PREDICTION ERROR DEPENDENT CHANGES IN BRAIN CONNECTIVITY DURING ASSOCIATIVE LEARNING.

HANNEKE DEN OUDEN

DISSERTATION SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

OF

UNIVERSITY COLLEGE LONDON

WELLCOME TRUST CENTRE FOR NEUROIMAGING

INSTITUTE OF NEUROLOGY

Declaration

I, Hanneke Eveline Maria den Ouden, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

17th April, 2009

"From their errors the wise and good learn wisdom for the future"

- Plutarch -

Abstract

One of the fundaments of associative learning theories is that surprising events drive learning by signalling the need to update one's beliefs. It has long been suggested that plasticity of connection strengths between neurons underlies the learning of predictive associations: Neural units encoding associated entities change their connectivity to encode the learned associative strength. Surprisingly, previous imaging studies have focused on correlations between regional brain activity and variables of learning models, but neglected how these variables changes in inter-regional connectivity. Dynamic Causal Models (DCMs) of neuronal populations and their effective connectivity form a novel technique to investigate such learning dependent changes in connection strengths.

In the work presented here, I embedded computational learning models into DCMs to investigate how computational processes are reflected by changes in connectivity. These novel models were then used to explain fMRI data from three associative learning studies. The first study integrated a Rescorla-Wagner model into a DCM using an incidental learning paradigm where auditory cues predicted the presence/absence of visual stimuli. Results showed that even for behaviourally irrelevant probabilistic associations, prediction errors drove the consolidation of connection strengths between the auditory and visual areas. In the second study I combined a Bayesian observer model and a nonlinear DCM, using an fMRI paradigm where auditory cues differentially predicted visual stimuli, to investigate how predictions about sensory stimuli influence motor responses. Here, the degree of striatal prediction error activity controlled the plasticity of visuo-motor connections. In a third study, I used a nonlinear DCM and data from a fear learning study to demonstrate that prediction error activity in the amygdala exerts a modulatory influence on visuo-striatal connections.

Though postulated by many models and theories about learning, to our knowledge the work presented in this thesis constitutes the first direct report that prediction errors can modulate connection strength.

Acknowledgements

Throughout my years at the FIL, my supervisor Klaas Enno Stephan gave me a combination of support, guidance, advice and freedom to pursue my own ideas, that was extraordinary and for which I am immensely grateful. Huge thanks also to Karl Friston, my other supervisor, whose comments and suggestions never failed to be insightful, and his exceptional ability to combine technical expertise with philosophical thinking will forever be an inspiring example. Both have fundamentally shaped my views on science and the sort of research I intend to pursue.

I have learnt from so many people at the FIL that I can't begin to mention them all, but I want to especially thank Jean Daunizeau for teaching me all I know about Bayesian observer models as well as instilling a general dislike of ad hoc non-principled models, Jon Roiser for interesting discussions and general statistical guidance, Nathaniel Daw for great help on the computational modelling, Quentin Huys for interesting discussions and patience explaining about Kalman filters, Guillaume Flandin for his truly endless patience with my even more endless SPM questions, Mohammed Seghier for all the interesting discussions about DCM, and everyone at the Monday afternoon Neurobiology meetings, for valuable comments and insights and general good fun. Of course I also want to thank Andre, Rosalyn, Justin, CC, Guillaume, Mohammend, Wako and Ben for being such great office mates!

In the last five years England and London became my home, and I would also like say thanks to the people who made my life there so amazing: my fellow Wellcome Trust PhD students Louise, Rosie, Kieran, Curtis and Dave, for our ever more lavish dinners, Sinterklaas poems and general moral support throughout the PhD; my climbing friends for introducing me to a world with vertical features; Nick, Paul, Al, Che, Tim, Jim, Ian, Adam, Dave (both), Duc, and especially all the girly climbing club members, Eu Lee, Manchi, Caroline, Fran and Nening; the Valentine Road crew; Marianne, Jonathan and Cas and of course Hermann and Marieke for having me stay over for the last months; Jovanka for the weekly Sunday morning calls,

Lonneke, Piet, Nienke, Jasper, and of course Frank and Marieke, who are always there.

Finally, I do not know how to express my gratitude to my family for their support and love. Mama, dankjewel voor de wandelingen en wekelijkse telefoontjes. Papa, jouw emails zijn altijd de liefste en leukste. Papa, Mama, Mies en Rein, ik ben weer thuis.

Table of Contents

List of Figures

List of Tables

Outline and Aims

The aim of this thesis was to assess the role of prediction errors and connectivity changes in associative learning, using a combination of formal learning models and DCM for fMRI. A range of associative learning tasks was used with increasing behavioural relevance of the associative relationships. This thesis is organized as follows:

Chapter 1 – Introduction – This chapter is divided into two parts. The first part gives a brief overview of the field of associative learning, and discusses in more detail the role of prediction errors and synaptic plasticity. The second part describes and compares classical reinforcement learning models and Bayesian ideal observer models, both of which were used to model the behavioural and fMRI data described in this thesis.

Chapter 2 –Methods – This chapter is divided into two parts. The first part describes DCM, including both the original formulation and a novel extension which allows for second order modulation. Both these tools will be used for hypothesis testing in the subsequent chapters. The second part describes Bayesian model selection, which is used to decide which of a group of models is the best model for a given dataset. In subsequent chapters this tool is applied to both DCMs and behavioural data.

Chapters 3-5 – Results chapters – These chapters describe the experimental work: the aims, the hypotheses / models tested, the set up and the outcomes of three studies. The specific goals of each study were the following:

- To investigate associative learning of task-irrelevant associations, at the level of the sensory cortex, and more specifically to test for changes in connectivity between the sensory areas involved (**Chapter 3**).
- To explore stimulus independent and stimulus bound surprise processing when subjects learn dynamically changing relationships between sensory stimuli and

to identify an underlying second order connectivity model for the SPM results (**Chapter 4**).

- To investigate prediction error processing in an aversive reinforcement learning paradigm, the connectivity parameters of the underlying causal model (**Chapter 5**).

Chapter 6 – General Discussion and Conclusion – This chapter provides a general discussion and the conclusions of this work; presents its contributions to the field; and suggests directions for future research.

Chapter 1

Introduction

In order to interpret incoming sensory information and predict future events, our brains need to construct models of the world that represent how external events are causally linked. In his 1898 dissertation on animal intelligence, Edward Thorndike first proposed a theory of associative learning in animals (Thorndike E.L., 1898). He posed the so-called 'law of effect', arguing that learning consists of the establishment of associations that are formed when responses are followed by rewards. This theory formed the basis of a century of stimulus-response and stimulus-stimulus associative learning.

It is easy to see how predicting relevant stimuli in the environment such as food and predators can boost adaptive fitness, allowing one to seek out juicy fruits and avoid painful shocks in cognitive neuroscience experiments. However phenomena like the mismatch negativity and sensory preconditioning (see **Section 1.1.2** and **3.1**) show that the brain's predictions about the environment are not limited to behaviourally relevant stimuli. One hypothesis is that sensory perception rests upon active prediction of the environment rather than just passive reception of sensory information. Here, sensory perception is a recurring input-match-prediction loop where beliefs about the environment are continuously updated to predict future sensory inputs.

After decades of animal research into the neural mechanisms of associative learning in animals, functional neuroimaging has allowed for extension of these investigations to human subjects. This thesis combines two recent developments in human functional magnetic resonance imaging (fMRI) methods: (i) the use of formal associative learning models to explain measured BOLD responses and (ii) physiologically motivated models of brain connectivity. Changes in connectivity have long been thought to be central to the physiological implementation of learning

(see (Hebb, 1949)), and in the work presented here, associative learning models are embedded in physiological models of connectivity to investigate changes during associative learning. Within this novel framework associative learning is investigated in three paradigms, in which the probabilistic stimulus associations range from affectively neutral to noxious, static to changing, and incidental to task relevant.

This chapter is divided into two parts. In the first part I will give an overview of the fundaments of associative learning, including the notion of prediction errors in both reward and non-reward based contexts, the hypothesis of predictive coding as a fundamental mechanism of brain functioning, and the neurophysiological mechanisms underlying associative learning. The second part of this chapter then describes classical and Bayesian learning models, and their advantages and limitations for investigating associative learning processes.

1.1 Associative learning

Behavioural research on how humans and animals learn to predict positive and negative stimuli in their environment was pioneered by Ivan Pavlov in the late nineteenth century. Originally studying the digestive system and the chemical composition of saliva, Pavlov observed that dogs started salivating before food was actually delivered. Upon closer examination it transpired that the salivating response commenced when a bell was rung by his assistant to indicate that the food was ready. Pavlov abandoned the study of saliva chemistry in favour of further investigating this ‘psychic secretion’ response, as he termed it.

In classical, or Pavlovian, conditioning, a motivationally significant unconditioned stimulus (US; the food stimulus, often also termed ‘reinforcer’), elicits an unconditioned response (UR; salivation). When an affectively neutral conditioned stimulus (CS; the bell) regularly precedes the US, the CS will eventually also elicit salivation as a conditioned response (CR) (see **Table 1.1**). This formation of stimulus-stimulus associations is fundamentally important as it allows animals and humans to predict and prepare for biologically important events.

Later experiments showed that temporal pairing of a cue and reinforcer alone is not enough to learn a cue-outcome association. This was demonstrated by a phenomenon called 'blocking' ((Kamin LJ, 1969), see **Table 1.1**). In the first stage of a blocking paradigm, an initially neutral cue A is paired with a reinforcer, and another neutral cue B is presented but never paired. After stage 1, A will elicit a conditioned response, but B will not. In a second stage, A is presented in combination with another cue X, and B with Y, and both compound cues are repeatedly paired with the reinforcer. After stage 2, Y will elicit a conditioned response, whereas X will not, even though both cues have been paired with a reinforcer equally often. This can be explained by noting that for the AX compound, The reinforcer could be fully predicted by A alone, rendering X redundant, whereas for the BY compound, B could not explain the reinforcer, leaving it 'free' to be associated to Y. This suggests that when a reinforcer is completely predicted by the cue(s), no further learning occurs; in other words, A had 'blocked' learning an association between X and the reinforcer.

Based on this effect, Kamin concluded that simply pairing of the cue and reinforcer is not enough; the presence of the reinforcer has to be surprising in order to establish an association (Kamin LJ, 1969). This notion of surprise lies at the heart of nearly all associative learning theories. The basic idea is that a mismatch between predicted and actual outcome signals that the internal model's predictions are wrong and need to be updated. Such surprising events are known as prediction errors. The next section reviews accumulating neurobiological evidence that the brain indeed processes surprising events differently from predicted events

1.1.1 Neuronal prediction errors

1.1.1.1 Dopamine & Ventral striatum

Dopamine (DA) neurons in the ventral striatum in macaques strongly increase their firing rate when salient or rewarding stimuli are presented. These rewarding stimuli can be primary rewards, such as food and water, but also arbitrary cues that are predictive of primary rewards (Ljungberg et al., 1992;Romo and Schultz, 1990). These DA responses generalise to stimuli that are perceptually similar to the reward-predicting cues, and the responses show other characteristics that parallel those

reported in behavioural studies, such as blocking (see **Table 1.1** ; (Waelti et al., 2001)).

In a seminal series of studies, Schultz and colleagues carefully investigated the nature of this phasic dopamine firing during classical conditioning using single unit recordings in the macaque ventral tegmental area (VTA) (Mirenowicz and Schultz, 1994;Mirenowicz and Schultz, 1996;Romo and Schultz, 1990;Schultz, 1998). When a monkey is first presented with an arbitrary visual cue followed by a juice reward, the dopamine neurons strongly increase firing in response to the reward, but not in response to the cue (see **Figure 1.1**). Over time, as the monkey learns the cue-reward association, firing rates increase when the cue itself is presented. This response parallels the behaviourally observed conditioned response to the cue, which has now become a reward in itself. Furthermore, as the association is learned, rewards evoke progressively smaller increases in firing; when the reward is fully predicted, firing rates no longer increase. Finally, firing rates decrease to below baseline when a predicted reward is omitted. This pattern of responses suggest that what the dopamine neurons respond to is not reward per se, but its prediction error. When the reward is presented before the association is learned, it is unpredicted and increased firing rates reflect the large difference between prediction and observed outcome. When the reward is fully predicted by the cue it elicits no response, but the presentation of the cue itself is surprising and does elicit an increased response. When a predicted reward is omitted, the difference between the outcome (no reward) and prediction (reward) is negative, leading to a depression of responses.

Further research has shown that this prediction error signal is sensitive to many different aspects of the reward stimuli. For example, prediction errors are specific to the context in which the association has been learned, (Nakahara et al., 2004), and firing rates in response to the cue are proportional to both the magnitude (Bayer and Glimcher, 2005) and the probability (Fiorillo et al., 2003) of the reinforcer. Furthermore, to maintain the reward sensitivity over a large range of values, the gain is adjusted to the variance of the reward value (Tobler et al., 2005). These and many other studies support the hypothesis that the DA neurons in the VTA signal aspects of reward prediction error (Schultz and Dickinson, 2000).

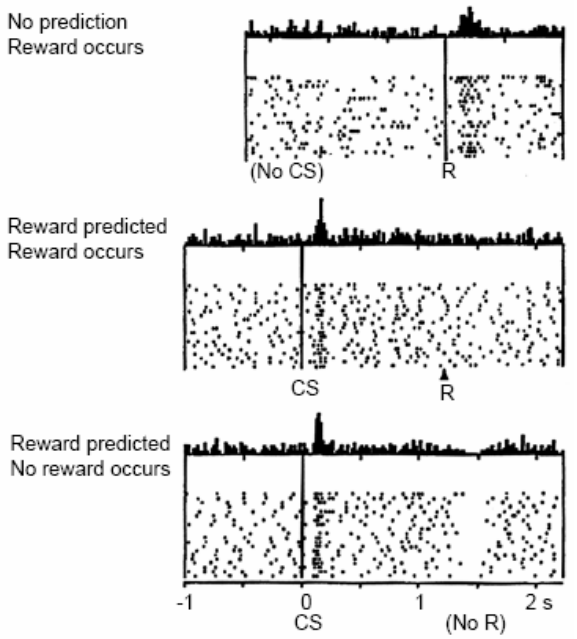

Figure 1.1. Dopamine firing reflects prediction errors. Changes in dopamine neuron firing reflect the prediction errors of appetitive events. For each panel, the top graph represents the accumulated spike count per time bin, and each dotted line in represents one recording session, where each dot is a spike. CS = conditioned stimulus, R = primary reward (juice) **Top.** Before learning, the juice drop is not predicted, resulting in a positive prediction error, and increased firing in response to the juice. **Middle.** After learning, the CS predicts the reward, and the dopamine neurons increase firing rate in response to the reward-predicting CS, but not to the predicted reward. **Bottom.** When after learning the reward-predicted CS is presented, but the reward is omitted, this results in a negative prediction error and suppressed firing of the DA neurons at the time the reward should have occurred. (From (Schultz et al., 1997)).

Inspired by the results from these animal experiments, fMRI studies have subsequently shown that in humans the VTA also responds to the difference between expected and actual rewards (D'Ardenne et al., 2008). More frequently and prominently, however, these studies found reward prediction error responses in the ventral striatum, e.g. in the context of primary food rewards (McClure et al., 2003;O'Doherty et al., 2003;Pagnoni et al., 2002;Rodriguez et al., 2006) to money (Abler et al., 2006;Hare et al., 2008;Seymour et al., 2007;Yacubian et al., 2007) and even attractive faces (Bray and O'Doherty, 2007). These findings can be explained by noting that (i) the ventral striatum is a primary target of dopaminergic projections from the VTA (Joel and Weiskopf, 2000), and (ii) the BOLD signal reflects

postsynaptic field potentials (and thus input to an area) more strongly than firing rate (and thus output from an area) (Logothetis et al. 2001). Thus, it is likely that ventral striatal BOLD activity in relation to reward prediction errors is partially, although probably not completely, a downstream reflection of reward prediction error responses of DA neurons in VTA. A recent study directly linked dopamine and reward-seeking behaviour in humans (Pessiglione et al., 2006). Behaviourally, subjects treated with DA agonist levodopa were more likely to choose more rewarding actions than subjects on DA antagonist haloperidol. Furthermore, the striatal prediction error response to rewarding stimuli as observed in previous studies was modulated by dopaminergic drugs. The degree of this modulation determined how much the subject's behaviour was affected by the drugs.

The human and animal studies described above strongly support the hypothesis that DA neurons encode reward prediction error and that ventral striatal activity as measured with fMRI reflects these prediction errors. This does not mean, however, that processing in the ventral striatum is limited to reward-based learning and is not involved in other forms of associative learning. Conditioning and associative learning has long been dominated by animal research, and training animals to do behavioural tasks is inherently reward-based; one cannot simply ask a monkey or rat to press a button, they have to be rewarded to do so. As a result there has been a bias towards reward-based tasks, and comparatively little interest in investigating associative learning and striatal processing of affectively neutral and task-irrelevant stimuli. This has changed somewhat since the advent of human neuroimaging, and a few recent studies have given some hints as to the nature of striatal processing of affectively neutral stimuli. fMRI results showed that the ventral striatum responds to nonrewarding, unexpected stimuli (Zink et al., 2003) proportional to the salience of the stimulus (Zink et al., 2006). Furthermore, activity in the ventral striatum increases in responses to cues predicting a novel, affectively neutral stimulus and to novel stimuli per se (Wittmann et al., 2007;Wittmann et al., 2008).

These results suggest that rather than just coding rewards or reward prediction errors, the striatum may have a more general role in processing salient and unexpected events. One of the proposed functions of this striatal response is to reallocate resources to unexpected stimuli in both reward and non-reward contexts (Zink et al., 2006), which will be discussed in the next section.

1.1.1.2 Affectively neutral prediction errors

In a pioneering fMRI study of prediction error signals regarding causal associative learning for affectively neutral stimuli subjects learned the relationships between various cues (fictitious drugs) and outcomes (fictitious syndromes) (Fletcher et al., 2001). At the start of the experiment, when the environment was still unpredictable, activity in the right dorsolateral prefrontal cortex (DLPFC) and the putamen was high, and decreased as the associations were being learned. Furthermore, activity in the DLPFC was higher on trials with unexpected compared to expected outcomes. The authors suggest this response pattern reflected 'cognitive' prediction errors because the learned associations were not reward-based, but nevertheless task-relevant.

In a subsequent study, Corlett et al. investigated prediction error responses for two well-known conditioning effects: backwards blocking and unovershadowing (see **Table 1.1**; (Corlett et al., 2004)). Subjects were instructed to predict allergic reactions of a fictitious patient in response to certain food items. In a first stage, two food items A and X were presented together, followed by an allergic response. Two other food items, B and Y were also paired and followed by an allergic response. In a second stage A was presented alone, followed by an allergic response, and B was also presented alone, but not followed by an allergic response. Because A could fully account for the allergic response observed with the AX compound, X became disassociated from the allergic response, which is known as backwards blocking. Conversely, because B could not explain the allergic response observed in the BY compound, Y alone will now explain the allergic response, so Y has been 'unovershadowed' by B. If these processes indeed occur, then X followed by an allergic response and Y not followed by an allergic response will be more surprising, i.e. have a larger prediction error, than vice versa. Indeed Corlett et al. showed exactly such a prediction error response in the striatum as well as in the same part of the right DLPFC (Corlett et al., 2004) as was demonstrated previously (Fletcher et al., 2001;Turner et al., 2004).

Table 1.1. Overview of conditioning paradigms. Overview of conditioning paradigms described in the main text, showing the different training stages, the testing phase and the conditioned response during testing

	Stage 1	Stage 2	Stage 3	Test	Response
Classical Conditioning	A → reinforcer			A	+
Blocking	A → reinforcer	AX → reinforcer		X	0
Control	B → nothing	BY → reinforcer		Y	+
Backwards Blocking	AX → reinforcer	A → reinforcer		X	0
Unover-Shadowing	BY → reinforcer	B → nothing		Y	+
Preventative Learning	A → reinforcer	AB → nothing		B	-
Superlearning	A → reinforcer	AB → nothing	BC → reinforcer	C	++
Higher order conditioning	A → Reinforcer	B→A→reinforcer		B	+

Given that the ventral striatum responds to novelty, (Wittmann et al., 2007) and that in the study by Fletcher et al. unpredictability and novelty were correlated, it is possible that the observed striatal and DLPFC prediction error responses were simply due to novelty of the outcomes. However, Turner et al. showed that this was not the case in a very carefully controlled study employing the phenomena of preventative and superlearning (see **Table 1.1**; (Turner et al., 2004)). Here, in a first phase a stimulus gets associated with an outcome (A+). In the next phase a novel stimulus, is combined with A, but not followed by the outcome (AB-). This generates a negative prediction error, and B acquires a strong negative causal potential. In a third phase, the B is paired with a third stimulus and followed by the outcome (BC+). This generates a strong positive prediction error, and C acquires a very strong positive causal potential. When contrasting this with a standard blocking paradigm, in which the stimuli were presented the same number of times, both superlearning and preventative learning events, which only differed from control events in terms of the

size of the prediction errors, but not in terms of novelty, showed increased activity in the DLPFC as well as in the striatum.

These studies demonstrated that prediction errors play a role in learning associations that do not involve reward prediction, and that both the striatum and the prefrontal cortex are involved in processing these 'cognitive' prediction errors. However, although unrelated to reward, these prediction errors are still relevant to the subject in the sense that their task is to make accurate predictions and making correct predictions is rewarding in itself. Therefore, one cannot claim that learning here is entirely unrelated to any form of reward. Importantly, thought, the reward of being correct is entirely orthogonal to the presence or absence of the allergy outcome. This is in contrast with reward based studies where presence of the outcome (i.e. the reward) always results in a positive reward prediction error, and absence of the outcome in a negative reward prediction error. Unlike reward-based prediction errors, the prediction errors observed in the studies discussed above are independent of whether the error was in the positive (unexpected presence of outcome) or negative (unexpected absence of outcome) direction. In other words, the prediction errors were unsigned, which might be explained by the fact that the actual outcome itself is not relevant, only how surprising this outcome was.

Summarising, in circumstances where the only relevant measure is *how much* surprise the outcome engenders, i.e. for affectively neutral contexts, the dorsolateral prefrontal cortex and the striatum encode a sign-independent prediction error ((Corlett et al., 2004;Fletcher et al., 2001;Turner et al., 2004). It should be emphasised again though that the learned associations here are still relevant to the task. However, in recent years it has been suggested that coding of prediction errors is at the heart of every cognitive process, including low-level sensory perception (Friston et al., 2006;Rao and Ballard, 1999). The next section will discuss this theory of predictive coding as a basic mode of brain function.

1.1.2 Predictive coding

Why would the brain aim to predict irrelevant events and stimuli? The theoretical notion of predictive coding proposes that the brain has two primary objectives: *inference* about the causes of sensory input and *learning* the relationship between the inputs and the causes. This is achieved by constructing a *generative* model of how

causes in the world elicit sensory inputs; given some sensory inputs, this model can be inverted to *recognise* the causes of this input. In this scheme, each level of the processing hierarchy receives bottom-up sensory input from the level below and top-down predictions from the level above (Garrido et al., 2009b). Prediction error, i.e. the difference between the true and estimated probability distribution of the causes, is minimised at all levels of the hierarchy by adjusting connection strengths through synaptic plasticity (Friston, 2005a)).

