala theory of autism. Neuroscience and Biobehavioural Reviews 2000; 24:355-364.

Bechara A, Damasio AR, Damasio H, Anderson SW. Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 1994; 50:7-15.

Bechara A, Tranel D, Damasio H, Damasio AR. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. Science 1995; 275:1115-1118.

Bechara A, Damasio H, Tranel D, Damasio AR. Deciding advantageously before knowing the advantageous strategy. Science 1997; 275:1293-1295.

Bechara A, Damasio H, Tranel D, Anderson SW. Dissociation of working memory for decision making within the human prefrontal cortex. Journal of Neuroscience 1998; 18:428-437.

Bechara A, Damasio H, Damasio AR, Lee GP. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. Journal of Neuroscience 1999; 19:5473-5481.

Bechara A, Damasio H, Damasio AR. Emotion, decision-making and the orbitofrontal cortex. Cerebral Cortex 2000; 10:295-307.

Bench CJ, Frith CD, Grasby PM, Friston KJ, Paulesu E, Frackowiak RS et al. Investigations of the functional anatomy of attention using the Stroop test. Neuropsychologia 1993; 31:907-922.

Bishop C. Neural networks for pattern recognition. Oxford: Clarendon, 1995.

Blair RJR. A cognitive developmental approach to morality: Investigating the psychopath. Cognition 1995; 57(1):1-29.

Blair RJR, Sellars C, Strickland I, Clark F, Williams A, Smith M et al. Theory of Mind in the psychopath. Journal of Forensic Psychiatry 1996; 7:15-25.

Blair RJR, Jones L, Clark F, Smith M. The psychopath: A lack of responsiveness to distress cues? Psychophysiology 1997.

Blair RJR, Morris J, Frith CD, Perrett D, Dolan RJ. Dissociable neural responses to facial expressions of sadness and anger. Brain 1999; 122:833-893.

Blair RJR, Curran HV. Selective impairment in the recognition of anger induced by diazepam. Psychopharmacology 1999; 147:335-338.

Blair RJR, Cipolotti L. Impaired social response reversal: A case of "acquired sociopathy". Brain 2000; 123:1122-1141.

Blair RJR, Colledge E, Mitchell D, Murray L. Selective impairment in the processing of sad and fearful expressions by children with psychopathic tendencies. Journal of Abnormal Child Psychology. In press.

Blair R, Frith U. Neurocognitive explanations of Antisocial Personality Disorders. Criminal Behaviour and Mental Health. In press.

Bolles RC. Reinforcement, expectancy, and learning. Psychological Review 1972; 79(5):394-409.

Bond A, Curran HV, Bruce M, O'Sullivan G, Shine P. Behavioural aggression in panic disorder after 8 weeks' treatment with aprazolam. Journal of Affective Disorders 1995; 35:117-123.

Borrill J, Rosen B, Summerfield A. The influence of alcohol on judgement of facial expressions of emotion. British Journal of Medical Psychology 1987; 60:71-77.

Bowler DM. Theory of mind in Asperger's syndrome. MPQ 1992; 1:1-2.

Brieter H, Etcoff N, Whalen P, Kennedy W, Rauch S, Buckner R et al. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 1996; 17:875-887.

Broks P, Young AW, Maratos EJ, Coffey PJ, Calder AJ, Isaac CL et al. Face processing impairments after encephalitis: amygdala damage and recognition of fear. Neuropsychologia 1998; 36:59-70.

Brothers L. Friday's footprint: How society shapes the human mind. New York: Oxford University Press, 1997.

Büchel C, Dolan RJ, Armony JL, Friston K. Amygdala-hippocampal involvement in human aversive trace conditioning revealed through event-related functional magnetic resonance imaging. Journal of Neuroscience 1999; 19(24):10869-10876.

Büchel C, Dolan RJ. Classical fear conditioning in functional neuroimaging. Current Opinion in Neurobiology 2000; 10:219-223.

Burgess PW, Shallice T. Response suppression, initiation and strategy use following frontal lobe lesions. Neuropsychologia 1996; 34:263-272.

Burgess PW, Shallice T. Bizarre responses, rule detection and frontal lobe lesions. Cortex 1996; 32:241-260.

Burgess PW, Alderman N, Evans J, Emslie H, Wilson BA. The ecological validity of tests of executive function. Journal of the International Neuropsychological Society 1998; 4:547-558.

Burns LH, Robbins TW, Everitt B. Differential effects of excitotoxic lesions of the basolateral amygdala, ventral subiculum and medial prefrontal cortex on responding with conditioned reinforcement and locomotor activity potentiated by intra-accumbens infusions of D-amphetamine. Behavioural Brain Research 1993; 55:167-183.