One of the most basic and robust paradigms to demonstrate neuronal responses to unexpected stimuli is the oddball paradigm. Here, presentation of an oddball stimulus in a sequence of standard stimuli elicits a negative potential as measured using EEG, which is known as the mismatch negativity (MMN) potential. The MMN is observed in all sensory domains (auditory (Baldeweg, 2006), visual (Stagg et al., 2004); somatosensory (Akatsuka et al., 2007)) and can be understood in light of a predictive coding framework (Garrido et al., 2009b). Prediction errors are minimised by adjusting connection strengths through synaptic plasticity upon repeated presentation of the stimuli. These adjustments are reflected neurophysiologically by the disappearance of the MMN (Baldeweg et al., 2006;Friston, 2005a), which is elicited again when an oddball is presented. This adjustment is also reflected by repetition suppression in the visual domain, as observed in fMRI (Summerfield et al., 2008). Here, the likelihood of stimulus repetition was manipulated and repetition suppression was reduced in response to improbable compared to probable repetitions.

There is increasing evidence that perceptual learning is just one of many processes that can be explained in a predictive coding framework (Friston et al., 2006;Garrido et al., 2009a;Rao and Ballard, 1999) and is also at the heart of higher level processing: In an fMRI study, an expectation to see faces was induced by asking subjects to report whether presented stimuli were faces or not. Forward connectivity between face sensitive visual areas (FFA) to the frontal cortex to was modulated by the prediction errors. Incorrect predictions increased FFA → prefrontal connectivity, whereas correct predictions increased prefrontal → FFA connectivity (Summerfield and Koechlin, 2008). Other studies have shown predictive coding like mechanisms for sensory integration (Bays and Wolpert, 2007;Blakemore et al., 1998), predictive attenuation of tactile stimulation (Bays et al., 2006), and even for social interactions (Shergill et al., 2003;Wolpert et al., 2003). These findings support the notion that the

fundamental function of the brain could be to encode an implicit and probabilistic model of the environment (Friston et al., 2006).

1.1.2.1 Functions of neuronal prediction error signals

The effect that various forms of prediction errors have on neuronal functioning depends on several factors. Firstly, the specificity and scope of the projections of the prediction error encoding neurons determine whether the signal is broadcasted widely, or selectively affects a small group of neurons. For example, cholinergic and dopaminergic projections from the nucleus basalis and VTA have widespread connections to the cortex (Lewis, 1991). Resulting global error messages could then selectively affect neurons involved in processing information at the same time as the prediction error signal via postsynaptic neurons that act as coincidence detectors. Alternatively, the prediction error signal could be relayed only to a selected group of neurons, directly affecting behavioural reactions.

Secondly, the way in which the neurons affect postsynaptic signalling might differ. The postsynaptic effects may be very short-lived and directly affect immediate behaviour or attention, or they might control storage of predictions by inducing short-term or long-term changes in synaptic strengths. Such learning-dependent plasticity will be discussed in the next section.

1.1.3 Plasticity during associative learning

1.1.3.1 Synaptic plasticity during associative learning

Already in 1949, Donald Hebb suggested that changes in connectivity are central to the physiological implementation of association learning (Hebb, 1949). The previous section described how the brain actively generates predictions of sensory signals based on an internal model of the world and compares those expectations to the actual incoming information. Predictive coding theories propose that prediction errors are minimised by adjusting the synaptic efficacies or connections strengths between different levels of the processing hierarchy.

Brain connectivity is defined by three key properties: i) the current strength of a connection ii) the change in the strength of this connection over time, and iii) how this change is controlled. These three aspects correspond to distinct

neurophysiological mechanisms. For glutamatergic synapses, the main excitatory synapses in the brain, connection strength depends on the number and state of AMPA receptors (Malinow & Malenka 2002). Changes in synaptic strength, i.e. plasticity, is regulated by NMDA-dependent mechanisms modulating the number of AMPA receptors expressed at the synapse (Genoux and Montgomery, 2007). Because of its unique molecular properties the NMDA receptor can function as a 'coincidence detector' of afferent and efferent activity, and as such initiate synaptic plasticity. Presynaptic transmitter release concomitant with postsynaptic depolarisation allows a calcium influx through the NMDA receptors, which triggers trafficking as well as phosphorylation of glutamatergic AMPA receptors. These properties make NMDA receptors ideally suited for associative learning processes that involve concomitant activity in different (e.g. sensory) areas of the brain. Indeed NMDA-dependent mechanisms have been found to play a key role plasticity in learning and memory processes in the brain (e.g. see (Genoux and Montgomery, 2007;Gu, 2002;Ji et al., 2005;Morris, 1989;Tye et al., 2008)).

Finally, synaptic plasticity itself is influenced by modulatory transmitters like dopamine, serotonin and acetylcholine, mainly through changes in NMDA receptor function ((Gu, 2002), see **Figure 1.1**). For example, dopamine (DA) and acetylcholine (ACh) regulate the trafficking, insertion and endocytosis of NMDA receptors into the cell membrane. As such, cholinergic mechanisms strongly modulate NDMA dependent LTP and LTD in visual cortex (Brocher et al., 1992;Kirkwood et al., 1999) and auditory cortex (Metherate and Hsieh, 2003). The phosphorylation of the NMDA receptors, which determines the conductance properties, is modulated by DA and serotonin (5HT) receptors (Jiao et al., 2007;Salazar-Colocho et al., 2007;Wolf et al., 2003). In summary, excitatory brain connectivity is determined by (i) AMPA receptors, expressing synaptic strength, (ii) NMDA receptors controlling synaptic strength, and (iii) modulatory transmitters regulating this control (see **Figure 1.2**).

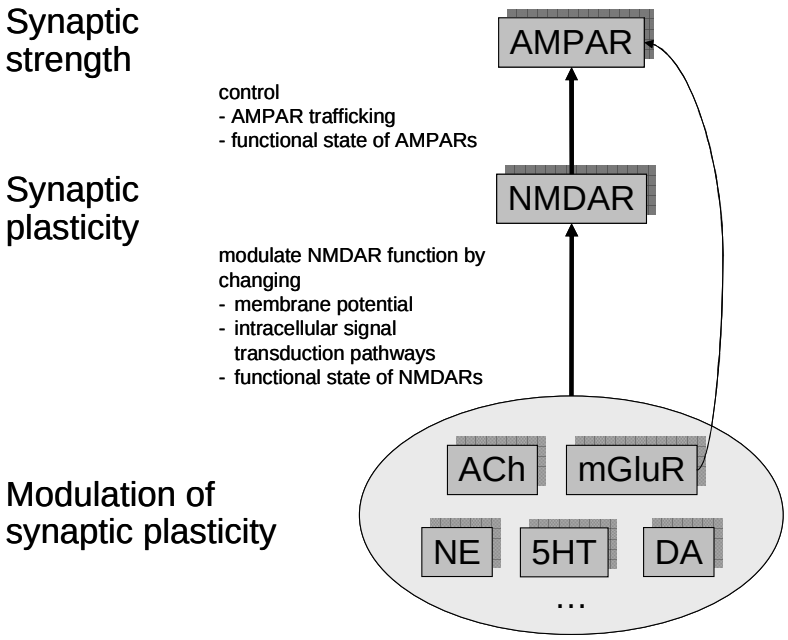

Figure 1.2. Modulation of synaptic plasticity. Modulation of synaptic plasticity of excitatory, glutamatergic synapses by several modulatory neurotransmitters via NDMA signalling. (Adapted from (Stephan et al., 2006)). NMDAR = NMDA receptor, AMPAR = AMPA receptor, ACh = acetylcholine, NE = norepinephrine, 5HT = serotonin, DA = dopamine.

1.1.3.2 Associative learning induced plasticity in the sensory cortex

The previous section described the cellular mechanisms of synaptic plasticity that underlie associative learning processes. In line with behavioural observations, Friston suggested that when 'value-dependent modulation is extended to the inputs of neural value systems themselves, initially neutral cues can acquire value' (Friston et al., 1994). Here, discriminative conditioned responses are accompanied by value-dependent plasticity of receptive fields, as reflected in the selective augmentation of unit responses to valuable sensory cues. Electrophysiological and fMRI measures of activity in the sensory cortices indeed show such changes in CS processing as the cues become associated with affective outcomes: Neurons in the superior colliculus and auditory cortex are known to have a frequency to which they fire preferentially, known as the best frequency (Weinberger, 2007). Electrophysiological studies in rats and big brown bats have shown a 'centripetal' best frequency shift towards the frequency of the conditioned tone when the tone is paired with an aversive stimulus such as a painful electric shock (Ma and Suga, 2005;Weinberger, 2004). Similarly, in humans, BOLD responses in the auditory cortex are enhanced in response to tones

paired with aversive outcomes (Thiel et al., 2002b;Thiel et al., 2002a). In contrast, however, Morris et al. observed a decreased response to CS+ (CS regularly and consistently paired with an outcome) compared to CS- (CS never or rarely paired with an outcome) stimuli in a similar fear conditioning study using PET (Morris et al., 1998). At first this seems at odds with the electrophysiology results. However, the electrophysiologically observed best frequency shift results in a small and narrowly tuned increase, which is accompanied by a decreased response in the surrounding frequencies. It is conceivable that this decrease swamps the small augmentation, so that a focal narrowing of the tuning curve as found in animals, actually results in a regional deactivation because of the courser spatial resolution. The direction of the observed changes notwithstanding, these results all point in the direction of conditioning dependent plasticity in CS processing in the auditory cortex.

Similar findings were reported in the visual cortex, where BOLD responses in the visual cortex increased to a visual stimulus that had been associated with a noxious outcome (Carlsson et al., 2006). Such changes in perceptual processing are not limited to aversive outcomes; Seitz et al. showed increased stimulus sensitivity when visual orientation gratings were paired with liquid rewards, even when subjects were unaware of the visual stimulus (Seitz et al., 2009).

In conclusion, recent studies show plasticity in the sensory cortices for CS processing in the context of aversive or appetitive conditioning. It is unclear as to whether such changes also occur when associations between neutral stimuli are learned, a question that will be addressed in **Chapter 3**.

1.1.3.3 The role of ACh

ACh is one of the most important modulators of synaptic plasticity in the context of associative learning in the perceptual domain. For example, the role of nucleus basalis dependent ACh release in auditory cortex plasticity has been extensively studied in aversive conditioning experiments in animals, showing receptive field plasticity and behavioural memory formation to be mainly dependent on muscarinic receptors (Weinberger, 2007). Also in humans, the enhanced BOLD response to tones associated with shocks is abolished upon administration of the ACh antagonist scopolamine (Thiel et al., 2002b;Thiel et al., 2002a). Furthermore, admistration of

scopolamine abolished both behaviourally observed repetition priming as well as the associated repetition suppression in the visual cortex (Thiel et al., 2002c). Finally, interaction of ACh and NMDA dependent mechanisms appear to be crucial for long term consolidation of conditioning-induced synaptic plasticity (Ji et al., 2005).

These findings are starting to elucidate the mechanisms by which cortical plasticity is regulated during conditioning and other forms of associative learning. The aim of the work presented in this thesis was to create biologically plausible models of connectivity to asses connectivity changes during associative learning. In the future these models could be used to directly test the influences of the neuromodulators discussed here on connection strengths and their learning dependent-changes, i.e. plasticity (see **Chapter 6**)

1.2 Models of associative learning

This section starts with a general discussion of model-based analysis methods, and then review how two classes of computational models can provide a framework to investigate aspects of associative learning at the behavioural and physiological level.

1.2.1 Model-based analysis methods

A model is a representation that contains the essential structure of some event or process in the real world. In psychology and biology, 'models' are often informal, consisting of boxes with arrows between them, such as protein synthesis cascade models or the working memory model (Baddeley and Della, 1996). In mathematics and physics, ‘models’ are more formal, in the form of equations that putatively underlie observed processes. In recent years systems neuroscience research has seen a strong increase in the use of formal modelling techniques concerning reinforcement learning (e.g. (Gläscher and Büchel, 2005;Pessiglione et al., 2006;Seymour et al., 2004)), decision making (e.g. (Beck et al., 2008;Behrens et al., 2007)), and brain connectivity assessed by fMRI and electrophysiology (Chen et al., 2008;Kiebel et al., 2008;Stephan et al., 2007a).

Indeed both types of models adhere to the definition of a model as ‘a representation of a process in the real world.’ An often overlooked fact is that any inferential analysis of data essentially tests a model. Because these models are not always

explicitly described as such, there is a perceived distinction between model-based and other research. A good example of such unrecognised model testing is classical statistical inference. This is essentially a test of very simple models of the world: A one-sample t-test on the effect of variable A on some measure B effectively compares a model of the world in which A affects the generation of B, to a model of the world where A does not affect B. Here, the model of the world is simplified to the extent that it only contains A to predict B; any other factors that might affect B are considered 'noise', or errors in the prediction of the model. In summary, the recent increase in 'model-based' approaches in neuroimaging does not break away from classical analysis methods, but merely constitutes an evolution towards more explicitly defined and complex models.

1.2.1.1 Model complexity

Models are by definition simplifications; if a model included every aspect of the real world, it would no longer be a model, but it would be the world itself. Simplification allows us to distil and probe those aspects of the world that we are interested in. A good model has the right balance between complexity and fit: on the one hand, it should be simple enough not to be misled by noise, i.e. experiment specific variations that do not generalise across experiments. On the other hand, if a model cannot account well for important aspects of the observations, then it may not be complex enough. In other words, in order to create a good model, one has to make simplifying assumptions about the causes of events in the real world, and all other causes that are not represented by the model will contribute to the noise, or error in the model predictions.

Critically, whether a given model is worse or better than another model depends on the phenomenon that is to be explained. In other words, there is no single 'true' model of the world, just different models with different (levels of) simplifications that can account for different situations and test different hypotheses. This underlines the importance of selecting the optimal model for a given question and data set. A generic statistical framework for handling this challenge is Bayesian model selection (Penny et al. 2004; Stephan et al. 2009). This approach was used for each study contained by this thesis and will be described in detail in **Chapter 2**.

1.2.2 Reinforcement learning models

1.2.2.1 RW model

The first and most influential theoretical model of associative learning was proposed by Rescorla & Wagner in the early seventies (Rescorla and Wagner, 1972), and is based on Pavlovian classical conditioning learning (see **Section 1.1**). The Rescorla-Wagner (RW) model describes this learning process in terms of the strength of the association formed between the CS and the US. The basic principle is that the change in associative strength ΔV_t at a particular trial *t* is directly proportional to the size of the prediction error:

$$\Delta V_t = \alpha\beta(\lambda_t - \sum V_t) \qquad (1.1)$$

Here, the prediction error $(\lambda_t - \sum V_t)$ is the difference between the actual outcome λ_t on a trial *t*, and the predicted outcome $\sum V_t$, which is based on the summed prediction across all cues present. α and β are learning constants that determine the weight of the incoming information (i.e. the prediction error) relative to the information accumulated on previous trials (i.e. the prediction). On each trial, the change in associative strength ΔV_t is added to the current associative strength V_t, such that the associative strength reflects the cumulative information from all observed trials.

$$V_{t+1} = V_t + \Delta V_t \qquad (1.2)$$

The values of the learning constants α and β reflect properties, such as salience and motivational value, of the CS and US stimuli, respectively. Note that each stimulus has an associated learning constant, but that in most paradigms there is only one type of CS and one type of US, or the properties of the CSs and USs are assumed to be constant across stimulus types. As a result, the product of these two constants is another constant, which is the overall learning rate. However, when for example the salience of two different cues is very different, one can use different associated learning rates (cf. **Chapter 3**).

During learning, when the outcome is incompletely predicted, the prediction error will be positive and the associative strength will increase. On the next trial, the prediction error will be slightly smaller, and thus the increase in associative strength will also be slightly smaller. Once the outcome is completely predicted, the difference between outcome and prediction is zero, and the associative strength will remain unchanged. This happens when learning has reached an asymptote. However, when an outcome is predicted but omitted, the prediction error is negative, and the associative strength will be reduced. Because the learning constants determine the size of the update, they will determine how quickly the asymptote is reached, but not the actual level of the asymptote. The level of the asymptote is determined by the conditioning schedule, and reflects the average value of λ.

Thus, when the association between the CS and the US is constant but probabilistic, say the CS predicts a US reward with an 80% probability, then association strength V will asymptote at 0.8 (given that λ is either 1 or 0). Note that the learning rate determines both the speed at which the asymptote is reached and the size of the fluctuations around the asymptote after learning is complete.

1.2.2.2 Delta rule model: Connectionist implementation of the RW model

Although the RW model describes how associations form between (internal representations of) CS and US stimuli, it does not provide a mechanistic explanation of the learning process. McLaren proposed a neural network that could compute the prediction error using a negative feedback assembly, as a potential mechanism underlying error-based learning (McLaren, 1989). In this model, the weight of the connections between the signal (CS input) and response (prediction based on the CS) are controlled by a facilitatory unit F that itself is controlled by direct excitation by the reinforcer and a negative feedback from the response unit (see **Figure 1.3**). Thus, the prediction based on the presentation of a CS is constantly updated depending on the prediction error. The response unit perfectly reproduces the predictions from the RW model. The modulation of connection strengths by prediction errors as proposed by this neural network is at the heart of the work presented in this thesis where we aimed to investigate the role of prediction errors in modulating connection strengths during different types of associative learning.

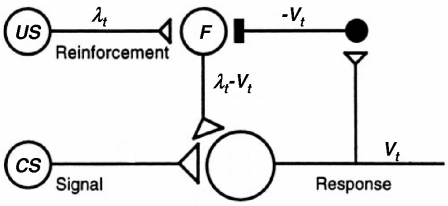

Figure 1.3. Neuronal network implementing prediction error signalling. Schematic simplification of the original neuronal assembly proposed by McLaren for how prediction errors shape neuronal connections encoding associations (after (McLaren, 1989)). The response unit delivers a prediction V_t driven by the signal unit which received the CS input. This driving input to the response unit is modulated by the facilitatory unit F. Activity in F itself is controlled by the excitatory output from the US unit, carrying information about the reinforcer λ and an inhibitory output from the response unit, carrying the negative of the prediction V_t. By summation of these two inputs, activity in F reflects the prediction error.

1.2.2.3 Determining the learning rate

Determining the learning rate is one of the most contentious issues with regard to the RW and related models. There are two main approaches to determine the learning rate or any other constant parameter. The first approach is to decide on a learning rate based on previous literature and knowledge about the particular task and stimuli at hand (e.g. see (O'Doherty et al., 2003;Petrovic et al., 2008)). For example, when the differences between a CS+ and CS- are small, learning is likely to be quite slow, so a small learning rate is appropriate (e.g. (Petrovic et al., 2008)). This approach is the easiest to implement but theoretically problematic, as one can never be sure that the used learning rate is indeed the optimal one. A more principled approach is to fit the parameters to the data (see **Chapter 2** for a more detailed discussion of model fitting using Expectation Maximisation algorithms, and see (Behrens et al., 2007) for an example). Once the optimal learning rate has been established, one can then test whether the predictions of the model explain a significant proportion of the variance. From a theoretical point of view this is the optimal approach, but it is sometimes difficult to implement. For example when the data is very noisy (e.g. reaction time data, or BOLD responses) it might be difficult to find the optimal values, and estimation might get stuck in a local maximum. One hybrid approach often used to determine the learning rate is to define a range of plausible learning rates, and do a

stepwise analysis of the proportion of variance explained by models with different learning rates within this range (e.g. see (den Ouden et al., 2009;Gläscher and Büchel, 2005;Seymour et al., 2004)).

1.2.2.4 Limitations, extensions and alternatives of the RW model

The RW model was a great step forward in formalising thought and theories about associative learning, and can explain a wide range of behaviourally observed learning phenomena, including classical conditioning and extinction, as well as blocking, preventative learning and superlearning (described in **Table 1.1**). However, the RW model is not appropriate in all situations. There are a number of behaviourally observed phenomena which cannot be explained by the reinforcement learning models discussed above. For example, the RW model predicts that the history of conditioning has no influence on its present status; only the current association value is important. However, experiments have shown that a previously conditioned stimulus actually needs fewer trials to reach the same level of conditioning, i.e. there is facilitated acquisition after extinction. Related to this is the observation that when a CS is not presented for a while after a CS-US association has been extinguished, that there is partial recovery from extinction, and furthermore that exposure to the US alone can reinstate the CS-US association. Another example that RW models cannot deal with is higher-order conditioning (see **Table 1.1**). When a novel cue is paired with a conditioned excitor, in the absence of a reinforcer, the RW model would predict that the novel cue becomes a conditioned inhibitor, but instead it becomes a conditioned excitor as well. This pairing effect also works when the CSs are paired before the CS-US association is learned (sensory preconditioning; discussed in **Chapter 3**). Finally, the RW model predicts that presenting a novel stimulus without a US should not affect later conditioning. However, latent inhibition (or CS-preexposure effect) is a well established observation that after exposure to a CS without the US, conditioning to the CS is retarded. Below we will discuss two alternative models that are originally based on the RW model and can model a number of the phenomena discussed above.

1.2.2.4.1 Pearce & Hall attentional model

A crucial property of the standard RW model is fact that the learning rate is constant. In other words, the balance between current observations compared to predictions

based on past observations is unchanging. One would expect, however, that once an association has been learned, the learning rate decreases. Attentional theories of associative learning are one class of models that do allow for changes in the learning rates. For example the model proposed by Pearce & Hall, the prediction error does not directly impact on the associative strength, but rather controls how much attention is allocated to the next stimulus (Pearce and Hall, 1980). The associability is determined by the following equation

$$\alpha_t = \gamma\left|\lambda_{t-1} - \sum V_{t-1}^{NET}\right| + (1-\gamma)\alpha_{t-1} \tag{1.3}$$

Here the associability of the CS with the outcome determined by the absolute value of the prediction error at the previous trial, and the associability at the previous trial. The relative contributions of these terms are determined by the parameter γ. The underlying idea is that the more attention is paid to a stimulus, the more readily it will become associated with the reinforcer, and once an association is learned and the US fully predicted, the associability is low. This model can explain certain phenomena like latent inhibition, where the RW model fails. However, one of the drawbacks of the Pearce-Hall model is that because only the absolute value of the prediction error is used, the association of the CS and US can only ever increase. This results in the rather inelegant solution of having to invoke a second learning process in the form of modelling the inhibitory association of the same CS, i.e. the CS-noUS association. The net prediction is then the sum of the associative and 'anti-association' of the CS and US:

$$\sum V_t^{NET} = \sum V_t - \sum \overline{V}_t \tag{1.4}$$

A further problem is that introducing the parameter γ to determine weighting of current and past information only shifts the problem caused by the constant learning rate in the RW model. One still needs to somehow determine the constant value of the parameter γ that determines the shape of the learning curve.

1.2.2.4.2 Temporal difference learning models

Because the RW model can only capture between trial effects, it cannot account for within-trial effects such as second-order conditioning or sensitivity to stimulus

timing (cf. **Figure 1.1**). Temporal difference (TD) learning models are basically a real-time extension of the RW models that allow one to model *within trial* timing effects, and are therefore particularly suited to explain the DA prediction error signal. The main assumption of TD models is that the prediction V_t should be interpreted as the total, discounted sum of future rewards expected from time *t* to the end of the trial (Sutton and Barto, 1990). The strategy is to use a vector *x* that describes the presence of a sensory cue for each time bin in a trial, and another vector *w* that carries weights of predicted rewards in those time bins. On each trial[1], the predicted value V_t is the linear product of the weights w_i and the presence or absence of the CS, as encoded by the stimulus vector $x_{i,t}$:

$$V_t = \sum_i w_i x_{i,t} \tag{1.5}$$

At each time point the prediction error δ_t can be calculated as the difference between the prediction and the reward at that time point plus all future rewards until the end of the trial. At the end of each trial, the weights are updated depending on the prediction errors and the learning rate:

$$\Delta w_i = \alpha \sum_t x_{i,t} \delta_t \tag{1.6}$$

As the outcomes become associated with the cues, the weights shift from the outcome to the cue. Thus, after learning, the cues, which itself are unpredicted, will elicit a positive prediction error because they predict a positive summed reward until the end of the trial. TD models can explain higher-order conditioning because the reward associated with a particular cue simply shifts to a preceding cue (Seymour et al., 2004). The behaviour of the TD models depend strongly on the number and the size the time bins into which a trial is divided. A TD model with only one time bin is exactly the same as an RW model.

1.2.2.5 Concluding remarks

The RW model formed the basis of a wide range of error-based learning models. This model can explain a number of observed learning phenomena, including

[1] Note that for TD models, t denotes time within rather than between trials

classical conditioning and blocking, although it fails to predict a number of others, such as higher-order conditioning and latent inhibition. Extensions of the model including TD models and the Pearce-Hall model can account for some of these. That does not mean, however, that the RW model is invalid, and that these alternatives are necessarily 'better' models. Whenever choosing a model it is important to keep in mind its limitations and properties and select the appropriate model for a particular dataset. For example, when one is not interested in modelling within trial learning effects, because the time resolution of the data does not allow for this (cf. **Chapter 3**), there is no point using a TD model, because it simply reduces to an RW model. In other words, it is important to choose a model that can capture the phenomena one is interested in, but is not unnecessarily complex.

1.2.3 Bayesian ideal observer models

Bayesian methods for reinforcement learning can be traced back to the 1960s, but until recently they have only been used very sporadically. Part of the reason for this is that non-Bayesian approaches described in the previous section tend to be easier to implement and work with. The main difference between the classical RW[2] vs. Bayesian learning models is that the former use point estimates of the associations, whereas Bayesian methods are based on using full posterior distributions, considering not only the probabilities of the associations, but also the uncertainty about theses probabilities. In other words, the mean of the posterior distribution reflects the current estimate of the association strength, and the variance of this distribution reflects the uncertainty about this estimate. This principled approach to balancing previous knowledge and current information is formalised by Bayes Theorem:

$$p(\vartheta \mid y) \propto p(y \mid \vartheta)\, p(\vartheta) \tag{1.6}$$

Here, $p(\vartheta \mid y)$ is the posterior belief about states ϑ (e.g. trial-by-trial estimates of associations) given the data y, based on the optimal combination of the likelihood $p(y \mid \vartheta)$ and the prior belief in the model parameters $p(\vartheta)$. Bayes theorem ensures

[2] Here we will refer to the term RW model for the sake of simplicity, but note that this can be replaced by any prediction error-based learning model with a fixed learning rate, including TD and Q-value learning models.

optimal integration of current beliefs based on past information, and current information by weighing the prior and the likelihood by their respective precisions (see **Figure 1.4**).

The application of Bayesian models to neuroscience and cognition research focused initially on domains of perception and sensorimotor integration (Bays et al., 2006;Kording et al., 2007;Rao and Ballard, 1999;Whiteley and Sahani, 2008;Wolpert et al., 1995). More recently, however, the Bayesian approach has been applied to modelling learning processes, for which it is exquisitely suited, because it specifies how to optimally update beliefs in light of new evidence. Thus, application of Bayesian techniques has been extended to investigate a range of learning processes, from sensorimotor learning (Bestmann et al., 2008) to conditioning and reinforcement learning (Behrens et al., 2007;Courville et al., 2006;Daw et al., 2005;Yoshida et al., 2008).

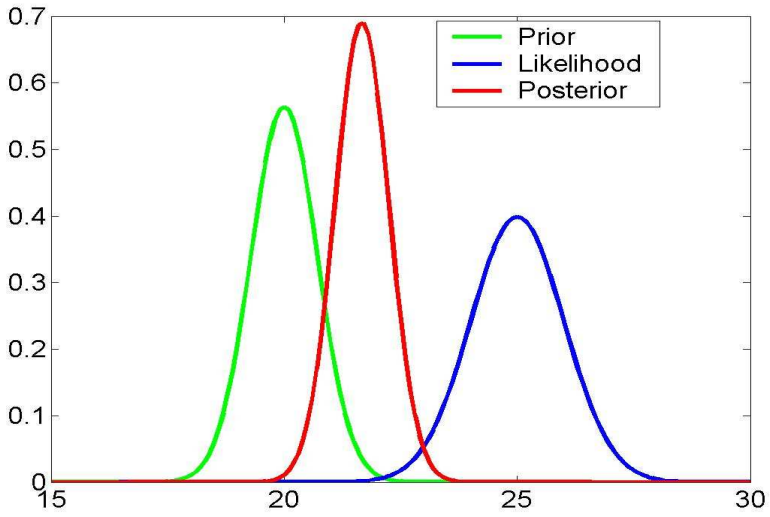

Figure 1.4. Bayes' Rule. Bayes' Rule optimally combines prior knowledge (green) with new data (blue) by weighing their respective uncertainties, to express a posterior belief (red). In this example the precision of the prior is greater than that of the likelihood, therefore the mean of the posterior distribution is closer to the prior than to the likelihood. Note that when two distributions are combined in a Bayesian fashion the resulting distribution always has a higher precision than each of the source distributions.

1.2.3.1 Prediction errors in Bayesian models

Although in the update equations of a Bayesian learning model there is no mathematical equivalent of the prediction error from RW models, surprise does play an important role. Like in RW models, a surprising event (i.e. a large distance between prior and likelihood) will result in a large shift of the posterior belief, whereas a fully predicted event (i.e. likelihood and prior fully overlap) results in no shift at all. Thus, Bayesian surprise is reflected by the distance between the posterior and prior distributions of the beliefs (Itti and Baldi, 2005).

1.2.3.2 Limitations of Bayesian ideal observer models

Bayesian ideal observer models are ideal in the sense that they follow Bayes rule, and thus generate predictions by optimally combining all available information. The catch here is the term 'available'. A Bayesian model, like any other model, is ideal only within the context of a given model structure. Consider for example the following situation: In a paradigm in which association probabilities fluctuate sinusoidally around an average of 0.5, then after a large number of trials the Bayesian observer will have a very strong belief that the association is at 0.5, and from that point onward, the belief will effectively be stuck at 0.5, failing to capture the sinusoidal shape of the fluctuations. This situation generalises to any belief after a very large number of observations; the prior will have a very small variance, because every new observation reduces the variance. Compared to an RW model, this is effectively equivalent to a learning rate that over time asymptotes to zero. One solution for this is to reset the priors every time the probability has changed (e.g. (Bestmann et al., 2008)), but this assumes that the model 'knows' when to reset itself. Thus, this Bayesian is clearly not a very good in this circumstance, even though it does, at every trial, optimally combine the available information.

A more generic solution to this particular problem is to introduce an extra parameter that estimates the variability of the environment, such that when variability is high, e.g. every time the contingencies have changed, the variance of the prior is increased. This is the Bayesian equivalent of introduction a changing learning rate (see (Behrens et al., 2007), and **Chapter 4**). With this extension, information about the variability of the environment will be optimally combined to weigh prior beliefs and new information.

In summary, ‘ideal’ observers are ideal in the sense that they optimally combine all the information that they can represent. One of the arguments against Bayesian models is that any phenomenon can then be explained in an "ideal" fashion given the right model structure. This illustrates how the often-posed question whether humans are ideal Bayesian observers is not a very informative one. Rather, the question should be what behaviour and brain processes can be modelled by which particular model.

1.2.4 RW vs Bayesian models

The most crucial distinction between standard reinforcement learning models and Bayesian update models, is that the former represent the state variables of the model (e.g. the contingencies) as point estimators, whereas the latter provides a full posterior distribution, where the mean reflects the belief about the estimated contingency and the variance denotes the uncertainty about this belief[3]. This difference has a number of important consequences.

First of all, because the history of observed events is now recorded by two quantities, the mean and the variance of the parameters, Bayesian models have a ‘memory’ of the trial history: For a point estimate model (with unchanging learning rate), each new trial carries the same weight with respect to updating the estimate. It does not matter whether it has just observed 10 or 100 instances of an 80% pairing of a stimulus and outcome; estimates on the 11^{th} or 101^{st} trial changed with equal ease. In contrast in the Bayesian models this weight is proportional to the number of trials that have been observed; the distribution of the belief about the association for a Bayesian observer model will be much narrower after 100 trials than after 10 trials (cf. **Figure 1.4**).

This leads to the second difference, which is that for the RW models, the learning rate, which determines how much a belief is updated based on new information, is determined by the researcher, or is estimated from the measured data (e.g. reaction times). For the Bayesian models, however, the balance between old and new information is determined by the structure of the model and the observed series of events.

[3] Note that one could also include higher order modes, such as the skewness and kurtosis, but most models make Gaussian assumptions about the distributions of state variables.

In summary, RW models are simpler to implement and can model a large number of observed conditioning and reinforcement learning phenomena. However, recently there has been a shift towards the use of Bayesian observer models which can take into account uncertainty about the estimates in a principled way. As such, Bayesian ideal observer models are part of a more general theoretical framework which will be discussed in more detail in **Chapter 2**.

1.3 Summary of experimental work

Chapters 3-5 describe the empirical research conducted for this thesis to investigate the role of prediction errors and connectivity changes in associative learning, using a combination of formal learning models and DCM for fMRI. A range of associative learning tasks was used with increasing behavioural relevance of the associative relationships. In **Chapter 3** a combination of bilinear DCMs and a RW learning model is used to investigate changes in connectivity between sensory cortices as unchanging and task-irrelevant relationships between affectively neutral sensory stimuli are being learned. **Chapter 4** describes a study with dynamically changing CS-US associations in which subjects responded to affectively neutral targets. These target stimuli were chosen to be preferentially processed by different visual areas, which made it possible to assess the stimulus specificity of the prediction errors in visual cortex and striatum. Nonlinear DCM was combined with a Bayesian observer model that could optimally account for the changing probabilistic associations to explore the role of the striatum in gating sensorimotor connections. Finally in **Chapter 5** a pre-existing dataset is used for a nonlinear DCM study to investigate the role of the amygdala modulating corticostriatal connections in a fear conditioning paradigm.

Chapter 2

Methods

Abstract

The first section of this chapter describes the bilinear and nonlinear implementations of DCM for measured BOLD time series data. Both implementations of DCM were used in this thesis to test different hypotheses related to changes in connectivity during associative learning. The second part describes Bayesian model selection, which was used to select the optimal model from sets of candidate models accounting for fMRI and behavioural data. In this chapter I will give a general description for both analysis tools; specific details for each implementation will be described in the results chapters **(Chapters 3 - 5).**

2.1 Dynamic Causal Modelling for fMRI

2.1.1 Connectivity models and DCM

Over the past decades, the predominant approach in cognitive neuroimaging has been to investigate functional specialisation of brain regions, based on the assumption that there is local specialisation of information processing. Although there is no doubt that such local specificity exists, this approach is clearly incomplete given that locally processed information must be integrated at some stage. The aim of connectivity models is to investigate experimentally induced changes in cortical pathways rather than cortical areas to look at functional integration rather than specialisation. There are two conceptually distinct approaches to connectivity analysis of fMRI timeseries. On the one hand there are models of *functional* connectivity, defined as statistical dependencies between spatially remote

neurophysiological events, and on the other hand there are models of *effective* connectivity, which is defined as the influence one neuronal system exerts over another (Friston et al., 1993a).

Models of functional connectivity describe statistical dependencies among the data, thereby providing a characterisation of the functional interactions, or context-dependent coherence between different timeseries (Friston et al., 1993b). A simple example of functional connectivity analysis would consist of generating brain maps that represent voxel-wise correlations of local activity with the timeseries of a particular 'seed' region of interest (ROI). These maps can then be compared e.g. under different experimental conditions. Functional connectivity analyses can be a useful exploratory device, because to characterise a functional network they do not rely on strong a priori assumptions about the underlying mechanisms. At the same time, this is also one of the main limitations, because this lack of specificity prevents one from testing detailed hypotheses about the connectivity of the underlying neural network. Moreover, the lack of causal (directed) effects precludes explanations at a mechanistic level as to the nature of the interactions between the different temporally correlated areas.

Effective connectivity explicitly models the influence that one neuronal system exerts over another, rather than just their statistical dependencies. It is congruent with the notion of 'synaptic efficacy' between individual neurons or neuronal populations. The aim of models of effective connectivity is to explain regional effects as detected by for example a voxel-wise GLM analysis, in terms of interregional connectivity. Unlike the exploratory approach of functional connectivity methods, models of effective connectivity are mechanistic models, which require a clear neuranatomical delineation of the areas that are modelled, as well as a clear hypothesis about how these areas affect each other. Any type of effective connectivity analysis involves two steps: A first step in which the anatomical areas will form the nodes in the model are selected, and a second step in which the relationships between the nodes are described.

DCM is a model of effective connectivity to make inferences about the neural processes underlying a measured time series. Like other models of effective connectivity DCM allows one to investigate the mechanisms underlying the observed

dynamics of complex neural systems in terms of connection strengths and their context-specific modulation. DCM views the brain as a deterministic nonlinear dynamic system that is subject to external inputs in the form of experimental manipulations, and that produces outputs (Friston et al., 2003). DCM assumes that brain responses are driven by changes in the input, rather than by endogenous noise or "innovations", as is assumed by other models of effective connectivity (e.g. SEM and autoregressive models (McIntosh and Gonzales-Lima, 1994)). These experimental inputs can enter the system and elicit responses in one of two ways: Firstly they can enter as driving inputs, for example the presence of an auditory stimulus would directly affect an auditory cortex node. The second way inputs influence the system is more indirectly, by modulating the coupling between the nodes, for example effects of attention to visual input could modulate a top-down connection from frontal to primary visual areas.

Moreover, while other models, such as structural equation modelling (SEM, (McIntosh and Gonzales-Lima, 1994)) operate at the level of the measured signal, implicitly assuming an identity mapping between neuronal processes and (hemodynamic) measurements, DCM accounts for the nonlinear coupling between the measured hemodynamic responses and the underlying neural activity of interest (Penny et al., 2004b). In DCM the generative model consists of two levels. Causal effects in a cognitive system are modelled at the underlying (hidden) neuronal level using a parsimonious but plausible neurobiological model. The modelled neuronal population dynamics are then transformed into area-specific BOLD signals using a biologically informed hemodynamic forward model (Stephan et al., 2007d). The general idea is to model interactions among cortical regions by optimising the parameters of a reasonably realistic underlying neuronal model such that the modelled BOLD timeseries matches the experimentally measured timeseries as closely as possible.

In a further diversion from conventional models of effective connectivity, which model instantaneous effects, DCM is a time series model in which the temporal evolution of the neural state vector is a function of the current state as well as the inputs and the system architecture. By modelling how *state changes* in a given node depend on the *current state* of any other node it is influenced by, DCM allows one to

determine the *directional* influence between areas, equivalent to causal relationships in the sense of control theory.

The most important application of DCM is that it can be used to answer questions about the modulation of effective connectivity. In the original formulation bilinear DCMs allow one to infer that a particular experimental manipulation (e.g. a cognitive set, learning, or a pharmacological manipulation) modulate a pathway, rather than a cortical region (see **section 2.1.2** for details and **Chapter 3 for implementation**). Bilinear DCMs preclude an important aspect of neuronal interactions, namely how the connection between two neuronal populations is enabled or gated by activity in other populations. Therefore in a recent extension to DCM the bilinear approach is extended such that now the effective connectivity between nodes is not only modulated by external inputs but also by activity in other nodes. In these nonlinear (second order) DCMs the modulation of connections can thus be assigned to a particular neuronal population in the system (see **Section 2.1.3** for details on nonlinear DCMs and **Chapters 4 and 5** for implementation).

2.1.2 Bilinear DCM

In the bilinear formulation of DCM, the states of multiple interacting brain regions are modelled as a set of coupled bilinear differential equations (Friston et al., 2003). The neuronal states, which represent the neuronal population activity of the modelled brain regions, change in time according to the system's connectivity and experimentally controlled inputs *u*. These inputs can enter the model in two different ways; they can either elicit responses through direct influences on specific regions (“driving inputs", e.g. sensory inputs) or they can change the strength of connections between regions (“modulatory inputs", e.g. task effects or learning). The hidden neural dynamics are modelled by the following bilinear differential equation:

$$\frac{dz}{dt} = \left(A + \sum_{i=1}^{m} u_j B^{(i)} \right) z + Cu \qquad (2.1)$$

Here, z is the state vector (with each state variable representing the population activity of one region in the model), t is continuous time, and u_i is the i-th input to the modelled system. In this state equation, the A matrix represents the fixed (endogenous) strength of connections between regions and the $B^{(1)}...B^{(m)}$ matrices

represent the modulation of these connections by (exogenous) inputs, as an additive change. Finally, the *C* matrix represents the influence of exogenous inputs on each area. Note that DCM allows one to make inferences about changes in effective connections between areas, which do not necessarily correspond to direct anatomical connections but may be via intermediary regions.

The hidden neuronal dynamics described by **Equation 2.1** are transformed to predicted BOLD responses by a hemodynamic forward model (Friston et al., 2003). Given measured BOLD responses, this model can be inverted, using a Bayesian estimation scheme, to obtain maximum a posteriori estimates of the parameters in **Equation 2.1** (Friston et al., 2003). Finally, the probability of the data given a particular model can be estimated by integrating out the dependency of the joint density on the model parameters. This estimate, known as the model evidence or marginal likelihood, can be used to compare the goodness of competing models, and thus to test different hypotheses of the underlying neural network generating the measured responses. This procedure, known as Bayesian model selection, will be described in detail in **Section 2.2**.

2.1.3 2nd Order DCM

As mentioned above, effective connectivity represents the influence of one neuronal population on another, corresponding to the notion of 'synaptic efficacy'. The bilinear term in DCM models the effect of experimental manipulations on connections between neuronal populations. However, this framework precludes an analysis with respect to the neuronal source of these modulations, and thus omits an important aspect of neuronal interactions: how connections between two neuronal populations are gated by activity in other populations. These gating mechanisms are known to be mediated through interactions between synaptic inputs and are central to learning and attentional modulation. Therefore, a nonlinear extension to DCM has been developed which allows one to assign the modulation of network interactions to specific neuronal populations (Stephan et al., 2008).

In the original bilinear implementation of DCM for fMRI, the temporal change of the neuronal state vector is modelled using a bilinear approximation that governs the dynamics of the system. In nonlinear DCMs, this bilinear approximation is extended

to second order such that the hidden neural dynamics are modelled by the following equation:

$$\frac{dx}{dt} = \left(A + \sum_{i=1}^{m} u_i B^{(i)} + \sum_{j=1}^{n} x_j D^{(j)} \right) x + Cu \qquad (2.2)$$

Here, **Equation 2.1** is extended with the $D^{(j)}$ matrices, which encode how connection strengths are modulated or gated by activity in area *j* (for details, see (Stephan et al., 2008)). In this thesis, the second order extension of DCM was employed to investigate the influence of the putamen and amygdala on network connectivity during associative audio-visual learning and during fear acquisition (see **Chapters 4 and 5,** respectively).

2.1.4 Parametric modulation of connections

In most DCM studies to date, the inputs constituting the bilinear modulations of the network interactions are context dependent, such as attention or task instructions (see e.g. (Grol et al., 2007;Stephan et al., 2007b)). These inputs are simply either 'on' or 'off', and are conceptually related to the main effect regressors in classical GLM analyses using mass univariate models (e.g. SPM, see www.fil.ion.ucl.ac.uk/spm). However, it is also possible to assess modulators that change parametrically, for example dosage of a pharmacological intervention or the temporal evolution of learning. The form of these parametric modulatory inputs corresponds to that of so-called "parametrically modulated regressors" in a classical GLM analysis. In the bilinear and nonlinear DCMs described in **Chapters 3 and 4**, association strengths were estimated using two different learning models, and entered the DCMs as direct or indirect modulatory input.

2.2 Bayesian Inference and Model Comparison

2.2.1 Within subject Bayesian inference

In order to estimate the parameters of the forward model the DCMs are inverted using a Bayesian inversion approach. The inversion of a particular DCM involves approximation of the posterior probability of the parameters of the model,

$p(\vartheta \mid y,m)$, given a particular dataset and model. The posterior is proportional to the product of the prior probability and the likelihood $p(y \mid \vartheta,m)$, following Bayes' rule:

$$p(\vartheta \mid y,m) \propto p(y \mid \vartheta,m)p(\vartheta \mid m) \tag{2.3}$$

The aim of the model inversion is to find the parameters ϑ that maximise the posterior probability, using empirical priors for the hemodynamic parameters and conservative shrinkage priors for the neural coupling parameters. The parameter estimation scheme uses a Gauss-Newton gradient descent embedded in an Expectation-Maximisation (EM) algorithm, which is described in detail elsewhere (Friston et al., 2002b). In short, in the E-step the posterior mean and covariance are updated, and during the M-step the hyperparameters of the noise covariance matrix are updated. The posterior densities of the neural parameters can then be used to make inferences about the effective connectivity, for example to test how certain one can be that a particular parameter exceeds a particular threshold (usually zero).

However, one typically needs to compare alternative models representing different hypotheses about the connectivity of the network, and select the optimal model before making inferences about the model parameters. The optimal model is the model that has the greatest probability of representing the underlying system that produced the measured dataset; this probability is known as the model evidence $p(y \mid m)$, and accounts for both model fit and model complexity (Pitt and Myung, 2002). The model evidence can be found by integrating out any dependencies on the estimated model parameters ϑ from **Equation 2.3**.

$$p(y \mid m) = \int p(y \mid \vartheta,m)p(\vartheta \mid m)d\vartheta \tag{2.4}$$

Unfortunately in most cases this integral cannot be solved analytically, and is difficult to compute numerically (for one exception see **Section 2.2.1.1**). Therefore, instead of evaluating the integral in **Equation 2.4**, approximations to the model evidence must be used (Friston et al., 2007;Penny et al., 2004a). Commonly used approximations are the Akaike Information Criterion (AIC, (Akaike, 1974)), the Bayesian Information Criterion (BIC, (Schwarz, 1978)) and the negative free energy (F). All of these methods approach the true model evidence by optimising a bound on the integral in **Equation 2.4**. The difference between these approximations is how

they treat the trade-off of model accuracy (model fit) and model complexity. It is important to penalise the model evidence for complexity, because model fit will increase monotonically with complexity, but at some point the mode will start fitting noise, thereby reducing the generalisability of the model. Therefore the optimal model provides the best balance between model fit and complexity. In all three approximations of the model evidence (AIC, BIC and F), the accuracy is the expected log likelihood of the data under an approximating posterior density on the parameters $q(\vartheta)$, which is optimised iteratively.

For the AIC and BIC the approximation to the log model evidence for model *m* can be given by

$$\begin{aligned} BIC &= accuracy(m) - \frac{p}{2}\log n \\ AIC &= accuracy(m) - p \end{aligned} \qquad (2.5)$$

where *p* is the number of parameters and *n* the log of the number of observations (e.g. scans). When looking at the complexity terms, it becomes clear that the BIC pays a heavier penalty than the AIC (when $\log n/2 > 1$, i.e. when $n > 8$). Therefore the BIC will favour simpler models whereas the AIC will be biased towards more complex models. Because this could lead to contradictory results, generally models are only considered to be different in fit when the results from the AIC and BIC concur. The AIC and BIC were used in **Chapter 3** to select the optimal DCM out of a number of models.