Burns LH, Everitt B, Robbins TW. Effects of excitotoxic lesions of the basolateral amygdala on conditional discrimination learning with primary and conditioned reinforcement. Behavioural Brain Research 1999; 100:123-133.

Butter C. Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in *Macaca Mulatta.* Physiology and Behavior 1969; 4:163-171.

Cador M, Robbins TW, Everitt B. Involvement of the amygdala in stimulus-reward associations: interaction with the ventral striatum. Neuroscience 1989; 30:77-86.

Calder AJ, Young AW, Rowland D, Perrett D. Facial emotion recognition after bilateral amygdala damage: Differentially severe impairment of fear. Cognitive Neuropsychology 1996; 13:699-745.

Cannon WB. The James-Lange theory of emotions: A critical examination and an alternative theory. American Journal of Psychology 1927; 39:106-124.

Carpenter G, Grossberg S. The ART of adaptive pattern recognition by a self-organizing neural network. IEEE 1988; March:77-88.

Carruthers P. Autism as mindblindness: An elaboration and partial defence. In: Carruthers P, Smith PK, editors. Theories of theories of mind. Cambridge, UK: Cambridge University Press, 1996: 257-273.

Carter CS, Mintun M, Nichols T, Cohen JD. Anterior cingulate gyrus dysfunction and selective attention deficits in schizophrenia: [15O]H2O PET study during single-trial Stroop task performance. American Journal of Psychiatry 1997; 154(12):1670-1675.

Castelli F, Happé FGE, Frith U, Frith CD. Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patters. Neuroimage 2000; 12(3):314-325.

Channon S, Daum I, Gray JA. Operant conditioning after temporal lobe lesions in man: conditional and simple discrimination learning. Cortex 1993; 29:315-324.

Cleckley H. The mask of sanity: an attempt to clarify some issues about the so-called psychopathic personality. St. Louis, MO: The CV Mosby Co., 1950.

Corcoran R, Cahill C, Frith CD. The appreciation of visual jokes in people with schizophrenia: A study of 'mentalizing' ability. Schizophrenia Research 1997; 24:319-327.

Cornell D, Warren J, Hawk G, Stafford E, Oram G, Pine D. Psychopathy in instrumental and reactive violent offenders. Journal of Consulting and Clinical Psychology 1996; 64(4):783-790.

Courchesne E. Brainstem, cerebellar and limbic neuroanatomical abnormalities in autism. Current Opinion in Neurobiology 1997; 7:269-278.

Critchley HD, Elliott E, Mathias CJ, Dolan RJ. Neural activity relating to generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. Journal of Neuroscience 2000; 20(8):3033-3040.

Critchley HD, Simmons A, Daly E, Russell A, van Amelsvoot T, Robertson D et al. Prefrontal and medial temporal correlates of repetitive violence to self and others. Biological Psychiatry 2000; 47(10):928-934.

Damasio AR, Tranel D, Damasio H. Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. Behavioural Brain Research 1990; 41(2):81-94.

Damasio AR, Tranel D, Damasio H. Somatic markers and the guidance of behavior: Theory and preliminary testing. In: Levin HS, Eisenberg HM, Benton AL, editors. Frontal lobe function and dysfunction. New York: Oxford University Press, 1991: 217-229.

Damasio AR. Descartes' error: Emotion, reason and the human brain. London: Picador, 1994.

Damasio AR. The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society of London, Series B 1996; 351:1413-1420.

Daum I, Schugens MM, Channon S, Polkey CE, Gray JA. T-maze discrimination and reversal learning after unilateral temporal or frontal lobe lesions in man. Cortex 1991; 27:613-622.

Dias R, Robbins TW, Roberts AC. Dissociation in prefrontal cortex of affective and attentional shifts. Nature 1996; 380:69-72.

Dickinson A. Contemporary animal learning theory. Cambridge: Cambridge University Press, 1980.

Dougherty D, Bjork J, Bennett R, Moeller F. The effects of cumulative alcohol dosing procedure on laboratory aggression in women and men. Journal of Studies on Alcohol 1999; 60(3):322-329.

Duncan J. Disorganisation of behaviour after frontal lobe damage. Cognitive Neuropsychology 1986; 3:271-290.

Edwards K. The interplay of affect and cognition in attitude formation and change. Journal of Personality and Social Psychology 1990; 59:202-216.

Elliott R, Dolan RJ, Frith CD. Dissociable functions in the medial and lateral orbitofrontal cortex: evidence from human neuroimaging studies. Cerebral Cortex 2000; 10:308-317.