The AIC and BIC are useful and easy approximations of the model evidence. However, because the complexity term scales linearly with the number of parameters, they both fail to account for redundant parameterisation; when adding a parameter that has identical effects than another parameter on predicting measurements, the complexity terms of both AIC and BIC would increase even though the 'true' complexity would not change. Often models will have (partial) dependencies amongst parameters, and in this case the AIC/BIC approach will overestimate model complexity. In the negative free energy approach the complexity is the Kullbach-Leibler divergence between the approximating posterior and the prior density, reflecting the amount of information obtained about the model parameters from the data.

$$F = accuracy(m) - KL[q(\vartheta), p(\vartheta \mid m)] \quad (2.6)$$

Under the Laplace approximation (i.e. assuming that the conditional density is a multivariate Gaussian), the complexity term splits into three terms

$$KL[q(\vartheta), p(\vartheta \mid m)] = \tfrac{1}{2}|C_\vartheta| - \tfrac{1}{2}|C_{\vartheta|y}| + \tfrac{1}{2}(\mu_{\vartheta|y} - \mu_\vartheta)^T C_\vartheta^{-1}(\mu_{\vartheta|y} - \mu_\vartheta) \quad (2.7)$$

where $|C_\vartheta|$ and $|C_{\vartheta|y}|$ are the determinants of the prior and posterior covariance matrices, and μ_ϑ and $\mu_{\vartheta|y}$ the prior and posterior means. The first term increases the complexity with the *effective* degrees of freedom, taking into account dependencies amongst parameters, i.e. additional redundant parameters do not increase the complexity. The second term decreases the penalty with the degree of independence that the parameters have *a posteriori*, because in a good model the parameters are as precise and independent as possible The third term shows that the complexity penalty increases the larger the difference between the prior and posterior means, i.e. when suboptimal priors are used. Thus the free energy F is often a better approach to approximate the log evidence than the AIC/BIC, and was used in **Chapters 4 and 5** to decide between different competing DCM models.

For any of these model evidence approximations, to determine how strong the evidence is in favour of one model, one can simply compute the model evidence ratio of the two models, also known as the Bayes Factor, or equivalently the difference between the log evidences. If the difference in the log evidences is greater than about three (i.e. the Bayes factor is larger than 20), this is considered as strong evidence in favour of a particular model (Raftery, 1995).

2.2.1.1 A special case: Bayesian GLM for response speed data

Linear Gaussian models constitute a rare case where there is an analytical solution to the model evidence, instead of having to resort to an approximation as described above. This analytical solution, under the assumption that the data and design matrix are Gaussian, can be viewed as the Bayesian version of a GLM. Like in a classical GLM, the model to be tested is described by a design matrix which includes

regressor for all explanatory variables of the model. Using flat priors, one can then calculate the model evidence for different models as a function of the model fit (sum of squared residuals) and the complexity (the number of regressors in the design matrix). This linear model has the following form:

$$\begin{aligned} &Y = X\beta + \varepsilon \Rightarrow \\ &p(Y \mid \beta, \sigma^2, m) = (2\pi)^{-d/2} \sigma^2 \exp\left(-\frac{(Y - X\beta)^T (Y - X\beta)}{2\sigma^2} \right) \end{aligned} \tag{2.8}$$

This is the probability of the data Y (e.g. response speeds), given the design matrix X, parameters β, and normally distributed errors $\varepsilon \sim N(0, \sigma^2)$. In order to compute the model evidence, or probability of the data given the model, the parameters and hyperparameter σ need to be integrated out:

$$\begin{aligned} p(Y|m) &= \iint p(Y, \beta, \sigma, m) d\beta d\sigma \\ &= (2\pi)^{(r-d)/2} \left| X^T X \right|^{-r/2} \Gamma(d - r - 1) \left(\lambda / 2 \right)^{r+1-d} \end{aligned} \tag{2.9}$$

Here r is the number of parameters in the design matrix, d is the number of data-points and

$$\lambda = Y^T \left(I - X \left(X^T X \right)^{-1} X^T \right) Y \tag{2.10}$$

is the sum of squared residuals. Therefore the log model evidence is

$$\begin{aligned} \log(p(Y \mid m)) = &\frac{r-d}{2} \log(2\pi) - r/2 \log\left(\left| X^T X \right| \right) \\ &+ \log(\Gamma(d - r - 1)) + (r + 1 - d) \log\left(\lambda / 2 \right) \end{aligned} \tag{2.11}$$

Model evidences can then be compared either at the level of the individual subject, using the Bayes Factor, or at the group level using one of the group Bayesian model selection tools described below (see **Section 2.2.2**).

2.2.2 Group level Bayesian inference

2.2.2.1 Fixed effects analysis: Group Bayes Factor

The Bayes Factor approach described in **Section 2.2.1** is suitable for comparing different models for one particular dataset, for example from a single subject. However, one will often want to make inferences about a group of subjects, and select the model that best explains multiple datasets. Assuming that the datasets of different subjects are independent, one can simply multiply the Bayes factors for each model across all subjects, known as the Group Bayes Factor (GBF, e.g. see (Stephan et al., 2007c)). This fixed effects analysis will be used in **Chapter 3** to select the optimal DCM from three competing models.

2.2.2.2 Bayesian random effects analysis

Combining BMS results from a group of subjects relying on fixed-effects analyses such as described above assumes that all subjects' data are generated by the same model. As a result they fail to account for group heterogeneity and are vulnerable to outliers. Stephan et al. developed a novel random effects Bayesian analysis framework to cope with these shortcomings (Stephan et al., 2009). This method allows one to quantify the probability that a particular model generated the data for any randomly selected subject, relative to other models. They showed that this approach of calculating a conditional density of model probabilities given the model evidence for individual subjects, is superior both to using the group Bayes factor (as described above), and to applying frequentist tests to the log evidences. This superiority was especially evident in the case of large intersubject heterogeneity and in the case of outliers (Stephan et al., 2009).

Instead of assuming that the data were generated by the same model for all subjects, this approach computes a density from which models are sampled to generate subject-specific data. In other words, it searches for the conditional estimates of model probabilities $r = [r_1,...,r_K]$, that generate indicator variables, $m_n = [m_{n1},...,m_{nK}]$, where $m_{nk} \in \{0,1\}$, and for any given $n \in \{1,...,N\}$, $\sum_{k=1}^{K} m_{nk} = 1$. These indicator variables prescribe the model for the *n*-th subject, where $p(m_{nk}) = r_k$. Since the model probabilities r follow a Dirichlet distribution $p(r \mid \alpha)$, the

conditional expectations $\langle r_k \rangle_q = \alpha/(\alpha_1 + ... + \alpha_K)$ encode the expected probability that the *k*-th model will be selected for a randomly selected subject. For details about the hierarchical Bayesian model that is inverted to obtain the Dirichlet parameters α of the posterior $p(r \mid y;\alpha)$ see (Stephan et al., 2009).

After optimisation of α, the posterior can be used for group level Bayesian model comparison, where the results can be reported in several different ways. Firstly one can simply report the estimate of $\alpha = [\alpha_1,...,\alpha_K]$ for each of the models, where $\alpha_k - 1$ represents the number of subjects in which m_k generated the observed data. One can also use the posterior $p(r \mid y;\alpha)$ to compute the expected multinomial parameters $\langle r_k \rangle$, and thus calculate the expected likelihood of obtaining a particular model for any randomly selected subject

$$\langle r_k \rangle_q = \alpha/(\alpha_1 + ... + \alpha_K) \qquad (2.12)$$

Either of these models can then be used to rank the models at the group level. A third option is to use $p(r \mid y;\alpha)$ to quantify an *exceedance probability*, defined as the belief that a particular model *k* is better than any other of the models tested given the group data. In this thesis we have adopted the approach to report both the Dirichlet parameters α and the exceedance probabilities when discussing the results of our analyses. We have used this novel random effects Bayesian model selection tool to show that behavioural data were well described by a sophisticated Bayesian learning model in **Chapter 4**, and to select the optimal DCM in **Chapters 4 and 5**.

2.2.2.3 Model space partitioning in Bayesian random effects analysis

The Bayesian random effects analysis can be used not only to compare specific models, but also to test for differences between parts of 'model space', provided that each subspace contains the same number of models, i.e. the design is fully factorial (Stephan et al., 2009). For example, one may wish to compare the effect of adding or leaving out a particular connection, irrespective of any other differences between the tested models. This model space partitioning can be regarded as the Bayesian equivalent of a main effects analysis in a classical ANOVA.

This analysis exploits the agglomerative property of the Dirichlet distribution: Once the parameters α_k for all *K* models have been estimated, for each subset of models a new Dirichlet density can be calculated simply by adding the α_k for all models belonging to that particular subset. The resulting Dirichlet can be used to compare subsets of models in exactly the same way as for individual models, for example to calculate the exceedance probabilities.

Model space partitioning will be used in **Chapter 5** to compare the addition and removal of endogenous connections and second order modulations in a 2x3 factorial design.

Chapter 3

A Dual Role for Prediction Error in Associative Learning

Abstract

In this fMRI experiment subjects implicitly learned the association between the presence (or absence) of a task-irrelevant visual stimulus and the presence (or absence) of a task-irrelevant auditory stimulus. Using a Rescorla-Wagner (RW) model to describe the evolution of fMRI responses during learning, it was shown that BOLD activity in primary visual cortex (V1) and the ventral striatum covaried with prediction errors, or surprising events, regardless whether this surprise concerned the unexpected presence of a visual stimulus or its unexpected absence. Furthermore, DCM analyses suggest that this response in V1 is due to prediction error dependent changes in connections from the auditory cortex (A1). To our knowledge, this is the first empirical evidence that (i) V1 responds to prediction errors engendered by audio-visual probabilistic relations, and, more generally, that (ii) prediction errors during associative learning drive synaptic plasticity. This finding has important implications for our understanding of general mechanisms of perceptual learning and inference in the human brain.

3.1 Introduction

Among the fundaments of adaptive behaviour is the ability to predict future events. This ability is crucial to functions ranging from sensory processing to decision making. In psychology and neuroscience, prediction has been studied most extensively in the context of Pavlovian and instrumental conditioning tasks, which measure how organisms anticipate (and act on) affectively significant events such as food delivery or electric shocks. A recent series of functional neuroimaging studies has investigated the neurophysiological basis of prediction and learning in humans. Using Pavlovian and instrumental conditioning tasks, these studies have identified several areas where BOLD signals correlate with trial-wise estimates from formal learning models like TD learning (Sutton and Barto, 1998) or the Rescorla-Wagner model (RW) (Rescorla and Wagner, 1972). In particular, BOLD activity in areas including the striatum and the dorsolateral prefrontal cortex (key dopaminergic targets) has been shown to covary with both *predictions* and *prediction errors* (Corlett et al., 2004;Fletcher et al., 2001;Gläscher and Büchel, 2005;Jensen et al., 2007;McClure et al., 2003;O'Doherty et al., 2004;Pessiglione et al., 2006;Seymour et al., 2004;Turner et al., 2004).

In all of these previous studies, the learned associations had direct relevance for behaviour, either because they were linked to rewarding or punishing outcomes (e.g. (McClure et al., 2003;O'Doherty et al., 2004;Seymour et al., 2004) or because subjects received feedback on their performance (Aron et al., 2004;Corlett et al., 2004;Fletcher et al., 2001;Turner et al., 2004). In contrast, it is unclear whether incidental learning of stimulus-stimulus associations, i.e. learning of associations that are irrelevant for current behavioural goals, draws upon the same neuronal mechanisms. A paradigm that shows that these types of associations are learned is ‘sensory preconditioning’. Here, in a first stage, the subject is exposed to behaviourally meaningless CS_1-CS_2 associations and, in a second stage, to CS_1-US pairings. In a third and final stage, the presentation of a CS_2 alone generates a conditioned response, indicating that the subject must have learned the initial CS_1-CS_2 association (Brogden, 1939;Gewirtz and Davis, 2000).

In this study we used a factorial design that extended the first stage of a classical sensory preconditioning paradigm. Healthy volunteers performed an audio-visual

target detection task, while being exposed to a stream of concurrent audio-visual "distractor" stimuli (**Figure 3.1**). These stimuli possessed statistical regularities, which enabled prediction of the visual distractor from the preceding auditory cue (**Figure 3.2**). Critically, however, these statistical associations were completely irrelevant to the target detection task. Any learning of these associations would therefore be of an incidental (task-unrelated) nature and, in the absence of behavioural responses to the learned associations, could only be inferred neurophysiologically. This paradigm capitalised on previous work by McIntosh et al. (McIntosh et al., 1998) who used positron emission tomography (PET) to show that learning of associations between sensory stimuli was reflected by activity in early visual cortex. However, the use of PET permitted only a simple conditioning scheme and precluded a full investigation of dynamic changes in the brain's representation of the learned association. Here, we employed a more refined conditioning scheme and used fMRI to study learning-dependent changes in brain activity over time. Additionally, we assessed learning-dependent changes in effective connectivity between auditory and visual cortex using DCM.

Using a 4-factorial design (cf. **Table 3.1**), this study characterised learning in terms of the temporal evolution (learning; factor 1) of both brain activity and interregional connectivity in response to a visual stimulus whose presence or absence (V^+ vs. V^-; factor 2) was predicted in 2 contexts, established by 2 types of auditory conditioning stimuli (CS+ vs. CS-; factor 3), each of which could be present or absent on each trial (A^+ vs. A^-; factor 4). In other words, in contrast to classical sensory preconditioning paradigms, we could not only investigate differential learning, depending on CS type but could also assess whether the consequences of an absent CS were learned. It should be noted that both the CS+ and CS- contexts (or blocks) were balanced in terms of stimuli; the *a priori* probabilities of the auditory CS and of the visual stimulus occurring on a given trial were always 50%. Critically, the task was not related to these auditory and visual stimuli; subjects performed a target-detection task on unrelated stimuli that were presented sporadically.

One of the features of our factorial paradigm is that on half the trials the auditory CS is absent. This necessitates an additional cue that marks the beginning of each trial, which was a visual trial onset (TO) cue. In other words, learning of stimulus associations in this paradigm has two components, one related to the auditory CS and

another related to the visual TO cue. As a consequence, any model of the learning process must be able to formulate how a net prediction is computed from the associative strengths of the two cue components. We chose the RW model since it is the simplest and most generic model of associative learning that accounts for cue interactions (see **Discussion Section 3.4** for details). The RW-model has been validated extensively, using behavioural data from both humans and animals and can account for many aspects of associative learning (Pearce and Bouton, 2001;Schultz and Dickinson, 2000). In our study, the trial-wise associative strength predicted by the RW model was used to construct regressors for a voxel-wise general linear model of fMRI data and modulatory inputs for DCMs (Friston et al., 2003) of the effective connectivity between auditory and visual areas. Specifically, we addressed the following two questions:

1. In the absence of any behavioural responses to the audio-visual stimulus associations, can we obtain neurophysiological evidence that the brain learns these associations? Specifically, can we find brain regions whose activity correlates with learning[4] as predicted by a generic model of associative learning (i.e., the RW model)? Candidate areas included early visual cortex and the striatum. Furthermore, do these areas show a response profile across cue-outcome combinations that reflects a match between prediction and outcome or rather a prediction error response?

2. Since the predictive auditory cue temporally precedes the visual outcome, learning should modify neuronal activity in early visual cortex in response to auditory cues. Can these putative learning-related changes in visual cortex activity be explained by changes in the effective connectivity from auditory to visual cortex (cf., (McIntosh et al., 1998;McLaren et al., 1989)? Specifically, do these changes conform to changes in associative strength under a RW model of learning?

Before describing our experiment, two important issues should be highlighted. First, the goal of this fMRI study was not to pinpoint the exact mathematical form of incidental learning by comparing different models of associative learning. Instead,

[4] Throughout the chapter, we will use the colloquial term "learning curve" to denote the vector of predicted associative strength over time, i.e. ϕ_t^j in Equation 1.

we used the simplest model of associative learning that could accommodate our paradigm. In the Discussion (**Section 3.4**), we argue why the RW model can be considered an appropriate *a priori* learning model for our particular paradigm, relative to other models of associative learning. Second, it is important to note that *within* a given experimental condition the predicted outcomes and prediction errors are perfectly anti-correlated when mean-corrected (see **Appendix A** for details). This means they cannot be distinguished as alternative predictors of observed brain responses. However, with our factorial design one can analyse the pattern of parameter estimates *across* experimental conditions, contrasting expected and unexpected cue-outcome combinations. This enabled us to distinguish, voxel by voxel, brain responses that reflected a match between predicted and actual trial outcomes from responses that encode prediction error or surprise.

3.2 Methods & Statistical analysis

3.2.1 Subjects

Sixteen healthy volunteers, 25.3 ±3.3 years of age, (mean age ± SD, 8 female) participated in the study. The subjects had no history of psychiatric or neurological disorders. Written informed consent was obtained from all volunteers prior to the study, which was approved by the National Hospital for Neurology and Neurosurgery Ethics Committee.

3.2.2 Experimental Design – fMRI

The central idea of this study was to present subjects with "distractor" stimuli that were linked by predictive associations: two auditory stimuli served as conditioning stimuli (CS) and differentially predicted whether or not a visual stimulus would follow. Critically, the volunteers performed an unrelated detection task on separate auditory and visual targets; for this task, the predictive relationships between the distractor stimuli were completely irrelevant. Stimuli were presented using Cogent2000 (www.vislab.ucl.ac.uk/Cogent/index.html). An initial sound matching task and the subsequent learning study (4 x 10 min) were all completed inside the scanner. Subjects were debriefed with a post-scan questionnaire to assess whether they had learned the experimental contingencies.

3.2.2.1 Sound matching

Preceding the learning experiment, subjects had to match the two CS (450 Hz and 1000 Hz) and the auditory target stimulus (white noise burst) for perceived loudness. Stimuli were presented sequentially and dichotically. Subjects adapted the volume of the 1000 Hz tone to the 450 Hz tone until they perceived them to be of equal loudness. This procedure was repeated eight times and the results averaged. Subsequently, subjects matched the perceived loudness of the white noise burst to the pure tones, each repeated four times. The adapted volumes, as a percentage of the volume of the low tone were 94.0 ± 6.2 % (mean ± SD) for the high tone, and 104 ± 4.9 % for the white noise burst.

3.2.2.2 Differential conditioning

During the experiment, subjects were exposed to alternating blocks of trials in which one of two auditory conditioning stimuli (high and low tone) predicted the presence (CS+) or omission (CS-) of a subsequent visual stimulus with a fixed probability of 80% (**Figure 3.1** and **Table 3.1**). On each trial, a CS was presented (A+) with 50% probability. On 50% of all trials, a visual stimulus was present (V+). Every trial was preceded by a visual trial-onset (TO) cue.

Our paradigm thus used a 4-factor design with the following factors for each trial: i) CS context (CS+ vs. CS-), ii) CS presence (A+ vs. A-), iii) visual outcome (V+ vs. V-) and iv) learning (or time). We used a mixed design in which CS type was blocked, whereas the presentation of the CS and visual outcome were randomized (event-related) within blocks. CS+ and CS- blocks were completely balanced so that in each block of 10 trials, five CS and five outcome stimuli were presented. Within each subject, the auditory CS+ and CS- and their probabilistic relation to subsequent visual stimuli were fixed throughout the experiment. The assignment of tones to the two CSs was counterbalanced across subjects, i.e. in half the subjects the high tone served as CS+ (and the low tone as CS-), and vice versa the other half of the subjects. Each of the foursessions consisted of 20 blocks of 10 trials, interspersed with periods of rest (12 s), in which subjects fixated on a fixation cross. Blocks and sessions were balanced across and within subjects.

3.2.2.3 Target detection task

To ensure continuous attention to auditory and visual targets per se (but not their statistical associations), subjects performed a concurrent target detection task. The target stimuli were randomly interspersed between trials and consisted of either a white noise burst or a circle. Target stimuli occurred on average once per block (at most twice). In total, 40 auditory and 40 visual target stimuli were presented, randomised within conditions and sessions.

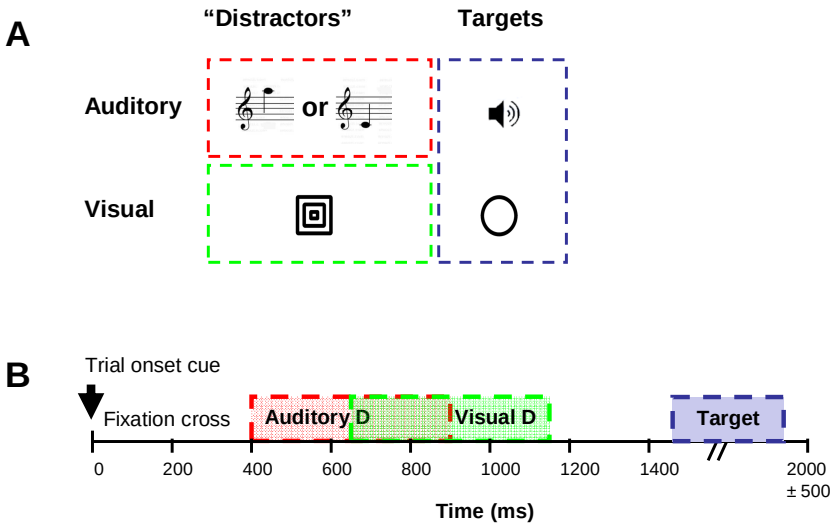

Figure 3.1. Experimental design. **A)** Stimuli presented during the experiment. The 'distractor' stimuli, whose associations were being learned incidentally, comprised two auditory conditioning stimuli (CS) corresponding to high- and low-frequency tones and one visual unconditioned stimulus (US) consisting of three concentric squares. The target stimuli, to which the subjects responded, comprised a white noise burst and a circle. **B)** Temporal sequence of a single trial. Both the CS and US could be either presented or omitted. The average trial duration was 2 seconds. The trial onset (TO) cue was a small central dot (100 ms); the auditory CS was presented for 500 ms, starting 400 ms after trial onset. The visual stimulus was presented 750 ms after trial onset, also for 500 ms. The inter-trial interval (ITI) was jittered, ranging from 350 – 1350 ms, and target stimuli were inserted only in the longest ITIs, lasting for 300 ms.

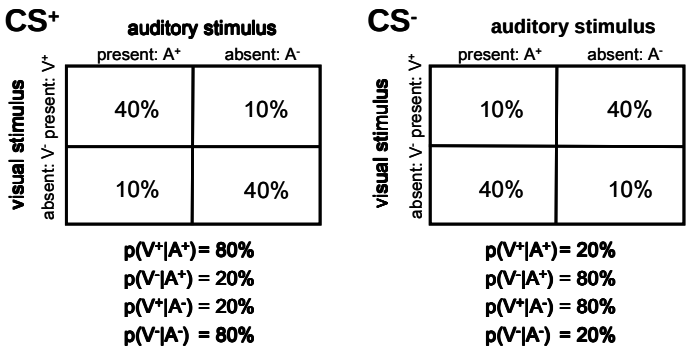

Table 3.1. Probabilistic relationship between auditory and visual stimuli. Contingency tables showing the proportion of a each trial type occurring during CS+ and CS- blocks respectively. Below the tables are the resulting conditional probabilities of the visual stimulus being present (or absent), given the presence (or absence) of the auditory conditioned stimulus (CS); these probabilities can be inferred by comparing the frequencies within each column of the table.

3.2.3 fMRI Data Acquisition

A 3 Tesla Siemens Allegra MRI scanner (Siemens, Erlangen, Germany) was used to acquire T1-weighted fast-field echo structural images and multi-slice T2*-weighted echo-planar volumes with blood oxygenation level dependent (BOLD) contrast (TR = 2.08 secs). For each subject, functional data were acquired in 4 scanning sessions of approximately 10 minutes each. 306 volumes were acquired per session (1224 scans in total per subject). The first 6 volumes of each session were discarded to allow for T1 equilibrium effects. Each functional brain volume comprised 34 2 mm axial slices with a 2 mm inter-slice gap, and an in-plane resolution of 3x3 mm. The field of view covered the whole brain, except for the cerebellum and brainstem. The total duration of the experiment was approximately 60 mins per subject.

3.2.4 Data Analysis

3.2.4.1 Functional neuroimaging analysis

fMRI data were analysed using the statistical software package SPM5 (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm). The 1200 images from each subject were realigned to correct for head movements, corrected for movement-by-distortion interactions (Andersson et al., 2001), spatially normalized to the Montreal Neurological Institute (MNI) template brain, smoothed

spatially with a 3-dimensional Gaussian kernel of 8mm full width half maximum and re-sampled to 3x3x3 mm voxels. The data were then modelled voxel-wise, using a GLM that included regressors for all experimental trials as well as regressors for the target detection task. Trial-specific effects were modelled by trains of delta functions convolved with three hemodynamic basis functions (a canonical hemodynamic response function, and its temporal and dispersion derivatives). Additionally, the time-dependent associative strengths from the Rescorla-Wagner model ($\phi_{i,t}^{j}$; see **Equation 3.1**) and their partial derivatives with respect to learning rate (see next Section) were used as parametric modulators of each trial-specific regressor. The data were high-pass filtered (cut-off 128 seconds) to remove low-frequency signal drifts, and a first-order autoregressive model was used to model the remaining serial correlations (Friston et al., 2002a). Contrast images of parameter estimates encoding trial-specific effects were created for each subject and entered separately into voxel-wise one-sample *t*-tests (df = 15), to implement a second-level random effects analysis. We report regions that survive cluster-level correction for multiple comparisons (family-wise error, FWE) across the whole brain at $P<0.05$. Since previous studies demonstrated the role of the striatum and the prefrontal cortex in associative learning (e.g. (Corlett et al., 2004;Fletcher et al., 2001;O'Doherty et al., 2004), we performed an additional search restricted to these areas, using anatomical masks generated from the PickAtlas toolbox (Maldjian et al., 2003). Again, we only report activations that survived a small volume correction (SVC) at $P<0.05$.

3.2.5 Rescorla-Wagner model

We used a RW model of associative learning to generate predictors of learning-dependent changes in brain activity (as indexed by the BOLD signal) and inter-regional connectivity over time. The basic principle of this model is that the size of the trial-specific prediction error, i.e. the degree of surprise incurred by an event, determines the change in associative strength. From the train of observed events a learning curve was computed and fitted to the fMRI data. Trial-specific cueing was modelled by means of two separate components (see **Figure 3.1**): the visual TO cue TO, which was present on every trial and the auditory CS *per se*, which was present on half the trials. This allowed us to model learning effects on trials where no CS

was present. In the RW framework, the predicted outcome on trial *t*, ϕ_t^j , is the sum of the associative strengths of each cue component:

$$\phi_{i,t+1}^j = \phi_{i,t}^j + \varepsilon_i(\lambda_t - \phi_t^j) \times u_{i,t} \quad (3.1)$$

where

$$\phi_t^j = \sum_i \phi_{i,t}^j \times u_{i,t} \quad (3.2)$$

On each trial *t*, **Equation 3.1** is calculated separately for each cue component, indexed by *i* (i.e., the auditory CS, and TO), while $u_{i,t}$ indexes which of the cue components is actually present on trial *t*. λ_t indicates the actual outcome at trial *t*, being 1 for V^+ and 0 for V^-; ε_i is the learning rate that determines how strongly the prediction error affects the update of the prediction. Separate components are summed in **Equation 3.2**, where ϕ_t^j is the summed prediction of whether a visual stimulus will be presented at trial *t*, and *j* indexes whether this is a CS+ or CS- trial.[5]

3.2.5.1 Learning rate

A challenge when applying the RW model to our experiment was to determine an appropriate learning rate. In principle this can be done by fitting the model to behavioural data and using the resulting learning rate to construct regressors for the fMRI analysis. However, our experimental design deliberately precluded behavioural responses; instead, learning could only be assessed neurophysiologically in terms of changes in cortical activity and inter-regional connectivity. Alternative strategies are to choose the learning rate based on principled considerations (e.g. (O'Doherty et al., 2004)) or using model comparison (Gläscher and Büchel, 2005). Since we knew from a previous study that learning should occur in the visual cortex (McIntosh et al., 1998), we adopted the approach by Gläscher and Büchel (Gläscher and Büchel, 2005) of optimising the value of ε_i to best explain putative learning-induced

[5] When considered for a single cue per trial, Equation 1 can also be seen as a simple model of Hebbian or associative plasticity. In this context, $\phi_{i,t}^j$ encodes the associative strength, which changes according to the second term in Equation 1. This associative term comprises a (pre-synaptic) input $u_{i,t}$ encoding the outcome on any trial, and a (post-synaptic) prediction error.

responses within the main area of interest, the visual cortex. Because our volunteers did not notice the statistical associations (and thus learning was presumably slow) and since another study of perceptual association learning showed small learning rates ε_{CS} below 0.1 (Gläscher and Büchel, 2005), we tested the following values of ε_{CS} in separate models: 0.01, 0.025, 0.05, 0.075, 0.1. We found that ε_{CS}=0.075 gave the best fit to the data in primary visual cortex for the main contrast of interest (i.e., the 4-way interaction in a random effects second level analysis); this learning rate was then used for further analysis across the entire brain and for the connectivity analyses described below. Importantly, we used a first-order Taylor expansion around the learning rate ε_{CS}=0.075 to make the model less dependent on the particular choice of learning rate and to account for inter-subject variability in the shape of the learning curves. This was implemented by including the partial derivative of the learning curve ϕ_t^j with respect to the learning rate ε_i as an additional parametric modulator in the GLM for the fMRI data.

Given that there was no prior hypothesis about differences between the learning rates of CS+ and CS- trials, the analyses described were performed using identical learning rates for both CS types. However, the results from the GLM analysis of the fMRI data showed that learning effects were largely driven by CS+ trials, which suggested that for CS- trials a smaller learning rate should have been chosen than for CS+. This prompted additional analyses to test this possibility. We examined whether (i) a selective decrease of the learning rate for CS- trials improved the ability to detect learning effects during this trial type, and, more generally, whether (ii) the parameter estimates for the partial derivatives (of the learning curve ϕ_t^j with respect to the learning rate ε_i) indicated that the learning rate for either CS+ or CS- trials was different from ε_{CS}=0.075.

With respect to the first point, the data were re-analysed using lower learning rates for CS- trials (ε=0.05 and 0.025, i.e. $^2/_3$ and $^1/_3$ of the learning rate ε=0.075 used for CS+ trials). Specifically, the critical interaction (CS presence × visual outcome × RW learning restricted to CS- trials) was examined to check whether these lower learning rates would give evidence of learning effects during CS- trials. This contrast

was first tested across the whole brain, and subsequently restricted to those regions which showed significant learning effects for CS+ trials.

To address the second issue, the parameter estimates for the partial derivatives (with respect to the learning rates) of the learning curves were examined. If the learning rate for either CS had been set too high or too low, the parameter estimates for the partial derivative would have deviated significantly, across subjects, from zero. All learning-related contrasts were tested for CS+ and CS- trials separately. Again, this analysis was first performed across the whole brain and subsequently restricted to those areas which showed significant learning effects.

Finally, because of its short duration and small size, the TO cue is less salient than the CS. Since in the RW model the learning rate reflects stimulus properties including salience (Rescorla and Wagner, 1972), ε_{TO} can be assumed to be considerably smaller than ε_{CS}. In this study ε_{TO} was assumed to be four times smaller than ε_{CS}. It should be noted that violations of this assumption are unlikely to have a dramatic effect because the inclusion of the derivatives enables the model to cope with deviations from the assumed learning rates, as was described above.

3.2.5.2 Statistical analysis of learning effects

The association strengths of the different cue components with the visual outcome were determined from the series of observed cue-outcome combinations using **Equation 3.1** and the learning rates established as described above. This resulted in the four "partial" learning curves shown in **Figure 3.2A**: two curves (TO and CS) for each CS type (CS+ and CS-). As described by **Equation 3.2**, the predicted outcome on a given trial is the sum of the predictions for each cue component that is present; **Figure 3.2B** shows this summed prediction for each CS type, either present or absent.

Each of the 8 trial types resulting from the three-factorial design (CS+/CS-) $\times$ (A^+/A^-) $\times$ (V^+/V^-) was represented by a separate regressor in the general linear model. Importantly, learning would be reflected by time-evolving, context-dependent brain responses to the visual stimuli. Learning is therefore a fourth experimental factor that changes, over time, how differential brain responses to visual stimuli depend on the presence of an auditory CS and whether it is presented in a CS+ or CS- context.

Specifically, the emergence of these differential responses should follow the time-course predicted by the RW model. In other words, learning is expressed as a 4-way interaction *CS type* × *CS presence* × *visual outcome* × *RW learning*[6]. The primary goal of our GLM analyses was therefore to test this interaction. To establish which CS was driving this interaction, we also tested the simple (3-way) interactions *CS presence* × *visual outcome* × *RW learning* within each CS type. Finally, to test for responses reflecting the prediction (ϕ_t^j) entailed by the auditory CS, independently of the prediction error ($\lambda_t - \phi_t^j$) elicited by the visual outcome, we tested the 3-way interaction *CS type* × *CS presence* × *RW learning*, which is independent of visual outcome.

In order to test for these learning effects, the partial learning curves served as parametric modulators for their respective regressors. Given that each trial always had a trial onset cue, all 8 trial type regressors were modulated by the TO learning curve. Because the CS was present on only half the trials (A^+ trials), these provided another 4 regressors, resulting in a total of 12 parametric modulators.

The linear summation of these partial learning curves (as predicated by **Equation 3.2**) was achieved by defining appropriate statistical contrasts for the general linear model. By assigning equal contrast weights to the regressors for both cue components (**Table 3.2**), it was possible to test for their summed influence, so that the interaction contrasts were effectively operating on the compound learning curves as shown in **Figure 3B**.

[6] Note that when the *CS* is absent on a specific trial, this trial can be assigned unambiguously to the CS^+ or CS^- factor because trials were blocked by CS type.

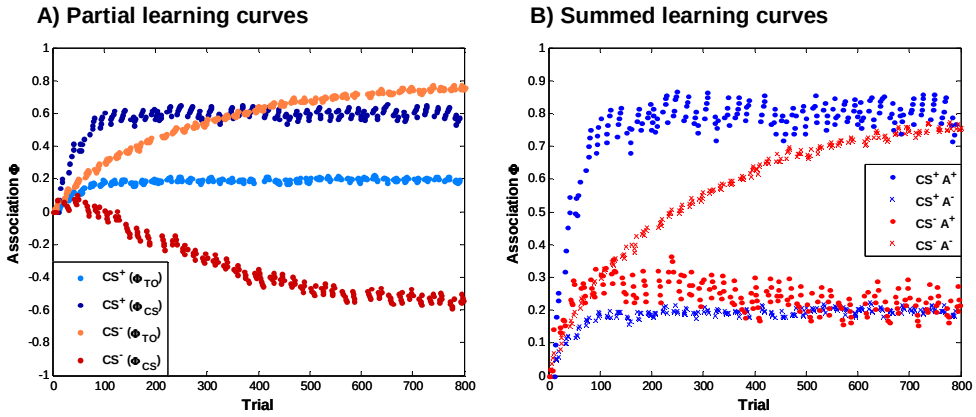

Figure 3.2. Cue-outcome association strengths. (A) Partial learning curves. Trial-specific cueing was modelled with two components: the visual trial onset cue (TO), which was present on every trial, and the auditory CS, which was present on half the trials. This allowed us to model learning effects on trials, when no CS was present. To yield the summed learning curves in **(B)** on each trial the associative strengths of the cues present on that trial are summed, as shown in Equation 3.2. For example, on a CS+A+ trial, both the auditory CS and the TO were present, therefore the total prediction would be the sum of the two blue curves. On a CS+A- trial, the auditory CS is not present, therefore the total prediction is identical to the light blue curve, as only the TO is present. These partial learning curves were used as regressors in the SPM analysis. Note that learning is slower in the absence of an auditory CS than in its presence and faster for CS+ than for CS- trials.

Table 3.2. Contrast weights for parametrically modulated regressors. Contrast weights to test for the 4-way and 3-way interactions, across all 12 modulators. This contrast definition effectively linearly sums all parametric modulators per trial type as described by Equation 3.2.

		CS+ block				**CS- Block**			
		A+		**A-**		**A+**		**A-**	
		V+	**V-**	**V+**	**V-**	**V+**	**V-**	**V+**	**V-**
4-way interaction	**TO**	-1	1	1	-1	1	-1	-1	1
	CS	-1	1			1	-1		
3-way interaction	**TO**	1	1	-1	-1	-1	-1	1	1
	CS	1	1			-1	-1		

3.2.5.3 Prediction error versus prediction

An important feature of our factorial design is that it enabled us to determine whether the responses of a particular brain region reflected the prediction of the visual target or the prediction error. This is important because one cannot include separate regressors based on predictions and prediction errors in the same design matrix. This is due to the form of the RW equation, in which predictions and prediction errors are perfectly correlated (*within* a given experimental condition), after mean-correction (see **Appendix A** for further details). However, in a factorial design like ours such a distinction can be made by analysing the pattern of parameter estimates *across* conditions, contrasting conditions that correspond to expected and unexpected cue-outcome combinations. Specifically, the factorial design provided us, in a mirror-symmetric fashion, with two expected outcomes and two unexpected outcomes for each CS type. For example, on CS+ trials, $A^{+}V^{+}$ and $A^{-}V^{-}$ trials represented expected cue-outcome combinations (conditional probability = 80%) whereas $A^{+}V^{-}$ and $A^{-}V^{+}$ trials consisted of unexpected cue-outcome combinations (conditional probability = 20%); cf. **Table 3.1**). This means one can effectively compare expected and unexpected trials (with low and high prediction error, respectively), with a contrast that is orthogonal to the presence or absence of the visual outcome and its prediction. This enabled us to distinguish, voxel by voxel, brain responses that reflected *expected visual outcomes* from those that represented *unexpected* or *surprising outcomes*. During learning, brain regions encoding prediction errors should show increasing activation on trials where the outcome was unexpected according to the learned contingencies and decreasing (or non-changing) activation on trials where the outcome was expected. We will call such an activation pattern a "prediction error response"; this activation pattern would be expected if surprise was the driving force for learning. In this case, surprising events, or prediction errors, signal the need for learning in order to update predictions. This idea is not only a core component of associative learning models (Schultz and Dickinson, 2000;Shanks, 1995), but is also central to predictive coding theories of perception (Friston, 2005a;Rao and Ballard, 1999): that the brain should concentrate resources on representing surprising sensory events.

Note that our factorial analysis was not geared towards detecting prediction error responses only. It was equally capable of finding opposite activation patterns, i.e.

increasing activation on trials where the prediction based on the learned contingencies matched the outcome and decreasing (or non-changing) activation on trials where the prediction did not match the outcome (cf. Baier et al. 2006). Notably, for our particular design, both types of responses could be identified by the same statistical test, i.e. the 4-way interaction *CS type* × *CS presence* × *visual outcome* × *learning* (see above). Since it is only the direction of the interaction that differs between the two types of responses, our factorial design enabled an analysis that simultaneously tested for these two aspects of associative learning.

3.2.6 DCM

3.2.6.1 Choice of areas and time series extraction

The goal of the present DCM analysis was to explain the (3-way) simple interaction *CS presence* × *visual outcome* × *RW learning* for CS+ trials in V1 (see SPM findings in the Results Section) by a simple model, in which the strength of the A1→V1 connection was modulated as a function of the RW predictions, ϕ_t^j (i.e., learning curves; **Figure 3.2**). Representative A1 time-series were chosen by testing for the main effect of CS presence, and V1 time series were selected by testing for the simple interaction described above. We did not model the 4-way interaction with DCM because the SPM analysis showed that the learning effect was driven by the CS+ (see **Section 3.3.1** for the full SPM results).

As the exact locations of activation maxima varied over subjects, we ensured the comparability of our models across subjects by using combined anatomical-functional constraints in selecting the subject-specific time series (cf. (Stephan et al., 2007c)). Specifically, we thresholded the subject-specific SPMs at $P<0.05$ and chose the local maximum within 8 mm of the group activation maxima in primary auditory cortex (A1) and primary visual cortex (V1) as inferred by a probabilistic cytoarchitectonic atlas in MNI space (Eickhoff et al., 2005). As a summary time-series, we computed the first eigenvector across all supra-threshold voxels within a radius of 4 mm around the chosen local maximum. Overall, we were able to extract time series in 14 out of 16 subjects. In 2 subjects, V1 could not be defined due to the lack of a significant interaction that met the anatomical and functional criteria described above. These 2 subjects were excluded from the DCM analysis.

3.2.6.2 DCM specification

The question addressed by DCM was whether learning effects in V1 could be explained by changes in the connectivity of a simple auditory-visual network. Our DCMs modelled the entire time-series, i.e. data from all trials or conditions, trying to explain regional activations by condition-dependent changes in connectivity. We tested three simple models that could potentially account for the interaction we found in V1. These models were fitted separately to each subject's data and compared using BMS (Penny et al., 2004a). In these models, auditory and visual stimuli from all trials elicited activity directly in their respective primary sensory areas (see **Figure 3.3**). These driving inputs were modelled as individual events. The first model only had a connection from A1 to V1, whereas the second and third model included the reciprocal connection (see **Figure 3.4**). The A1→V1 connection in model 1 and 2, and the V1→A1 connection in model 3 were modulated by the Hadamard product (point-wise multiplication) of the RW associative strength ϕ_t^j and a vector encoding visual outcome (1 for visual stimulus present, -1 for visual stimulus absent) during CS+ trials. In the first two models, this modulatory effect corresponds to the interaction of the auditory CS+ prediction with the visual outcome and models a learning-dependent contribution from CS+ responses in auditory cortex to visual cortex responses that depends on whether the visual stimulus was present or not (cf., a prediction error that rests on top-down signals from auditory areas). In the third model, which represented a control suggested by one of the reviewers, this modulatory effect acted on the reverse connection, i.e. V1→A1.

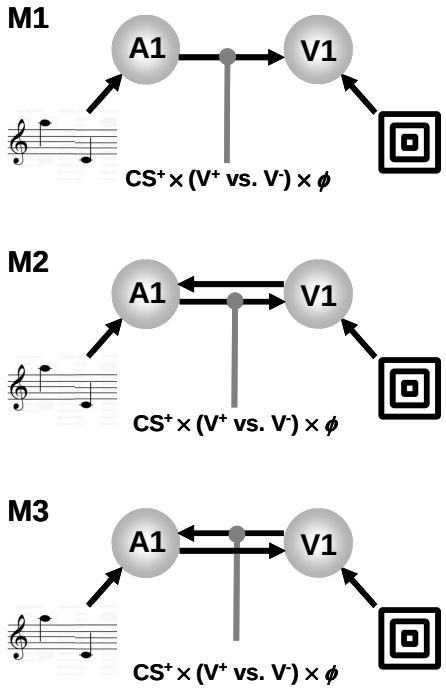

Figure 3.3. DCMs of learning effects on audio-visual connectivity. For all three models, the primary auditory (A1) and visual (V1) areas are both driven by their respective sensory inputs. The first model (M1) had a single connection from A1 to V1; in model 2 (M2), the V1→A1 connection was added. In both M1 and M2, the A1→V1 connection was allowed to change during CS+ trials as a function of the visual outcome (V+ vs. V-) and the Rescorla-Wagner learning curve (ϕ). This modulatory effect corresponds to the interaction of the auditory CS+ prediction with the visual outcome and models a learning-dependent contribution to V1 responses by CS+ related activity in A1; this contribution depends on whether the visual stimulus was present or not (in other words, a prediction error mediated by top-down signals from A1). In the third model, instead of the A1→V1 connection, the V1→A1 connection is modulated by the learning signal.

3.3 Results

The post-scan debriefing questionnaire showed that none of the subjects had become aware of the contingencies between the auditory and visual stimuli. Prior to the fMRI data analysis subjects' performance on the target detection task was verified. On average, subjects responded to 93 ± 3% of the target stimuli.

3.3.1 SPM results

First, we examined the 4-way interaction *CS type* × *CS presence* × *visual outcome* × *RW learning*. We found learning-dependent responses in the primary visual cortex and bilateral putamen that survived whole-brain correction for multiple comparisons (see **Figure 3.4 A, B**). To characterise the nature of this interaction, we tested the simple interaction (*CS presence* × *visual outcome* × *RW learning*) within each CS type. This showed that the 4-way interaction was driven mainly by learning during the CS+ blocks (see **Figure 3E** for the parameter estimates of the visual cortex). As shown in **Figure 3.4 A,B**, testing the simple interaction for CS+ trials afforded almost identical results in the visual cortex and the putamen as the 4-way interaction (see also **Table 3.3**). In contrast, no evidence of learning, i.e. no significant interaction of CS presence and outcome with learning, was found for CS- trials.

The nature of the simple 3-way interaction was such that V1 and the putamen showed an increased response when an expected visual stimulus was omitted, or when an unexpected visual stimulus was presented (i.e. $A^{+}V^{-}$ and $A^{-}V^{+}$ trials). Critically, this response to surprising visual outcomes increased over time as the association was learned, following the form of the RW learning curve. Conversely, V1 responses to predicted stimuli diminished during learning. The putamen showed the same pattern of responses bilaterally; this activation extended into the insula bilaterally (see **Table 3.3**).

Because previous studies have implicated the right DLPFC in prediction (error) processing (Corlett et al., 2004;Fletcher et al., 2001), we used an anatomically defined fronto-striatal mask to test the 3-way interaction *CS type* × *CS presence* × *RW learning*, which characterizes responses to the prediction entailed by the auditory CS, independent of the visual outcome. During learning, the right dorsolateral prefrontal cortex (DLPFC) became increasingly active when a visual stimulus was predicted compared with when it was not; activity was higher for $CS+A^{+}$ and $CS-A^{-}$ trials compared with $CS+A^{-}$ and $CS-A^{+}$ trials (compare the probabilities in **Figure 3.2**). As above, we characterized the nature of the 3-way interaction by testing the associated simple interactions, confirming it was also driven by CS+ trials (**Figure3.4 C**). The same pattern of activation was found in the left putamen, but this activation did not survive correction for multiple comparisons.