Eslinger PJ, Damasio AR. Severe disturbance of higher cognition after bilateral frontal lobe ablation: patient EVR. Neurology 1985; 35:1731-1741.

Everitt B, Cador M, Robbins TW. Interactions between the amygdala and ventral striatum in stimulus-reward associations: Studies using a second-order schedule of sexual reinforcement. Neuroscience 1989; 30:63-75.

Everitt B, Robbins TW. Amygdala-ventral striatal interactions and reward-related processes. In: Aggleton JP, editor. The amygdala: Neurobiological aspects of emotion, memory and mental dysfunction. New York: Wiley, 1992.

Everitt B, Cardinal R, Hall J, Parkinson JA, Robbins T. Differential involvement of amygdala subsystems in appetitive conditioning and drug addiction. In: Aggleton J, editor. The amygdala. In press.

Fine C, Blair RJR. The cognitive and emotional effects of amygdala damage. Neurocase 2000; 6. In press.

Fisher L, Blair RJR. Cognitive impairment and its relationship to psychopathic tendencies in children with emotional and behavioural difficulties. Journal of Abnormal Child Psychology 1998; 26:511-519.

Flaherty CF. Incentive contrast: A review of behavioral changes following shifts in reward. Animal Learning and Behavior 1982; 10(4):409-440.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS et al. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 1995; 57:109-128.

Fonseca A, Yule W. Personality and antisocial behaviour in children and adolescents: An enquiry into Eysenck's and Gray's theories. Journal of Abnormal Child Psychology 1995; 23:767-781.

Fowles D. The three arousal model: Implications of Gray's two-factor learning theory for heart rate, electrodermal activity, and psychopathy. Psychophysiology 1980; 17:87-104.

Frith CD, Corcoran R. Exploring "theory of mind" in people with schizophrenia. Psychological Medicine 1996; 6:521-530.

Frith CD, Frith U. Interacting minds – a biological basis. Science 1999; 286(5445):1692-1695.

Frith U, Happé F, Siddons F. Social Development 1994; 3(2):108-124.

Frith U. Autism: Explaining the enigma. Oxford: Blackwell, 1989.

Frith U, Morton J, Leslie A. The cognitive basis of a biological disorder: Autism. Trends in Neuroscience 1991; 14:433-438.

Frye D, Zelazo PD, Palfai T. Theory of mind and rule-based reasoning. Cognitive Development 1995; 10:483-527.

Frye D, Zelazo PD, Brooks PJ, Samuels MC. Inference and action in early causal reasoning. Developmental Psychology 1996; 32:120-131.

Fudge J, Powers J, Haber S, Caine E. Considering the role of the amygdala in psychotic illness: a clinicopathological correlation. Journal of Neuropsychiatry 1997; 10:383-394.

Fuster JM. The prefrontal cortex. New York: Raven Press, 1980.

Gaffan D, Murray EA. Amygdalar interaction with the mediodorsal nucleus of the thalamus and the ventromedial prefrontal cortex in stimulus-reward associative learning in the monkey. Journal of Neuroscience 1990; 10(11):3479-3493.

Gallagher M, Schoenbaum G. Functions of the amygdala and related forebrain areas in attention and cognition. Annals of the New York Academy of Sciences 1999; 877:397-411.

Gallagher H, Happé F, Brunswick N, Fletcher P, Frith U, Frith C. Reading the mind in cartoons and stories: an fMRI study of "theory of mind" in verbal and nonverbal tasks. Neuropsychologia 2000; 38(1):11-21.

Goel V, Grafman J, Sadato N, Hallett M. Modeling other minds. Neuroreport 1995; 11:1741-1746.

Grafman J, Schwab K, Warden D, Pridgen B, Brown H, Salazar A. Frontal lobe injuries, violence, and aggression: A report of the Vietnam head injury study. Neurology 1996; 46:1231-1238.

Gray JA. The neuropsychology of anxiety. New York: Oxford University Press, 1982.

Gray JA. Issues in the neuropsychology of anxiety. In: Tuma A, Maser J, editors. Anxiety and the anxiety disorders. Hillsdale, NJ: Lawrence Erlbaum Associates, 1985: 5-25.

Gray JA. The psychology of fear and stress. 2 ed. Cambridge: Cambridge University Press, 1987.

Gray JA. The contents of consciousness: A neuropsychological conjecture. Behavioral and Brain Sciences 1995; 18:659-722.

Grossberg S. Processing of expected and unexpected events during conditioning and attention: A psychophysiological theory. Psychological Review 1982; 89(5):529-572.