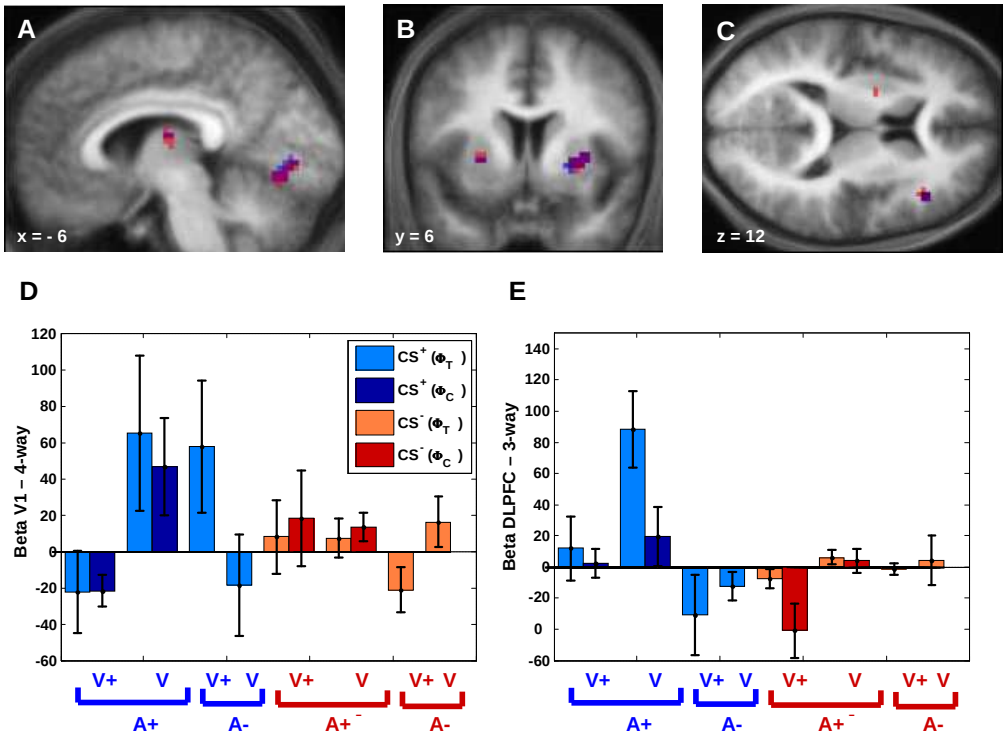

Figure 3.4. fMRI results. **(A)** Significant activations in V1 as a function of RW learning, for the 4-way interaction (CS type × CS presence × visual outcome × RW learning; red), and the same interaction restricted to the CS+ trials (simple 3-way interaction, blue) displayed on the mean structural image across all subjects. The caudate activation is also shown. **(B)** The same interaction in the putamen bilaterally. **(C)**. Significant 3-way interaction CS type × CS presence × RW learning in the DLPFC and left putamen (red). Again the interaction is driven by the CS+ trials, as shown by the simple interaction CS presence × RW learning for CS+ trials only (blue). **(D)** shows the parameter estimates for the 4-way interaction in peripheral V1: (CS type × CS presence × visual outcome × RW learning), where error bars denote standard error across subjects. For all trials on which an auditory CS was presented (A+), the modulatory effects of both the TO (light colours) and the auditory CS (dark colours) were estimated, whereas on A- trials only the TO was present. The estimates show that (mainly for CS+ trials, in blue), there is an increased (summed) response to trials with a surprising outcome, (for CS+ these are the A+V- and the A-V+ trials) and a decreased response to the unsurprising trials (A+V+ and A-V-). The activation in the putamen showed the same pattern of responses. **(E)** shows the parameter estimates for the 3-way interaction (CS type × CS presence × RW learning) in the dorsolateral prefrontal cortex (DLPFC). This represents increased responses when a visual stimulus was predicted to be presented, regardless of the visual outcome. Again, the estimates show that the interaction effect is mainly driven by CS+ trials (blue), showing an increased response to A+ trials.

3.3.1.1 Learning rate

Following Gläscher and Büchel (Gläscher and Büchel, 2005) the optimal learning rate for the RW model was determined, evaluating the primary contrast of interest (that is the 4-way interaction in a random effects second level analysis) under different learning rates in the primary visual cortex (as defined by a probabilistic cytoarchitectonic atlas (Eickhoff et al., 2005). Model fits under five different learning rates, suggested that ε_{CS} = 0.075 was the optimal learning rate (for details on the selection of the learning rates, see **Figure 3.2** and **Section 3.2.5.1**). Given that the learning effects were driven by the CS+ trials, we examined whether any learning effects could be detected for lower CS- learning rates. No learning effects were found at either a corrected level (P<0.05) or at an uncorrected level (P<0.001) for the CS- trials at either of the two lowered learning rates ($^1/_3$ and $^2/_3$ of the original learning rate), either across the whole brain, or when restricted to those regions showing significant learning effects for CS+ trials (i.e., V1, the striatum and dorsolateral prefrontal cortex - DLPFC).

Furthermore, none of the trial-type specific tests of the partial derivatives indicated a learning rate that was different from ε_{CS}=0.075 for the CS- or CS+ trials. If the learning rate had been set too high or too low, the parameter estimates for the partial derivative would have deviated significantly, across subjects, from zero. Again, this analysis was first performed across the whole brain and subsequently restricted to those areas in which significant learning effects had been found, and did not show any significant effects, neither at a corrected threshold (P<0.05) nor at uncorrected thresholds (P<0.001).

Taken together, these additional analyses showed that a selective decrease of the learning rate for CS- trials did not improve our ability to detect learning effects during this trial type, and that there no evidence for either CS+ or CS- trials that a learning rate different from the chosen ε=0.075 is more appropriate for modelling learning effects in the data. This means that the lack of learning effects during CS- trials was not due to a suboptimal choice of learning rate for CS- trials.

Table 3.3. MNI coordinates and Z-values for significantly activated regions.

	MNI coords.				
Foci of activation	x	y	z	Z score	Cluster size
Four-way interaction:					
CS type × CS presence × visual outcome × RW learning					
L occipital lobe*	-6	-75	-9	4.25	41
L insula and putamen*	-30	18	6	4.84	84
L putamen**	-24	12	6	3.85	20
R insula and putamen*	36	12	3	4.72	82
R putamen**	27	6	-3	4.48	35
L caudate/thalamus*	-9	-15	15	4.70	40
L S2 cortex*	-51	-27	24	4.39	93
L middle temporal gyrus*	-57	-39	-3	3.88	26
Simple (3-way) interaction:					
CS presence × visual outcome × RW learning (restricted to CS+)					
L occipital lobe*	-9	-78	-3	4.31	36
L insula and putamen*	-33	12	3	4.55	57
L putamen**	-27	12	6	3.63	10
R insula and putamen*	36	12	3	3.98	57
R putamen**	27	9	0	3.94	32
L caudate/thalamus*	-21	-9	9	4.32	54
L caudate**	-15	-9	21	4.19	14
R caudate**	15	12	18	4.24	7
L S2 cortex*	-60	-33	15	4.15	87
L middle temporal gyrus*	-57	-36	-6	4.30	34
R posterior insula*	39	12	-12	5.01	38
3-way interaction:					
CS type × CS presence × RW learning					
R inferior frontal gyrus**	42	27	12	4.39	10

*significant at P<0.05 (FWE whole-brain cluster-level corrected)

** significant at P<0.05 (SVC)

3.3.2 Learning dependent changes in connectivity

Since the learning effect was mainly driven by CS+ blocks, we focused on changes in connectivity between auditory and visual cortices during incidental learning of the predictive attributes of CS+ trials (see **Figure 3.5**). Bayesian model comparison showed that a DCM with a single connection from A1 to V1 (model 1, cf. **Figure**

3.3) was superior to alternative models with reciprocal connections (GBF in favour of model 1: 2.1×10^{17} and 2.2×10^{18} when compared to model 2 and model 3, respectively). Across subjects, the A1→V1 connection in the optimum model had an average strength of 0.10 s^{-1} (P = 0.003, df = 13, t = 3.57). During CS+ trials, this connection was significantly modulated by learning, depending on whether the visual stimulus was present or not (i.e., CS+ × (V^+ vs. V^-) × ϕ in **Figure 3.5**). Note that the modulatory variable in the DCM corresponds to the interaction of the auditory prediction with the visual outcome during CS+ trials. It accounts for a learning-dependent contribution from CS+ responses in auditory cortex to visual cortex responses that depends on whether the visual stimulus was present or not (cf., a prediction error mediated by top-down signals from auditory areas). Quantitatively, the strength of this modulation was -0.01 s^{-1} (P = 0.028, df = 13, t = 2.49). This corresponds to learning-induced changes in connectivity ranging from 2% (for $CS+A^-$ trials) to 8% (for $CS+A^+$ trials)[7] (**Figure 3.5**).

Critically, the negative sign of the modulatory parameter reflects the nature of the visual responses to auditory afferents under CS+ trials: V1 responses to predicted visual stimuli diminished during learning and the DCM explained this through a decrease in the strength of the A1→V1 connection. This is exactly consistent with an increase in the 'explaining away' of predicted visual input under predictive coding; in other words, if top-down predictions ϕ_t^j (see **Equation 3.2**) from auditory cues decrease the amplitude of V1 prediction error $|\lambda_t - \phi_t^j|$, a better prediction corresponds to a decrease in effective connectivity. Conversely, V1 responses to unpredicted, (i.e., absent) visual stimuli increased during learning. This was modelled in the DCM through an increase in the A1→V1 connection strength; again this is consistent with an increase in V1 prediction error amplitude $|\lambda_t - \phi_t^j|$, when predictions are violated. In summary, A1→V1 influences depended on whether the visual outcome was expected or surprising and were consistent with an 'explaining away' role. The emergence of this effect conformed to the learning curve provided by the RW model.

---

[7] As shown by **Equation 3.2**, the overall strength of a connection, given a single modulatory parameter, is the sum of the intrinsic connection strength (*A*) and the modulatory parameter (*B*) multiplied with its associated input (*u*). In the present case, the asymptotic magnitude of the input function is 0.8 for CS^+A^+ trials and 0.2 for CS^+A^- trials (see **Figure 3.5**).

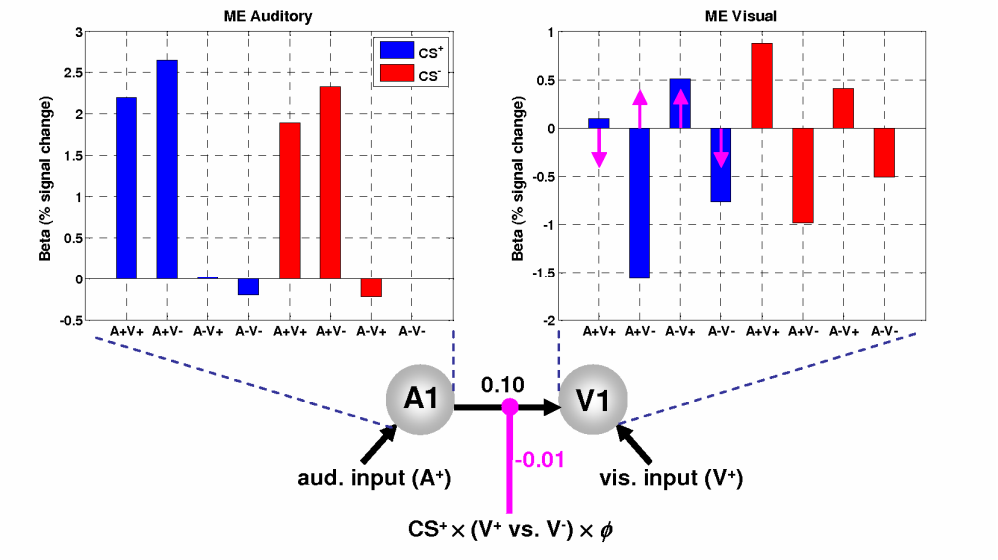

Figure 3.5. Learning effects on audio-visual connectivity. Bayesian model comparison showed that the DCM with a single connection from A1 to V1 was superior to the other models. Across subjects, there was a significant "endogenous" or "fixed" strength of the A1→V1 connection (0.10s-1, P=0.003) and a significant learning-induced modulation (magenta arrows) of this connection (P=0.028). The insets show the parameter estimates for the main effects in both A1 and peripheral V1. The magenta arrows indicate how the main effect in peripheral V1 is modulated by changes in connectivity from A1 to V1 during CS+ trials: over time the response to surprising visual outcomes is up-regulated, whereas the response to unsurprising visual outcomes is down-regulated. Note that in this plot the magenta arrows designate the direction in which V1 responses change due to modulation of connectivity; for quantitative information on this modulatory effect, see the main text.

3.4 Discussion

McIntosh and colleagues showed that after a predictive relationship between an auditory stimulus and a visual stimulus had been learned, the auditory stimulus alone was able to evoke responses in the visual cortex (McIntosh et al., 1998). The current study extended this work, pairing a visual stimulus with a predictive auditory stimulus in a 4-factorial design, with the factors CS type (CS+, CS-), CS presence (A^+, A^-), visual stimulus presence (V^+, V^-) and learning (over time). Both CS+ and CS- blocks were exactly balanced in terms of sensory stimulation, so that the *a priori*

probabilities of the auditory CS and of the visual stimulus occurring on a given trial were always 50%. Critically, the volunteers did not make any responses to the stimuli whose associations were being learned; instead, they performed a target-detection task on unrelated stimuli. Our factorial design enabled us (i) to characterise changes in neurophysiological responses due to learned associations that were incidental to behaviour, and (ii) to investigate whether activity in specific brain areas, and the connection strengths amongst them, reflected a match between predictions and outcome or prediction errors respectively.

The results demonstrate that during incidental learning of audio-visual associations changes in both regional activity and underlying connectivity reflect prediction errors. Furthermore, learning-dependent responses in visual cortex were elicited, even in the absence of visual stimuli. This finding can be explained by changes in top-down influences from auditory regions that are consistent with predictive coding models of perceptual inference.

3.4.1 RW model: predictions & prediction error

The goal of this study was not to pinpoint the exact mathematical form of learning by comparing different models of associative learning. Instead, we focused on changes in regional activity and interregional connectivity that could be explained by a specific learning model, namely the RW model. The RW model is a generic and well-established model of associative learning that has been successful in modelling a wide range of learning processes (Pearce and Bouton, 2001;Rescorla and Wagner, 1972;Schultz and Dickinson, 2000). We chose this model because it is the simplest learning model appropriate for our particular paradigm. In the absence of interactions among multiple cues per trial, the RW model is mathematically equivalent to a Hebbian model of associative learning (Montague and Berns, 2002). A crucial aspect of our paradigm, however, is that on each trial the net prediction resulting from two interacting cue components (the auditory CS and the visual trial onset cue) must be considered (see Methods Sections for details). This excludes the use of any associative learning model that cannot accommodate cue interactions (e.g. Hebbian models). In contrast, the RW model accommodates this aspect gracefully.

The RW model has one problematic limitation, however: as detailed in **Appendix A**, in its equation predictions and prediction errors are perfectly correlated under mean-

correction. In situations where mean-correction is mandatory (e.g., when using them to form interaction terms) this makes it impossible to disambiguate/interpret their contributions to a dependent variable. However, the factorial design in our study allows us to circumvent this problem, as it comprises conditions that correspond to congruent and incongruent prediction/outcome combinations, respectively. Analysing the 4-way interaction between our experimental factors, we found that responses in the primary visual cortex and the putamen were sensitive to surprising events; over time, these areas became significantly more active when presented with a surprising cue-outcome combination. Learning was stronger for the CS+ blocks than for the CS- blocks, which is in line with previous behavioural evidence (Fletcher et al., 2001;Wasserman et al., 1993). Previous fMRI studies in humans have demonstrated that BOLD activity in the striatum is correlated with (signed) prediction errors during reinforcement learning (Jensen et al., 2007;McClure et al., 2003;Menon et al., 2007;O'Doherty et al., 2004;O'Doherty et al., 2003;Seymour et al., 2004) and other associative learning tasks (Corlett et al., 2004). In these studies, the learned associations, and the sign of the resulting prediction errors, were of direct relevance for behaviour. The current study shows that the putamen is sensitive to unexpected outcomes even when the cue-stimulus association is learned incidentally and has no relevance to behaviour. However, in contrast to the previous studies, the pattern of putamen activity does not appear to be sensitive to the direction of the prediction error, only to its amplitude. This difference may reflect the fact that learning was perceptual as opposed to operant. In other words, the occurrence of an unpredicted or surprising event may play the role of negative reward, irrespective of whether the surprising event entailed the presence of absence of a stimulus. This issue will be discussed further in the Section on predictive coding below.

3.4.2 Role of prediction errors beyond reinforcement learning

Our finding that learning-induced responses in primary visual cortex and the putamen reflected prediction errors accords with a basic principle emerging from many previous studies: prediction errors, or surprise, constitute a driving force for learning because they signal the need for learning in order to update predictions (Schultz et al., 1997;Schultz and Dickinson, 2000;Shanks, 1995). Although the role of prediction errors has been mainly explored for reinforcement learning so far, there

is growing evidence that prediction errors may be equally important for learning statistical relationships that are affectively neutral and behaviourally irrelevant. In other words, the same mechanisms that optimise the learning of stimulus-response links may operate during the perceptual learning of stimulus-stimulus associations (Friston, 2005a;Rao and Ballard, 1999). Evidence that organisms learn predictive associations between initially neutral stimuli is seen in classical conditioning effects such as sensory preconditioning (Brogden, 1939). Some forms of sensory learning also exhibit such features, e.g. the mismatch negativity paradigm, in which responses to sensory stimuli decrease with predictability (Baldeweg, 2006;Friston, 2005a), regardless of whether stimuli are attended. A mechanism similar to predictive coding has been proposed in the motor domain for cancellation of self-generated events (Blakemore et al., 1998;Shergill et al., 2005;Wolpert et al., 1995). Moreover, the learning of predictive relationships that are affectively neutral and task-irrelevant may engage similar computational and neural mechanisms as those for predicting significant events (Wittmann et al., 2007;Zink et al., 2006).

The results of the present study support the notion that the role of prediction errors in learning transcends the simple reinforcement of stimulus-response links and plays a more pervasive and general role in various forms of learning. Indeed a hallmark of adaptive systems is their ability to minimise surprising exchanges with their environment (Friston et al., 2006). This entails adjustments to their internal models of the environment so that potentially surprising event can be predicted. Almost universally, this adjustment involves changes in the system's connections; it is therefore perhaps a little surprising that most previous imaging studies on learning and conditioning have exclusively searched for brain areas whose activity correlated with specific variables of a particular learning model (e.g., prediction or prediction error), but have not investigated how these variables change interactions among areas (although some studies have investigated learning-dependent changes in connectivity without using a learning model; (Büchel et al., 1999;McIntosh et al., 1998)). Changes in connectivity are central to the physiological implementation of learning; it has long been suggested that plasticity in connection strengths between neurons underlies the learning of predictive associations (Hebb, 1949). Put simply, two neural units encoding associated entities increase their synaptic connections to encode the learned associative strength of the stimuli. More precisely, for RW and similar ‘caching’

models (Daw et al., 2005) the connection strength at time *t* should carry the predicted association at time *t* (McLaren et al., 1989;Schultz and Dickinson, 2000). This hypothesis requires models of effective connectivity, in which connection strengths vary as a function of the associative strength predicted by the learning model. To our knowledge, the present study has implemented this approach for the first time, modelling how learning, as described by a RW model, modulates the effective connectivity, as assessed by a DCM, between primary auditory and visual areas.

3.4.3 Changes in connectivity between auditory and visual areas

In accordance with the considerations above, we investigated whether the learning-related changes in visual cortex responses could be explained by a simple model of effective connectivity, in which the strength of A1→V1 connection changed as a function of the associative strength predicted by the RW model. We modelled observed responses in the primary visual cortex by means of a simple 2-area DCM in which activity in the visual cortex was modelled by two components, (i) a direct effect of visual stimulation and (ii) a modulation of the A1→V1 connection by the interaction of the time-evolving prediction with the visual input (in CS+ blocks; see **Figure 6**). Across subjects, this DCM showed a significant change in the strength of the A1→V1 connection congruent with the pattern of responses in V1: the A1→V1 connection strength increased on trials where the visual outcome did not match the auditory prediction and decreased on trials where prediction and outcome matched. In other words, the learning-induced changes in A1→V1 connection strength reflected the same pattern of surprise or prediction errors as the regional activity in V1. This demonstrated that the response of V1 to visual stimuli was modulated by learning-dependent changes in top-down auditory influences that were consistent with the notion of predictive-coding, a general framework for perceptual inference and learning that is discussed in the next section (Friston, 2005a).

Although connections in models of effective connectivity do not need to correspond to monosynaptic anatomical connections, it is of interest to note that the surprise-related response in visual cortex appears to be in the peripheral visual field (**Figure 3.3 A**), and anatomical connections from primary auditory cortex to peripheral visual cortex have been demonstrated in recent monkey studies (Falchier et al., 2002;Rockland and Ojima, 2003). Additionally, numerous fMRI studies have

demonstrated that auditory stimulation or auditory attention affect activity in visual cortices during simultaneous processing of visual stimuli (e.g. (Baier et al., 2006;McIntosh et al., 1998;Watkins et al., 2006)).

3.4.4 Predictive coding in visual cortex

In previous neurophysiological studies of reinforcement learning, a negative prediction error, i.e. the unexpected absence of a reinforcer (e.g. a reward), often led to a decrease in neuronal or BOLD activity (McClure et al., 2003;Schultz, 1998;Tobler et al., 2007). Such directed excursions are thought to reflect the fact that the prediction error is a signed quantity: it signals not just that predictions need to be updated, but in which direction. In contrast, in our study we found an increase in striatum and visual cortex activity not only for unexpectedly presented stimuli, but also for the unexpected absence of a stimulus. Similarly, the strength of the A1→V1 connection decreased whenever the visual outcome was expected, and it increased whenever the outcome was surprising.

A useful perspective that explains our two main findings (the implicit encoding of surprise by V1 responses and its mediation by learning-dependent changes in input from the auditory cortex) is provided by the framework of predictive coding. Predictive coding posits a hierarchy of connected brain areas in which each level strives to attain a compromise between information about sensory inputs provided by the level below and predictions (or priors) provided by the level above (Friston, 2003;Murray et al., 2002;Rao and Ballard, 1999;Summerfield et al., 2006). The central learning principle is to establish a good model of the world, which is achieved by changing connection strengths such that prediction errors are minimised at all levels of the hierarchy. The hierarchy of a predictive coding architecture is often defined anatomically (in terms of forward and backward connections) and within one sensory modality, but it is equally possible to examine cross-modal predictive coding relationships (cf. (von Kriegstein and Giraud, 2006)). In the present study, a temporal hierarchical relation between auditory and visual areas is induced by presenting the auditory cue prior to the visual stimulus.

Predictive coding may be a general principle of brain function in which statistical relationships in the world are monitored, even when they are not attended and not relevant for ongoing behaviour. This would allow the brain to ignore predictable and

therefore uninteresting events in the environment, thereby enhancing the saliency of unexpected events. A good example of this notion is given by the so-called mismatch negativity (MMN), the difference between the event-related potential to an unexpected "deviant" and predictable "standard" stimuli (Naatanen et al., 2001). Importantly, the relationship between the MMN and learning was not established on the basis of behavioural data; in fact, it was initially not even recognised (Naatanen et al., 1978). This relationship was only subsequently inferred from striking relationships between the probability of deviants and neurophysiological time-series (e.g. (Csepe et al., 1987;Pincze et al., 2002). Current theories of MMN, which interpret it as a paradigmatic example of learning based on predictive coding (Baldeweg, 2006;Friston, 2005a), have recently received empirical support by DCM studies of electroencephalographic measurements (David et al., 2006;Garrido et al., 2007). These studies demonstrated that MMN can be understood as a prediction error signal, which results from deviant-induced changes in inter-regional connection strengths. A similar conclusion is offered by the present study. Here, we found that, at least during CS+ trials, BOLD responses in area V1 increased when the prediction provided by the auditory cue did not match the subsequent visual stimulus (analogous to MMN elicited by deviants). This surprise signal progressively increased as the predictive properties of the auditory cue were learnt. Moreover, in direct analogy to DCM studies of the MMN (David et al., 2006;Garrido et al., 2007), we found a decrease in the A1→V1 connection strength on "standard" trials (where the prediction by the auditory cue was correct), and an increase on "deviant" trials where the visual outcome did not match the prediction by the auditory cue. In the context of predictive coding, learning involves a more efficient suppression of sensory events, which is manifest by an apparent reduction in evoked responses, mediated by top-down predictions (which explain away bottom-up sensory afferents). Within the framework of our bilinear DCM, this is modelled as a decrease in top-down effective connectivity for visual stimuli that match the current prediction.

3.4.5 Limitations and future directions

We conclude this chapter by discussing a number of limitations of the present study. First, because we wished to study brain responses to stimulus associations that were

irrelevant to behaviour, we did not obtain behavioural evidence for learning. Instead, as with the MMN paradigm described above, learning is characterised neurophysiologically as a change in activity over time. **Chapter 4** will describe the results of a follow up experiment with stimuli that do require a behavioural response and thus provide a behavioural assessment of the learning process. It might be useful to emphasise that a neurophysiological characterisation of incidental associative learning processes only requires that the statistical associations between the CS/US stimuli are irrelevant for task performance. In contrast, it is not essential that the CS and US stimuli themselves are behaviourally irrelevant. In fact, the stimuli had some behavioural relevance insofar as they constitute distractors to which responses must be suppressed.

Secondly, the DCM presented here does not make any assumptions about where in the brain the predicted associative strength is calculated; i.e. which brain area exerts the modulatory influence onto the A1→V1 connection. Given the responses that were observed in the putamen, it is possible that the modulation of the A1→V1 connection is mediated via this region. Testing this hypothesis, however, requires the inclusion of non-linear terms in the neuronal state equation of DCM which goes beyond its bilinear mathematical framework. **Chapter 2** described a nonlinear extension of DCM (Stephan et al., 2008), which allows one to investigate the source of the modulatory influences. **Chapter 4** describes how this approach has been applied to a different associative learning study. Nevertheless, notwithstanding these limitations, the current study has presented a novel combination of dynamic system models and formal learning theory, which were used to model human neuroimaging data. This is a further step towards the long-term goal of constructing invertible models that unite the neurophysiological and computational aspects of learning (cf. (Stephan, 2004)).

Chapter 4

Striatal Prediction Error Activity Drives Cortical Connectivity Changes During Associative Learning

Abstract

Both perceptual inference and motor responses are shaped by our estimates of probabilistic relations among events in the world. Here, we investigated how (failures of) learned predictions about sensory stimuli influence subsequent motor responses. In an associative learning paradigm auditory cues differentially predicted subsequent visual stimuli. Critically, the predictive strengths of cues were unknown and varied over time, requiring subjects to continuously update estimates of stimulus probabilities. This dynamic inference, which we modeled using a hierarchical Bayesian observer, was reflected behaviourally: speed and accuracy of motor responses significantly increased with trial-by-trial predictability of visual stimuli. Dynamic causal modeling of fMRI data showed that activity in the putamen (i) increased the more surprising the current visual stimulus was and (ii) enhanced the strength of connections from visual areas to dorsal premotor cortex by non-linear gating. Thus, the degree of striatal trial-by-trial prediction error activity controlled the plasticity of visuo-motor connections.

4.1 Introduction

One of the major reasons for the remarkable flexibility and adaptive repertoire of human behaviour is that the human brain can construct, and rapidly update, estimates of conditional probabilities that describe causal relationships in the world. For

example, human subjects can infer changing conditional probabilities among sensory events (Behrens et al., 2007;Brodersen et al., 2008), even when these probabilities are currently not relevant for behaviour (den Ouden et al., 2009). Such learning of stimulus probabilities has been shown to be reflected by activity changes in visual (Summerfield et al., 2008;Summerfield and Koechlin, 2008), auditory (Pincze et al., 2002) and somatosensory areas (Akatsuka et al., 2007). The general principle, across all modalities, is that sensory responses increase with the size of prediction error, i.e. the more surprising they are. This is in accordance with current theoretical accounts of brain function, e.g. predictive coding (Friston, 2005a;Rao and Ballard, 1999), which posit a fundamental role of prediction errors for adaptive behaviour and learning.

Efficient learning of probabilities can be used to form predictions which guide motor behaviour. For example, once the predictive strength of a cue has been learned, the premotor cortex shows preparatory activity (Crammond and Kalaska, 2000;Tanji and Evarts, 1976;Wise and Mauritz, 1985) and reaction times decrease (e.g. (Bestmann et al., 2008;Requin and Granjon, 1969;Strange et al., 2005).

A critical question is what neurobiological mechanisms underlie the adaptive changes in motor behaviour that are needed when predictions fail, e.g. in rapidly changing environments. According to predictive coding theories, any prediction error should induce learning and thus synaptic plasticity, reconfiguring connection strengths in somato-motor networks such that prediction error is eventually minimised both at sensory and motor levels (Friston and Stephan, 2007). In a similar vein, Bestmann et al. (Bestmann et al., 2008) suggested that "... the brain tries to minimise prediction error ... that is then continuously channelled into motor regions to control the excitability of expected motor outputs". In this study, we provide direct empirical evidence for this idea, exploiting recent advances in computational models of learning (Behrens et al., 2007) and nonlinear DCMs of fMRI data (Stephan et al., 2008). In particular, we link the physiological mechanisms proposed by predictive coding, i.e. prediction error dependent changes in connectivity, to a large body of literature which have described prediction error responses in the striatum (Corlett et al., 2004;Jensen et al., 2007;McClure et al., 2003;Menon et al., 2007;O'Doherty et al., 2004;O'Doherty et al., 2003;Seymour et al., 2004). Specifically, we show that the observed learning-dependent changes in BOLD activity are compatible with a

mechanistic model in which the strengths of visuo-motor connections are modulated by prediction error related activity in the striatum.

4.2 Methods & Statistical analysis

Twenty healthy right-handed volunteers, 24.4 ± 2.1 years of age, (mean age ± SD, 10 female) took part in this study. The participants had no history of psychiatric or neurological disorders. Written informed consent was obtained from all volunteers prior to participation, which was approved by the National Hospital for Neurology and Neurosurgery Ethics Committee.

The central idea of this paradigm was to present participants with auditory stimuli that differentially predicted upcoming visual stimuli. Participants had to report whether the visual stimulus was a face or a house. They were instructed that the relation between auditory and visual stimuli was probabilistic, that these probabilistic relations were changing unpredictably in time and that there was no underlying rule to be learned or discovered. They were neither informed about the magnitude or distribution of the probabilistic relations nor about the temporal intervals at which they changed.

4.2.1 Conditioning

On each trial, one of two auditory cue stimuli (CS_1 and CS_2) was followed by a visual target stimulus (**Figure 4.1A**). Participants were instructed to respond as quickly as possible by button press (right middle and index finger, counterbalanced across subjects) and report whether the target stimulus was a face (F) or a house (H). Auditory and visual stimuli were presented for 300 ms and 150 ms, respectively. In order to prevent automatic responses or guesses, both the inter-trial interval (2000 ± 650 ms) and visual stimulus latency (150 ± 50 ms) were jittered randomly (**Figure 4.1A**).

The two tones differentially predicted the identity of the visual target stimulus, and these contingencies were changing in time (**Figure 4.1B**). Because each CS was followed by one of two stimuli (F or H), the probability of one visual stimulus, given a particular auditory CS, was one minus the probability of the other visual stimulus:

$$p(F \mid CS_i) = 1 - p(H \mid CS_i), \quad i \in \{1,2\} \tag{4.1}$$

To prevent that participants' responses could be biased by learned expectations (e.g. about the relative frequencies of the visual stimuli), we constrained the sequence of changes in probabilities such that at any point in time the marginal probabilities of faces and houses were identical. First, the probability of one visual outcome given CS_1 was the same as the probability of the other visual outcome given CS_2 (compare **Figure 4.1B**):

$$p(F \mid CS_1) = p(H \mid CS_2) \tag{4.2}$$

Secondly, each block contained equal numbers of randomly intermixed CS_1 and CS_2 trials. With these two manipulations, we ensured that on any given trial, before the CS was presented, the *a priori* probability of a face (or house) occurring was always 50%. Thus, any expectations about the visual stimulus could exclusively be evoked by and were time-locked to (the onset of) the auditory stimulus.

Each subject completed five sessions of 200 trials each. In each session, the predictive strengths of the two CS types were changing pseudorandomly over time, taking one of 5 different discrete levels of predictive association; the probability of the visual outcome stimulus ($p_{outcome}$) could be (i) strongly predictive (p = 0.9), (ii) predictive (p = 0.7), (iii) non-predictive (p = 0.5), (iv) anti-predictive (p = 0.3), and (v) strongly anti-predictive (p = 0.1). Each predictive level was presented as a block of stimuli once per scanning session. Predictive block lengths varied between 14-20 trials per CS type, so that participants could not predict when exactly a change in contingencies would occur. Furthermore, blocks with predictive cues alternated with short blocks (6-10 trials) containing non-predictive cues (i.e. p = 0.5) in order to avoid complete reversals of the contingencies.

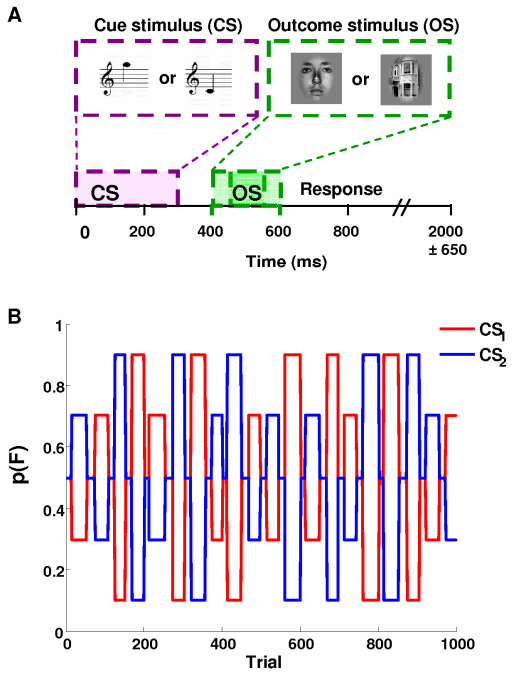

Figure 4.1. Experimental design. **(A)** Timeline for a single trial. At trial onset the auditory cue stimulus (CS) was presented for 300 ms. The visual stimulus lasted for 150 ms and was presented 150 ± 50 ms after the CS. The inter-trial interval lasted for 2000 ms on average (± 650 ms). **(B)** Temporal evolution of the probability of a face, *p*(F), occurring given either CS. Note that the probability of a house being presented is simply the mirror image of this sequence.

4.2.2 Stimuli

Eight pictures of neutral facial expressions drawn from the Ekman Series of Facial Affect (Ekman and Friesen, 1976) and eight pictures of houses were used as visual stimuli. Stimuli were matched for overall luminance and presented on a gray background. The auditory stimuli were matched for perceived loudness under scanning conditions as described previously (**Chapter 3**). The frequencies of the auditory stimuli used in this experiment were 1125 Hz and 500 Hz, and the adapted volume of the high tone was 98 ± 4.1 % (mean ± SD) with respect to the low tone. To maintain identical visual input conditions, all visual stimuli were presented

centrally and for a duration of 150 ms to prevent saccades, and subjects were required to fixate a central cross throughout the experiment. Stimuli were presented using the software package Cogent (www.vislab.ucl.ac.uk/Cogent).

4.2.3 fMRI Data Acquisition

A 3 Tesla head scanner (Allegra Magnetom, Siemens Medical, Erlangen, Germany) was used to acquire a T1-weighted fast-field echo structural image and multi-slice T2*-weighted echo-planar volumes with blood oxygenation level dependent (BOLD) contrast (TR = 2.73 sec, TE = 30 ms). Furthermore, prior to the functional scans, a B_0 field map was acquired using a gradient echo field map sequence. Functional data were acquired in five scanning sessions of approximately eight minutes each. 189 volumes were acquired per session (945 scans in total per subject). The first six volumes of each session were discarded to allow for T1 equilibrium effects. Each functional brain volume comprised 42 axial slices with 2 mm thickness and a 2 mm inter-slice gap and an in-plane resolution of 3x3 mm. The field of view was chosen to cover the whole brain, except for the brainstem. The total duration of the experiment was approximately 90 minutes per subject.

4.2.4 Data Analysis

4.2.4.1 Behavioural data analysis

First, the data were screened for outliers in reaction times. Responses faster than 150 ms were excluded. We then tested whether the distributions of reaction times (RT) and response speeds (RS; i.e. inverse reaction times) showed significant deviations from normality using a Kolmogorov-Smirnov test. Since RS, but not RT, were well described by a Gaussian distribution, the former were entered into a repeated-measures analysis of variance (ANOVA) with outcome probability, CS type (CS_1/CS_2) and outcome type (F/H) as within-subject factors. The Greenhouse-Geisser correction was employed where significant non-sphericity was detected. We tested for any main effects and interactions between these factors that were expressed in the RS. Furthermore we also assessed the main effect of outcome probability on error rates.

4.2.4.2 Bayesian learning model

The above ANOVA indicated that there was a significant acceleration of reactions with increasing probability (for details, see **Results Section 4.3** and **Figure 4.3C**). This simple linear model of the behavioural data is not very realistic, however, because it assumes instantaneous and precise knowledge of the probabilities that generated the stimulus sequence. In reality, the participants had to estimate these unknown probabilities from the observed stimulus sequence. One possibility is that subjects behave like Bayesian observers which continually update their estimates of the hidden contingencies by combining prior information from the past with current observations in the present. As described in **Chapter 1**, in standard Bayesian observer models, the learning rate, and thus the relative influence of past vs. current observations on the estimates, is unchanging. This, however, is not an ideal approach for our experimental paradigm where the underlying probabilistic associations are changing in an unknown and irregular fashion. In such an environment, an optimal learner would not only estimate the probabilities, but also their instability in time, i.e. volatility, and would increase the weight of current observation relative to past experience with increasing volatility of the environment.

Behrens et al. (2007) developed a hierarchical Bayesian learning model that represents such an ideal observer (Behrens et al., 2007). Given a series of observed events, this model estimates, at any given point in time, the posterior probability density function (PDF) of both the probabilistic associations and the volatility of the environment (**Figure 4.2**). Here, we adopted this model (see **Appendix B** for implementation details) and used the posterior mean of the PDFs as estimates of the probability and volatility. In order to verify that the probability estimate of this Bayesian model were better linear predictors of the behavioural RS than the true probabilities that generated the stimulus sequence, we used Bayesian model selection as described in the next section. Given the clear superiority of probability estimates from the Bayesian model in explaining the behavioural data, they were subsequently used in the analyses of the fMRI data.

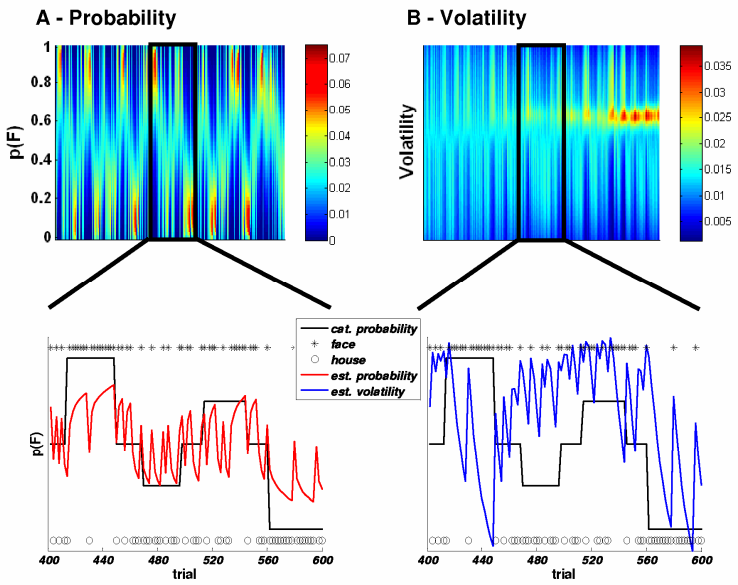

Figure 4.2. Trial-by-trial probability and volatility estimates. (A, top): evolution of the posterior probability density function (PDF) of $p(F|CS_1)$ across the entire experiment. **(A, bottom)** The posterior mean of $p(F|CS_1)$ (solid line) for session three clearly tracks the underlying blocked probabilities (dashed line). Because blocks of stable probabilities are short, however, the estimated probabilities never quite reach their true values during a given block. Note that the estimates change rapidly at block transitions. When an unexpected stimulus occurs, the estimates briefly move towards $p = 0.5$ (visible as "spikes" in the trajectory of the posterior mean).

(B, top) The posterior PDF of the volatility shows the initially high uncertainty about the volatility of the environment, which converges in the course of the experiment. The estimated posterior mean of the volatility **(B, bottom)** decreases over the course of a block, particularly when the probability is very high or very low ($p = 0.9$ and $p = 0.1$), and spikes between blocks. Additional spikes within blocks are present when an unexpected stimulus occurs.

4.2.4.3 Bayesian model selection (BMS)

When comparing different models for observed data, it is critical that the decision is not only based on the relative fit, but also on the relative complexity of the competing models (Pitt and Myung, 2002). For comparing competing models, both of behavioural and of fMRI data, BMS provides a principled foundation for such model comparisons (Penny et al., 2004a). In this study, we used a novel hierarchical

method for BMS that allows for group-level random effects inference about the relative goodness of multiple competing models (cf. **Chapter 2** and (Stephan et al., 2009)). In brief, for all models considered we computed the evidence $p(y \mid m)$, i.e. the probability of the data *y* being generated by model *m*, for each subject. Integrating out the model parameters, the model evidence balances fit and complexity, enabling one to compare non-nested models with different levels of complexity. For the linear models that were applied to the behavioural data, there is an analytic expression for the model evidence (see **Chapter 2** for details). For the nonlinear DCMs of the fMRI data described below, we used the negative free energy approximation to the log model evidence (cf. (Friston et al., 2007;Stephan et al., 2007d).

Subsequently the models were compared at the group level, using a new method for random effects BMS (Stephan et al., 2009). This method uses hierarchical variational Bayes to infer the posterior density of the models *per se*. This rests on treating the model as a random variable and estimating the parameters α of a Dirichlet distribution describing the models' probabilities *r*. One can then use the cumulative probability density of $p(r \mid y;\alpha)$ to quantify an *exceedance probability* φ_k, i.e. our belief that a particular model *k* is more likely than any other model (of the *K* models tested), given the group data. Exceedance probabilities are particularly intuitive when comparing two models (as in our analysis of the behavioural data; see **Figure 4.3D**). For example, when comparing two models, m_1 and m_2, the probability that m_1 is a more likely model than m_2 can be written as

$$\varphi_1 = p(r_1 > 0.5 \mid y;\alpha) \qquad (4.3)$$

This hierarchical Bayesian approach has been shown to be considerably more robust than either the conventional fixed effects analysis using group Bayes factors (Penny et al., 2004a), or frequentist tests applied to model evidences, especially in the presence of outliers (Stephan et al., 2009).

4.2.4.4 Functional neuroimaging analysis

fMRI data were analysed using the SPM5 software package (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm). The 915 EPI images from each subject were corrected for geometric distortions caused by

susceptibility-induced field inhomogeneities. A combined approach was used which corrects for both static distortions and changes in these distortions due to head motion (Andersson et al., 2001;Hutton et al., 2002). The static distortions were calculated for each subject by acquiring a B_0 field map and processing it using the FieldMap toolbox implemented in SPM5 (Hutton et al., 2004). The images were then realigned and unwarped using SPM5 (Andersson et al., 2001) which allows the measured static distortions to be included in the estimation of distortion changes associated with head motion. The data were temporally interpolated, using the middle slice in time as a reference, to account for slice-timing effects. The structural image was then coregistered with the mean unwarped functional image and processed using the unified segmentation procedure implemented in SPM5, with the default tissue probability maps. This procedure combines segmentation, bias correction and spatial normalization through the inversion of a single unified model (Ashburner and Friston, 2005). The same normalisation parameters were then applied to normalise the unwarped and realigned EPI images. Finally the EPI images were smoothed spatially with a three-dimensional Gaussian kernel of 8 mm full width half maximum and re-sampled to 3x3x3 mm voxels.

The data were then modelled voxel-wise, using the GLM for each of the 20 participants. In the GLM, correct and error trials were modelled as separate events. For correct trials, face and house trials were modelled as the two main conditions of interest. These were collapsed across the two different CS types, because the predictive strengths of the two CSs were counterbalanced over time and thus no differential effects were to be expected (analysis of the behavioural data also indicated the absence of such effects). Condition-specific effects were modelled in an event-related fashion, convolving a sequence of delta functions with a canonical hemodynamic response function. The probability estimates from the Bayesian observer as well as the subject-specific response speeds were included as first-order parametric modulators of face and house trials such that the delta functions representing the presence of a face were modulated by the trial-specific probability estimate that a face should have occurred on this trial (equivalently for house trials). We also included the volatility estimates from the Bayesian observer as parametric modulators (orthogonalised to the probability estimates). Finally the 6 parameter

vectors from the realignment procedure were included as regressors of no interest to account for variance caused by head motion.

After computing subject-specific contrast images of interest, random effects group analyses across all 20 subjects were performed (Friston et al., 2005), using one-sided one-sample *t*-tests and testing for both positive and negative activations. We report any activations that survived whole brain correction at the cluster-level ($P<0.05$). For anatomically constrained *a priori* hypotheses concerning stimulus-specific visual areas, putamen and ACC, we used masks and report activations that survived correction at the cluster-level within the region of interest ($P<0.05$). For the putamen and ACC, these masks were generated using the PickAtlas toolbox (Maldjian et al., 2003); for stimulus-specific visual areas, we used in-built localiser contrasts that were orthogonal to all other contrasts of interest.

Firstly we assessed the main effect of probability, that is, in which brain regions the activity reflected the probability of the stimulus occurring, independently of which stimulus it was. We tested both for activations that increased with the likelihood of the outcome and for activations that increased the less likely, i.e. more surprising, the outcome was. In other words, this contrast tested for stimulus-independent responses that reflected predicted or surprising outcomes, respectively. Given the results from our previous study (den Ouden et al., 2009), our *a priori* hypothesis was that activity in the putamen would increase the more surprising the outcome was.

Secondly, we tested for stimulus-by-probability interactions, that is, probability-dependent responses that differed between faces and houses. Our *a priori* hypothesis was that activity in stimulus-specific areas should scale inversely with the probability of the presented stimulus. In other words, responses of the fusiform face area (FFA) to face stimuli should decrease the more likely the presentation of a face had been on a given trial, and responses of the parahippocampal place area (PPA) to houses should decrease with the probability of a house being presented. This can be regarded equivalently as testing for surprise-dependent increases in the activity of stimulus-specific areas. To accommodate inter-subject variability in the exact location of FFA and PPA, we performed a region-of-interest analysis. Concerning the functional definition of FFA and PPA, we did not need a separate localiser scan since our factorial design provided an in-built localiser contrast (i.e. the main effect

of faces versus houses). Note that this contrast is orthogonal to the contrast testing for interactions and can thus be used to define regions of interest. In each subject the individual maximum within an 8 mm radius from the group maximum of face- and house-specific responses (see **Table 4.1**) was determined. Subsequently, given these voxels with individually maximal stimulus-specificity, we tested for (orthogonal) stimulus-by-probability interactions by entering the parameter estimates of regressors encoding trial-by-trial stimulus probability estimates into two-tailed one-sample *t*-tests. In other words, this procedure tested whether face- and house-specific responses in FFA and PPA, respectively, were modulated by the trial-by-trial probability estimate of a face or a house occurring.

4.2.4.5 Nonlinear DCMs

Numerous studies have demonstrated previously that activity in the putamen reflects prediction errors or surprise (e.g. (den Ouden et al., 2009;Jensen et al., 2007;McClure et al., 2003;O'Doherty et al., 2004;Pessiglione et al., 2006)). According to theoretical models of learning, the size of prediction errors should control the magnitude of synaptic plasticity, and thus changes in connection strength, that underlies the learning process (Friston, 2005a;McLaren et al., 1989;Schultz and Dickinson, 2000). In this study, we tested this notion directly by modelling how activity in the putamen gated the information flow from visual areas to the dorsal premotor cortex (PMd; **Figure 4.5**). We expected that increased activity in the putamen, induced by a surprising face, should gate the strength of the FFA→PMd connection, thus enhancing the influence of face information on PMd activity and facilitating an update of the motor plan. This type of analysis, which requires one to study non-linear (second order) modulatory effects on connectivity, has become possible with the recent introduction of nonlinear DCMs described in **Chapter 2** (Stephan et al., 2008).