Grossberg S, Schmajuk N. Neural dynamics of attentionally-modulated Pavlovian conditioning: Conditioned reinforcement, inhibition and opponent processing. Psychobiology 1987; 15:195-240.

Grossberg S, Gutowski S. Neural dynamics of decision making under risk: Affective balance and cognitive-emotional interactions. Psychological Review 1987; 94:300-318.

Grossberg S, Levine D. Neural dynamics of attentionally modulated Pavlovian conditioning: Blocking, inter-stimulus interval, and secondary conditioning. Applied Optics 1987; 26:5015-5030.

Grossberg S. The imbalanced brain: from normal behavior to schizophrenia. Biological Psychiatry 2000; 48:81-98.

Hamann S, Stefanacci L, Squire L, Adolphs R, Tranel D, Damasio H et al. Recognizing facial emotion. Nature 1996; 379:497.

Happé FGE. Communicative competence and theory of mind in autism: A test of Relevance theory. Cognition 1993; 48:101-119.

Happé FGE. An advanced test of Theory of Mind: Understanding of story characters' thoughts and feelings in able autistic, mentally handicapped, and normal children and adults. Journal of Autism and Developmental Disorders 1994; 24:129-154.

Happé F, Frith U. The neuropsychology of autism. Brain 1996; 119:1377-1400.

Hare RD. Acquisition and generalization of a conditioned-fear response in psychopathic and nonpsychopathic criminals. Journal of Psychology 1965; 59:367-370.

Hare RD. Temporal gradient of fear arousal in psychopaths. Journal of Abnormal Psychology 1965; 70(6):442-445.

Hare RD, Frazelle J, Cox D. Psychopathy and physiological responses to threat of an aversive stimulus. Psychophysiology 1978; 15(2):165-172.

Hare RD. The Hare Psychopathy Checklist-Revised. Toronto: 1991.

Harpur TJ, Hare RD, Hakstein A. Two-factor conceptualization of psychopathy: Construct validity and assessment implications. Psychological Assessment: A Journal of Consulting and Clinical Psychology 1989; 1(1):6-17.

Harpur TJ, Hare RD. Assessment of psychopathy as a function of age. Journal of Abnormal Psychology 1994; 103(4):604-609.

Hatfield T, Han J, Conley M, Gallagher M, Holland P. Neurotoxic lesions of basolateral, but not central, amygdala interfere with pavlovian second-order conditioning and reinforcer devaluation effects. Journal of Neuroscience 1996; 16(16):5256-5265.

Hayman L, Rexer J, Pavol M, Strite D, Meyers C. Klüver-Bucy syndrome after bilateral selective damage of amygdala and its connections. Journal of Neuropsychiatry and Clinical Neurosciences 1998; 10:354-358.

Hitchcock J, Davis M. Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. Behavioral Neuroscience 1986; 100:11-22.

Hitchcott P, Phillips G. Double dissociation of the behavioural effects of R(+) 7-OH-DPAT infusions in the central and basolateral amygdala nuclei upon Pavlovian and instrumental conditioned appetitive behaviours. Psychopharmacology 1998; 140:458-469.

Hughes C, Russell J, Robbins TW. Evidence for executive dysfunction in autism. Neuropsychologia 1994; 32:477-492.

Hughes C, Russell J. Autistic children's difficulty with mental disengagement from an object: Its implications for theories of autism. Developmental Psychology 1993; 29:498-510.

Hughes C. Executive function in preschoolers: Links with theory of mind and verbal ability. British Journal of Developmental Psychology 1998a; 16:233-253.

Hughes C. Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? Developmental Psychology 1998b; 34:1326-1339.

Iversen S, Mishkin M. Perseverative interference in monkey following selective lesions of the inferior prefrontal convexity. Experimental Brain Research 1970; 11:376-386.

Jacobsen R. Disorders of facial recognition, social behaviour and affect after combined bilateral amygdalotomy and subcaudate tractotomy - a clinical and experimental study. Psychological Medicine 1986; 16:439-450.

Jenkins R, Lavie N, Driver J. Ignoring famous faces: Category-specific dilution of distractor interference. In prep.

Johnsrude IS, Owen AM, Zhao WV, White NM. Conditioned preference in humans: A novel experimental approach. Learning and Motivation 1999; 30:250-264.

Jones B, Mishkin M. Limbic lesion and the problem of stimulus-reinforcement association. Experimental Neurology 1972; 36:362-377.

Killcross, Robbins T, Everitt B. Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. Nature 1997; 388:377-380.

Kim J, Rison R, Fanselow M. Effects of amygdala, hippocampus, and periacqueductal gray lesions on short- and long-term contextual fear. Behavioral Neuroscience 1993; 107(6):1093-1098.

LaBar JS, Gatenby J, Gore J, LeDoux JE, Phelps EA. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron 1998; 20, 937-945.

LaBar JS, LeDoux J, Spencer D, Phelps E. Impaired fear conditioning following unilateral temporal lobectomy in humans. Journal of Neuroscience 1995; 15:6846-6855.

Lang PJ, Greenwald MK, Bradley MM, Hamm AO. Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology 1993; 30:261-273.

LaPierre D, Braun CMJ, Hodgins S. Ventral frontal deficits in psychopathy: neuropsychological test findings. Neuropsychologia 1995; 33(2):139-151.

Lawrie SM, Abukmeil SS. Brain abnormality in schizophrenia: a systematic and quantitative review of volumetric magnetic resonance imaging studies. [Review]. British Journal of Psychiatry 1998; 172:110-120.

LeDoux JE, Cicchetti P, Xagoraris A, Romanski LM. The lateral amygdaloid nucleus: Sensory interface of the amygdala in fear conditioning. Journal of Neuroscience 1990; 10(4):1062-1069.

LeDoux JE. The emotional brain. New York: Weidenfeld & Nicolson, 1998.

Leslie A, Roth D. What representation teaches us about metarepresentation. In: Baron-Cohen S, Tager-Flusberg H, Cohen DJ, editors. Understanding Other Minds: Perspectives from Autism. Oxford: Oxford University Press, 1993: 83-111.

Leslie AM, Thaiss L. Domain specificity in conceptual development: neuropsychological evidence from autism. Cognition 1992; 43:225-251.

Lewis M, Goldberg S. The acquisition and violation of expectancy: an experimental paradigm. Journal of Experimental Child Psychology 1969; 7:70-80.

Luria A. Higher cortical functions in man. New York: Basic Books, 1966.

Lykken DT. A study of anxiety in the sociopathic personality. Journal of Abnormal and Social Psychology 1957; 55:6-10.

MacDowell KA, Mandler G. Construction of emotion: Discrepancy, arousal and mood. Cognition and Emotion 1989; 13(2):105-124.

Mackintosh NJ. A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review 1975; 82:276-298.

Malkova L, Gaffan D, Murray EA. Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. Journal of Neuroscience 1997; 17(15):6011-6020.

Mandler G. The interruption of behaviour. In: Levine D, editor. Nebraska Symposium on Motivation. Lincoln, Nebraska: University of Nebraska Press, 1964.

Mandler G. Mind and emotion: Psychology of emotion and stress. New York: Norton & Company, 1984.

Mandler G. Emotions, evolution, and aggression: Myths and conjectures. In: Strongman KT, editor. International review of studies on emotion. Chichester: John Wiley & Sons, 1991.

Maren S. Overtraining does not mitigate contextual fear conditioning deficits produced by neurotoxic lesions of the basolateral amygdala. Journal of Neuroscience 1998; 18(8):3088-3097.

Martinius J. Homicide of an aggressive adolescent boy with right temporal lesion: a case report. Neuroscience and Biobehavioral Reviews 1983; 7:419-422.

McDonald R, White N. A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum. Behavioral Neuroscience 1993; 107, 3-22.

Miller E. Verbal fluency as a function of a measure of verbal intelligence and in relation to different types of cerebral pathology. British Journal of Clinical Psychology 1984; 23:53-57.

Mineka S, Cook M. Mechanisms involved in the observational conditioning of fear. Journal of Experimental Psychology: General 1993; 122:23-38.

Mitchell D, Colledge E, Blair RJR. Psychopathy: A selective expression recognition impairment? In prep.

Montague PR, Dayan P, Sejnowski J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience 1996; 16(5):1936-1947.

Morris J, Friston KJ, Buchel C, Frith CD, Young AW, Calder AJ et al. A neuromodulatory role for the human amygdala in processing emotional facial expressions. Brain 1998; 121:47-57.

Morris J, Frith CD, Perrett D, Rowland D, Young A, Calder AJ et al. A differential response in the human amygdala to fearful and happy facial expressions. Nature 1996; 383:812-815.

Nelson H. A modified card sorting task sensitive to frontal lobe defects. Cortex 1976; 12:313-324.

Nelson HE, Willison J. The National Adult Reading Test. 2nd ed. Windsor (UK): NFER-Nelson, 1991.

Newman JP, Widom CS, Nathan S. Passive avoidance in syndromes of disinhibition: Psychopathy and extraversion. Journal of Personality and Social Psychology 1985; 48:1316-1327.

Newman JP, Kosson DS. Passive avoidance learning in psychopathic and nonpsychopathic offenders. Journal of Abnormal Psychology 1986; 95:252-256.

Newman JP, Patterson CM, Howland EW, Nichols SL. Passive avoidance in psychopaths: The effects of reward. Personality & Individual Differences 1990; 11:1101-1114.

Newman J, Patterson C, Kosson D. Response perseveration in psychopaths. Journal of Abnormal Psychology 1987; 96(2):145-148.

Newman J, Schmitt W, Voss W. The impact of motivationally neutral cues on psychopathic individuals: Assessing the generality of the response modulation hypothesis. Journal of Abnormal Psychology 1997; 106:563-575.

Oatley K, Johnson-Laird PN. Towards a cognitive theory of emotions. Cognition and Emotion 1987; 1:29-50.

Otsuka H, Harada M, Mori K, Hisaoka S, Nishitani H. Brain metabolites in the hippocampus-amygdala region and cerebellum in autism: an H-MR spectroscopy study. Neuroradiology 1999; 41:517-519.

Ozonoff S, Rogers S, Pennington B. Executive function deficits in high-functioning autistic children: Relationship to theory of mind. Journal of Child Psychology and Psychiatry 1991; 32:1081-1106.

Ozonoff S. Components of executive function in autism and other disorders. In: Russell J, editor. Autism as an executive disorder. Oxford: Oxford University Press, 1997: 179-211.

Parkinson JA, Robbins TW, Everitt BJ. Dissociable roles of the central and basolateral amygdala in appetitive emotional learning. European Journal of Neuroscience 2000; 12:405-413.

Patrick CJ, Bradley M, Lang PJ. Emotion in the criminal psychopath: startle reflex modulation. Journal of Abnormal Psychology 1993; 102(1):82-92.

Patterson C, Newman J. Reflectivity and learning from aversive events: toward a psychological mechanism for the syndromes of disinhibition. Psychological Review 1993; 100(4):716-736.

Pearce JM, Hall G. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review 1980; 87:532-552.

Pennington BF, Ozonoff S. Executive functions and developmental psychopathology. Journal of Child Psychology and Psychiatry 1996; 37:51-87.

Perner J, Wimmer H. "John thinks that Mary thinks that ... " Attribution of second-order beliefs by 5-10 year old children. Journal of Experimental Child Psychology 1985; 39:437-471.

Perner J, Frith U, Leslie AM, Leekam SR. Exploration of the autistic child's theory of mind: Knowledge, belief and communication. Child Development 1989; 60:689-700.

Perner J. The meta-intentional nature of executive functions and theory of mind. In: Carruthers P, Boucher J, editors. Language and thought: Interdisciplinary themes. Cambridge: Cambridge University Press, 1998: 270-316.

Perner J, Stummer S, Lang B. Executive functions and theory of mind: Cognitive complexity or functional dependence? In: Zelazo PD, David P, Astington JW, et al., editors. Developing theories of intention: Social understanding and self control. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1999: 133-152.

Perner J, Lang B. Theory of mind and executive function: is there a developmental relationship? In: Baron-Cohen S, Tager-Flusberg H, Cohen DJ, editors. Understanding other minds: perspectives from developmental cognitive neuroscience. Oxford: Oxford University Press, 2000: 150-181.

Petrides M. Nonspatial conditional learning impairment in patients with unilateral frontal but not unilateral temporal lobe excisions. Neuropsychologia 1990; 28:137-149.

Phelps E, LaBar J, Anderson A, O'Conner K, Fulbright R, Spencer D. Specifying the contributions of the human amygdala to emotional memory: a case study. Neurocase 1998; 4:527-540.

Phillips M, Young A, Senior C, Brammer M, Andrew C, Calder A et al. A specified neural substrate for perceiving facial expressions of disgust. Nature 1997; 389:495-498.

Ploghaus A, Tracey I, Gati J, Clare S, Menon RS, Matthews P et al. Dissociating pain from its anticipation in the human brain. Science 1999; 284:1979-1981.

Premack D, Woodruff G. Does the chimpanzee have a "theory of mind"? Behavioral and Brain Sciences 1978; 4:515-526.

Quay HC. The psychobiology of undersocialized aggressive conduct disorder: A theoretical perspective. Development and psychopathology 1993; 5:165-180.

Rahman S, Sahakian B, Hodges J, Rogers R, Robbins T. Specific cognitive deficits in mild frontal variant frontotemporal dementia. Brain 1999; 122:1469-1493.

Raine A, Buchsbaum M, LaCasse L. Brain abnormalities in murderers indicated by Positron Emission Tomography. Biological Psychiatry 1997; 42:495-508.

Raine A, Meloy J, Bihrle S, Stoddard J, LaCasse L, Buchsbaum M. Reduced prefrontal and increased subcortical brain functioning assessed using Positron Emission Tomography in predatory and affective murderers. Behavioral Sciences and the Law 1998; 16:319-332.

Raine A, Lencz T, Bihrle S, LaCasse L, Colletti P. Reduced prefrontal gray matter volume and reduced autonomic activity in Antisocial Personality Disorder. Archives of General Psychiatry 2000; 57(2):119-127.

Rescorla RA, Wagner AR. A theory of Pavlovian conditioning. Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II: Current research and theory. New York: Appleton-Century-Crofts, 1972.

Robbins TW. Dissociating executive functions of the prefrontal cortex. In: Roberts A, Robbins T, Weiskrantz L, editors. The prefrontal cortex: Executive and cognitive functions. New York: Oxford University Press, 1996: 117-130.

Rogers RD, Everitt B, Baldacchino A, Blackshaw AJ, Swainson R, Wynne K et al. Dissociable deficits in decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: Evidence for monoaminergic mechanisms. Neuropsychopharmacology 1999; 20(4):322-339.

Rolls ET. A theory of emotion, and its application to understanding the neural basis of emotion. Cognition and Emotion 1990; 4:161-190.

Rolls ET, Treves A. Neural networks and brain function. Oxford: Oxford University Press, 1998.

Rolls E, Hornak J, Wade D, McGrath J. Emotion related learning in patients with social and emotional changes associated with frontal lobe damage. Journal of Neurology, Neurosurgery, and Psychiatry 1994; 57:1518-1524.

Rolls E. The orbitofrontal cortex. Philosophical Transactions of the Royal Society of London, Series B 1996; 351:1433-1444.

Rolls E. The orbitofrontal cortex and reward. Cerebral Cortex 2000; 10:284-294.

Roth D, Leslie AM. The recognition of attitude conveyed by utterance: A study of preschool and autistic children. British Journal of Developmental Psychology 1991; 9:315-330.

Russell J. At two with nature: Agency and the development of self-world dualism. In: Bermudez J, Marcel AJ, Eilan N, editors. The body and the self. Cambridge, MA: MIT Press, 1995: 127-151.

Russell J. Agency: Its role in mental development. Hove: The Psychology Press, 1996.

Russell J. How executive disorders can bring about an inadequate 'theory of mind'. In: Russell J, editor. Autism as an executive disorder. Oxford: Oxford University Press, 1997: 256-304.

Saver JL, Damasio AR. Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. Neuropsychologia 1991; 29:1241-1249.

Scerbo A, Raine A, O'Brien M, Chan C, Rhee C, Smiley N. Reward dominance and passive avoidance learning in adolescent psychopaths. Journal of Abnormal Child Psychology 1990; 18(4):451-463.

Schmajuk N, Lam Y, Gray JA. Latent inhibition: A neural network approach. Journal of Experimental Psychology: Animal Behavior Processes 1996; 22(3):321-349.

Schmauk PJ. Punishment, arousal, and avoidance learning in psychopaths. Journal of Abnormal Psychology 1970; 76:325-335.

Schmitt W, Brinkley C, Newman J. Testing Damasio's somatic marker hypothesis with psychopathic individuals: risk takers or risk averse? Journal of Abnormal Psychology 1999; 108(3):538-543.

Schoenbaum G, Chiba AA, Gallagher M. Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. Journal of Neuroscience 1999; 19(5):1876-1884.

Scott S, Young A, Calder A, Hellawell D, Aggleton J, Johnson M. Impaired auditory recognition of fear and anger following bilateral amygdala lesions. Nature 1997; 385:254-257.

Shallice T, Evans ME. The involvement of the frontal lobes in cognitive estimation. Cortex 1978; 14:294-303.

Shallice T. Specific impairments of planning. Philosophical Transactions of the Royal Society of London B 1982; 298:199-209.

Shallice T. From neuropsychology to mental structure. Cambridge: Cambridge University Press, 1988.

Shallice T, Burgess PW. The domain of supervisory processes and temporal organization of behaviour. Philosophical Transactions of the Royal Society of London B 1996; 351:1405-1412.

Shapiro SK, Quay HC, Hogan AE, Schwartz KP. Response perseveration and delayed responding in undersocialised aggressive conduct disorder. Journal of Abnormal Psychology 1988; 97:371-373.

Shimamura A. The role of prefrontal cortex in dynamic filtering. Psychobiology 2000; 28(2):207-218.

Sokolov EN. Neuronal models and the orienting reflex. In: Brazier MAB, editor. The central nervous system and behavior. Madison, NJ: Madison Printing Co, Inc., 1960: 187-276.

Sprengelmeyer R, Young AW, Schroeder U, Grossenbacher P, Federlein J, Büttner T et al. Knowing no fear. Proceedings of the Royal Society of London B 1999; 266:2451-2456.

Stroop JR. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 1935; 18:643-662.

Sutton RS, Barto AG. Toward a modern theory of adaptive networks: Expectation and prediction. Psychological Review 1981; 88(2):135-170.

Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, Mass.: MIT Press, 1998.

Thornquist M, Zuckerman M. Psychopathy, passive-avoidance learning and basic dimensions of personality. Personality & Individual Differences 1995; 19(4):525-534.

Thorpe SJ, Rolls ET, Maddison S. The orbitofrontal cortex: Neuronal activity in the behaving monkey. Experimental Brain Research 1983; 49:93-115.

Tolman E. Purposive behavior in animals and men. New York: London: The Century Co., 1932.

Tranel D, Hyman B. Neuropsychological correlates of bilateral amygdala damage. Archives of Neurology 1990; 47(3):349-355.

Volkow N, Tandredi L. Neural substrates of violent behaviour: A preliminary study with Positron Emission Tomography. British Journal of Psychiatry 1987; 151:668-673.

Wallace J, Vitale J, Newman J. Response modulation deficits: Implications for the diagnosis and treatment of psychopathy. Journal of Cognitive Psychotherapy: An International Quarterly 1999; 13[1], 55-70.

Warrington EK. Recognition memory test. Windsor (UK): NFER-Nelson, 1984.

Warrington EK. The Camden memory tests. Hove (UK): Psychology Press, 1996.

Warrington EK, McKenna P, Orpwood L. Single word comprehension: a concrete and abstract word synonym test. Neuropsychological Rehabilitation 1998; 8.

Watanabe M. The appropriateness of behavioral responses coded in post-trial activity of primate prefrontal units. Neuroscience Letters 1989; 101(1):113-117.

Whalen P, Shin L, McInerney S, Rauch S. Greater fMRI activation to fearful vs. angry facial expressions in the amygdaloid region. Neuroscience Abstracts 1998; 24, 692.

Whalen P. Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. Current Directions in Psychological Science 1998; 7:177-188.

Whalen P, Rauch S, Etcoff N, McInerney S, Lee M, Jenike M. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. Journal of Neuroscience 1998; 18:411-418.

Whitelaw R, Markou A, Robbins TW, Everitt B. Excitotoxic lesions of the basolateral amygdala impair the acquisition of cocaine-seeking behavior under and second order schedule of reinforcement. Psychopharmacology 1996; 127:213-224.

Whiteley A, Warrington EK. Whiteley A, Warrington EK. Selective impairment of topographical memory: a single case study. Journal of Neurology, Neurosurgery and Psychiatry 1978; 41, 575-578.

Wilson B, Alderman N, Burgess P, Emslie H, Evans J. Behavioural assessment of the dysexecutive syndrome (BADS). Bury St Edmonds, UK: Thames Valley Test Company, 1996.

Wilton KM, Boersma FJ. Eye movements, surprise reactions and cognitive development. Rotterdam: Rotterdam University Press, 1974.

Wimmer H, Perner J. Beliefs about beliefs: Representation and the constraining function of wrong beliefs in young children's understanding of deception. Cognition 1983; 13:103-128.

Young AW, Ellis AW, Flude BM, McWeeny KH, Hey DC. Face-name interference. Journal of Experimental Psychology: Human Perception and Performance 1986; 12:466-475.

Appendix

Tasks 8 & 9: Example of an Advanced Theory of Mind story.

Simon is a big liar. Simon's brother Jim knows this, he knows that Simon never tells the truth! Now yesterday Simon stole Jim's ping-pong bat, and Jim knows Simon has hidden it somewhere, though he can't find it. He's very cross. So he finds Simon and he says, "Where is my ping-pong bat? You must have hidden it either in the cupboard or under your bed, because I've looked everywhere else. Where is it, in the cupboard or under your bed?" Simon tells him the bat is under the bed.

The participant is asked:

Q1: "Was it true, what Simon told Jim?"

Q2: "Where will Jim look for his ping-pong bat?"

Q3: "Why will Jim look there for his bat?"

Task 10: Example of a non-literal speech comprehension story

Karen is very thin, but thinks she needs to go on a diet. She tells her friend, Jen, that she is going on a diet. Jen thinks that Karen is too thin and says,

"That's good, because you're so enormous, Karen."

The participant is asked: "What does Jen mean by this?"

She continues,

"You're a stick."

The participant is asked: "What does Jen mean by this?"