4.2.4.5.1 DCM specification

Based on our SPM results, we constructed a nonlinear DCM including the right putamen, PPA and FFA, and the left PMd. As shown in **Figure 4.4** and **Table 4.1**, several other areas showed a surprise dependent response and are likely to be involved in the visuomotor transformation; the present model with the above four regions should be regarded as the most parsimonious model that enabled us to test

whether surprise-related activity in the putamen gated visuomotor connections. While putamen, FFA and PPA showed peak activations in the right hemisphere, we included the left premotor cortex as participants were responding with their right hand.

We constructed and compared several alternative models. A basic DCM shown by **Figure 4.5A** included connections from FFA and PPA to the PMd, and modulations of these connections by activity in the putamen, which was driven by the trial-by-trial probability estimates provided by the Bayesian learning model. The endogenous connectivity structure of this DCM was subsequently optimised systematically by BMS (see **Figure 4.5** for a graphical representation of all models tested).

After the endogenous connections had been optimised, we conducted a final and critical model comparison. Since the putamen and the PMd showed similar surprise–related activations (**Figure 4.4**), we wanted to establish the specificity of our model and demonstrate that putamen activity gated visuo-motor connections, instead of PMd gating visuo-putamen connections. We therefore tested a DCM, in which the roles of the PMd and the putamen were reversed, and in which PMd activity modulated the connection between the visual areas and the putamen (**Figure 4.5C**).

4.2.4.5.2 Time series extraction

Since the exact locations of activation maxima varied across participants, we ensured the comparability of our models across participants by combining anatomical and functional constraints in selecting the subject-specific time series (cf. (Stephan et al., 2007c)). In brief, a regional time series was extracted if (i) it passed a threshold of $P<0.05$ (uncorrected) and (ii) was located within the same anatomical structure as and within a certain radius from the group maximum. For FFA and PPA (identified by the contrast testing for main effect of faces vs. houses, F>H and H>F, respectively) the individual maxima were required to be within an 8 mm radius around the group maxima. For the putamen and PMd (identified by the contrast testing for a [negative] main effect of probability) the individual maxima were required to be within a 16 mm radius around the group maximum (PMd) and within the putamen as defined by the participants' individual structural scan. As a summary time series, we computed the first eigenvector across all supra-threshold voxels within a radius of 4 mm around the chosen local maximum. Overall, following this

procedure, we were able to extract time series for all four areas in 15 out of 20 participants. We could not obtain a putamen time series in three participants and a PMd time series in two participants due to the lack of an activation that met the anatomical and functional criteria described above. Since we could not specify the complete model in these participants, they were excluded from the DCM analysis.

4.3 Results

4.3.1 Behavioural data

On average, subjects responded correctly on 91 ± 3.4% (mean ± SD) of the trials, on 5% of the trials they gave the wrong response or pressed multiple buttons, and on the remaining 4% they did not respond before the end of the trial.

Averaging reaction times (RT) across the blocks of different association levels showed that subjects did learn the changing contingencies, such that subjects responded faster to more likely outcomes (**Figure 4.3A**). The difference in average RT between unexpected (p = 0.1) and expected outcomes (p = 0.9) across subjects was 32 ms (**Figure 4.3A**). However, the Kolmogorov-Smirnov test for normality showed that the RT distributions differed significantly from a normal distribution in 13 out of 20 subjects; they were skewed towards the larger RT ($P<0.05$, Bonferroni corrected). However, in accordance with previous work (e.g. Carpenter & Williams 1995), response speed (RS) distributions were not significantly different from normal in all but 4 subjects, and therefore these were used for further analysis.

A repeated measures ANOVA significantly refuted the null hypothesis that RS did not differ across experimental conditions ($F(2.4; 45.4)=43.9$; $P<0.001$). A post hoc *t*-test showed that RS increased linearly with the probability of the outcome target ($P<0.001$, **Figure 4.3A**). Furthermore, subjects responded slightly faster to faces than to houses, ($P<0.05$), and slightly faster to trials with a high-frequency CS compared to trials with a low-frequency CS ($P<0.05$). However, in neither case was there an interaction with probability ($F=1.03$; $P=0.4$ and $F=0.69$; $P=0.6$), nor a threeway interaction between all factors ($F=1.17$, $P=0.33$), showing that there was no differential learning for the different event types. A repeated measures ANOVA also rejected the null hypothesis that error rates did not differ across experimental

conditions ($F(1.5; 334.0)=12.52$; $P<0.001$). Again, there was a significant main effect of probability ($P<0.001$), such that subjects made more errors to more unexpected outcomes (**Figure 4.3B**).

Finally, we used BMS to decide whether the trial-by-trial probability estimates of the Bayesian learning model or the true (but unknown) probabilities that had generated the stimulus sequence were better linear predictors of the RS. The distribution of the log evidences across subjects (**Figure 4.3C**) and the subsequent BMS at the group level indicated that the Bayesian learning model was vastly superior: the exceedance probability in favour of the Bayesian learning model was 100% (**Figure 4.3D**).

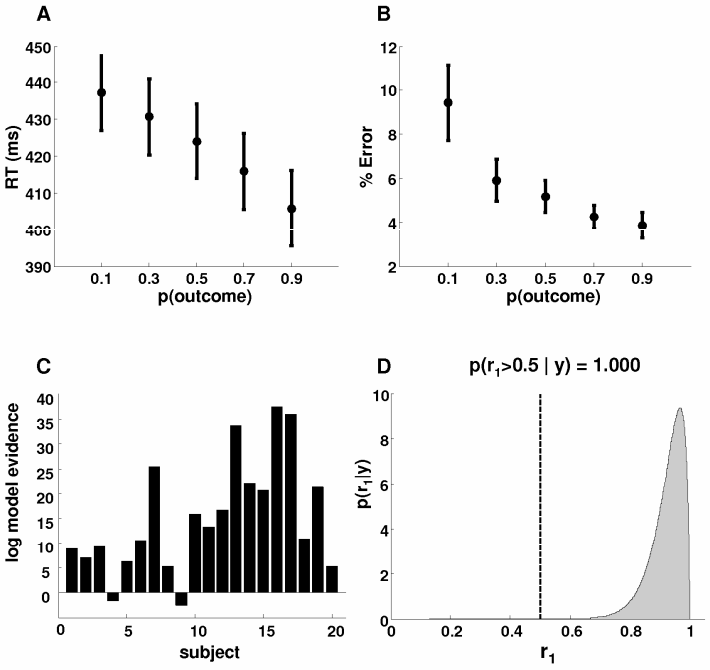

Figure 4.3. The effect of outcome probability on RTs and error rates. RTs **(A)** and percentage of errors **(B)** are shown as a function of outcome probability (mean ± standard error (SE)). Correct trials were averaged within each level of probability and collapsed across CS and visual outcome type (F/H). Subjects speed up and make fewer errors the higher the probability of the outcome. **(C)** Difference in log model evidence for using the trial-by-trial probability estimates from the Bayesian model versus the true probabilities as linear predictors for behavioural measured response speeds. In all but two subjects, there is greater evidence for the Bayesian model. **(D)** The Dirichlet density describing the probability of model m_1 (based on the probability estimates from the Bayesian learning model) relative to the alternative model m_2 (based on the true, blocked probabilities), given the measured response speeds across the group. The shaded area represents the exceedance probability of m_1 being a more likely model than m_2. This exceedance probability of Φ_1 = 100.0% was strongly favouring m_1 as a more likely model than m_2.

4.3.2 Analyses of fMRI data

The main results of our SPM analysis are summarised graphically in **Figure 4. 4**. Note that panels A, B and F show the results of a whole-brain analysis, whereas panels C, D, E and G result from region of interest analyses that were either defined by orthogonal localiser contrasts (panels C and D) or an anatomical mask (panel E,G); see the **Section 4.2** for details.

The key questions of interest for this study is characterization of stimulus-independent and stimulus-specific surprise responses and connectivity analyses. For completion, the results of additional analyses are reported, including a detailed analysis of the main effects of the stimuli as well as an analysis of regional responses associated with the volatility of the probabilistic associations. Although the use of a volatile environment was not a phenomenon of primary interest for this study, but merely a means of enforcing continuous learning (and thus maximising induction of synaptic plasticity and hence connectivity changes), it is noteworthy that our analysis of volatility effects replicated previous results by Behrens et al. (2007).

4.3.2.1 Stimulus main effects in FFA and PPA

As expected, the mid fusiform gyrus was activated more strongly to *face* stimuli than to *house* stimuli (FFA, **Figure 4.**4A and **Table 4.1**), and the parahippocampal gyrus showed the opposite effect (PPA, **Figure 4.4B** and **Table 4.1**). In the random effects analysis main effect for houses in the PPA has a much greater spatial extent than the main effect for faces in the FFA. This is possibly due to the greater variability in the location of the FFA than the PPA: At the group level the FFA activation is significant at whole brain corrected level only in the right hemisphere, but at the left FFA is significant within an ROI for the fusiform gyrus (**Table 4.1**).

4.3.2.2 Regional responses reflecting stimulus-independent surprise

Activity in the bilateral putamen decreased significantly with increasing probability of the visual stimulus, regardless whether face or house stimuli were presented (**Table 4.1** and **Figure 4.4E**). In other words, putamen activity increased the more surprising the presented stimulus was, given the trial-by-trial probability estimates of the Bayesian learning model. Several areas that are involved in preparation of motor responses showed equivalent stimulus-independent surprise-related responses. These

included the left dorsal premotor cortex (PMd), right intraparietal sulcus and right superior parietal gyrus (**Table 4.1** and **Figure 4.4F**). The homotopic counterparts of these areas in the opposite hemisphere also showed increased responses to surprising stimuli, but these activations did not survive whole brain correction (**Table 4.1**).

4.3.2.3 Surprise-related responses in stimulus-specific areas

Using the main effect of stimuli, we functionally defined FFA and PPA in each participant (for details on group main effects, see supplementary material). In each subject, we then determined the peak voxels in right FFA and right PPA that showed maximally selective face and house responses, respectively, and tested for (orthogonal) stimulus × probability interactions, i.e. a difference in the modulation of stimulus-specific responses by the probability of that stimulus occurring. In the FFA, there was a pronounced negative modulation of its responses to faces by the trial-by-trial probability estimates for faces (β = -2.05 ± 0.52). In other words, FFA responses to faces *increased* with the magnitude of prediction error, i.e. the more surprising the occurrence of a face was. In contrast, the modulation of FFA responses to houses by the trial-by-trial probability estimates for houses was marginal (β = -0.09 ± 0.78; see **Figure 4.4C**). This interaction was significant (p = 0.037).

When examining activity in PPA, we found that its responses to houses showed a strongly negative modulation by the trial-by-trial probability estimates for houses (β = -2.29 ± 0.54; see **Figure 4.4D**). That is, in analogy to the FFA results, PPA responses to houses *increased* the more surprising the presentation of a house was. In contrast, PPA responses to faces were positively modulated by the trial-by-trial probability estimates for faces (β = 1.91 ± 0.67); this corresponds to a decrease in activity the more surprising the presentation of a face was. Again, as for FFA, this interaction was significant (p < 0.001).

In summary, responses of PPA and FFA to their preferred stimuli were strongly modulated by surprise (or prediction error) about these stimuli (i.e. showed a negative modulation by trial-by-trial probability estimates for these stimuli as provided by the Bayesian observer model), and this modulation by surprise was significantly higher than for their non-preferred stimuli.

4.3.2.4 Volatility dependent brain activations

For completion, we also tested in which areas activity increased or decreased with the trial-by-trial volatility estimates. Following the results by Behrens et al., (2007), who demonstrated that ACC activity correlated with volatility estimates during reward learning, we tested whether volatility encoding in the ACC would also be present in our learning paradigm which did not include any rewards. Indeed, activity in the dorsal and rostral ACC and the ventromedial prefrontal cortex correlated significantly with the volatility estimates (**Table 4.1** and **Figure 4.4G**).

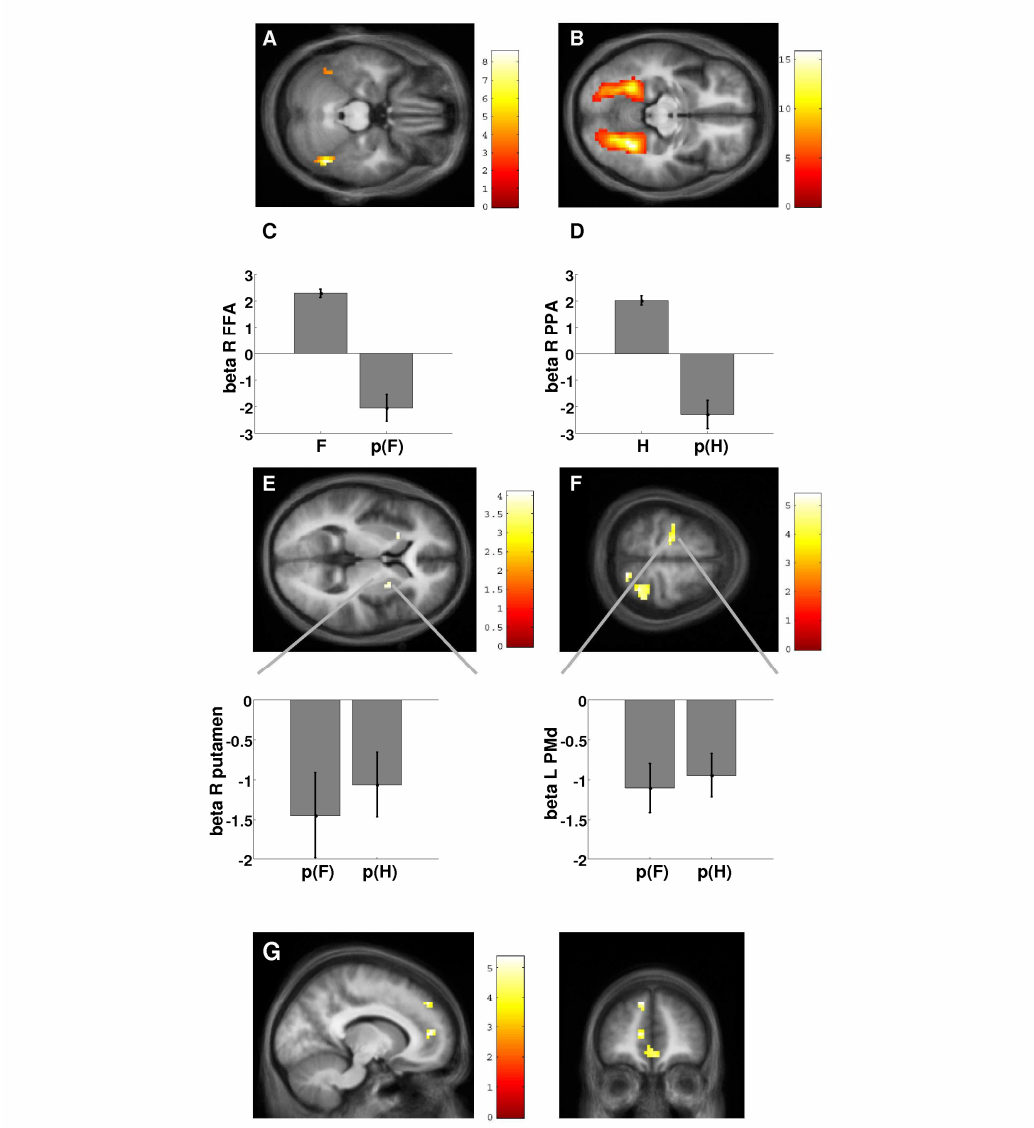

Figure 4.4. Main effects and modulation of outcome stimulus processing. All parameter estimates show mean ± SE across all subjects and all activations are displayed on the average anatomical scan. **A)** Main effect of F>H in the right FFA, also showing the left FFA activation (see supplementary material). **B)** Main effect of H>F in the bilateral PPA. **C)** Parameter estimates across subjects (located at the individual maxima in the F>H contrast in the FFA) of the modulatory effect of stimulus probabilities. There was a pronounced negative modulation of FFA responses to faces by the trial-by-trial probability estimates for faces (β = -2.05 ± 0.52). In contrast, the modulation of FFA responses to houses by the trial-by-trial probability estimates for houses was marginal (β = -0.09 ± 0.78). This interaction was significant (p = 0.037). **D)** Parameter estimates across subjects (located at the individual maxima in the H>F contrast in the PPA) of the modulatory effect of stimulus probabilities. PPA responses to houses showed a strongly negative modulation by the trial-by-trial probability estimates for houses (β = -2.29 ± 0.54). In contrast, PPA responses to faces were positively modulated by the trial-by-trial probability estimates for faces (β = 1.91 ± 0.67). This interaction was significant (p < 0.001). **(E, top)** Bilateral effect of surprise in the anterior putamen. **(Bottom)** Parameter estimates from the putamen showing the negative dependency on both p(F) and the p(H). **(F, top)** Bilateral effects of surprise in dorsal premotor cortex (PMd) and the parietal cortex. **(Bottom)** Parameter estimates for the left PMd, showing the same surprise dependent effect as the putamen. **G)** Parametric modulation of the VMPFC/ ACC by volatility.

Table 4.1. MNI coordinates and Z-values for significantly activated regions.

	MNI coords.			
Foci of activation	X	y	z	Z score
Surprise effects: negative correlation with p(F) and p(H)				
Motor areas				
L precentral gyrus (dorsal premotor cortex)*	-18	-18	60	4.13
R precentral gyrus (dorsal premotor cortex)**	33	-15	57	3.40
R intraparietal sulcus*	42	-33	39	4.02
L intraparietal sulcus **	-42	-39	39	3.72
R superior parietal gyrus*	15	-60	63	4.16
L superior parietal gyrus**	-15	-57	63	3.42
Striatum				
R putamen**	27	3	6	3.42
L putamen**	-24	15	3	3.39
Probability effects: positive correlation with p(F) and p(H)				
No significant activations.				
Volatility effects: positive contrast				
Ventromedial prefrontal ctx*	3	48	-9	3.64
ACC**	-12	45	9	4.11
Ventral ACC / subgenual ctx **	-6	36	-3	3.57
L caudate/thalamus*	-21	-9	9	4.32
Volatility effects: negative contrast				
No significant activations.				
Main effects of sensory stimulation				
House>Face				
R parahippocampal gyrus*	30	-51	12	7.01
L parahippocampal gyrus*	-24	-57	-18	6.70
Face> House				
R mid fusiform gyrus*	45	-57	-24	5.42
L amygdala*	-21	-12	-9	4.31
L mid fusiform gyrus**	-45	-54	-21	3.47

* significant at P<0.05 FWE cluster-level corrected across the whole-brain

** significant at P<0.05 cluster-level corrected for a priori region of interest

4.3.3 Nonlinear DCM

Based on our SPM results, we constructed a nonlinear DCM including the right putamen, PPA and FFA, and the left PMd as parsimonious model for testing whether surprise activity in the putamen gates visuomotor connections. This initial DCM included connections from FFA and PPA to PMd and a modulatory influence from the putamen on these connections (**Figure 4.5** - m1), and thus included the minimal number of connections necessary to test the hypothesis. All additional models were derived by expanding this basic architecture. Hierarchical BMS was then used to select the optimal model at the group level (Stephan et al., 2009).

In a first step, all possible combinations of endogenous connections between the PPA, FFA and PMd were compared using Bayesian model comparison (**Figure 4.5 -** m_{1-4}). Compared to all other models, there was greater evidence for the model with full connectivity between all these areas (**Table 4.2**). Model 4, including full reciprocal connectivity between the sensory and premotor areas was clearly the best model (exceedance probability $\phi_4 = 0.99$).

In a subsequent step, two more models were tested to look at endogenous connections to the putamen from the sensory and premotor cortex. Firstly a model was tested in which connections from the sensory areas to the putamen were included, to test whether there was any direct influence of these areas on the putamen (**Figure 4.5** - m_5). This model turned out to be worse than the model that did not include these connections ($\phi_4 = 0.99$). Secondly, because there are known to be direct projections from the premotor cortex to putamen (Leh et al., 2007;Takada et al., 1998), m_6 included a direct connection from the PMd to the putamen (**Figure 4.5** - m_6). Here, the evidence for m_4 was still greater than for m_6, although less decisively than for the other models ($\phi_4 = 0.64$). Note that this does not mean that this connection does not exist anatomically, but just that it is unlikely to play a major role in the process modelled here.

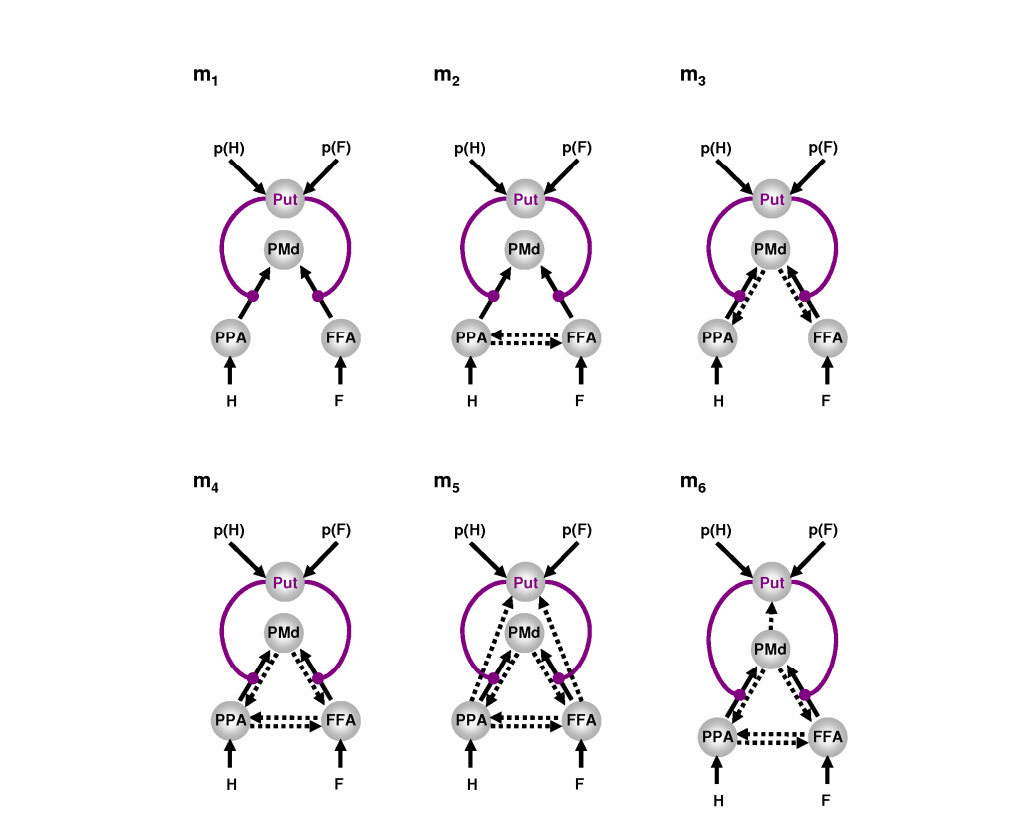

Figure 4.5. DCMs tested to establish the optimal endogenous connectivity. Set of 6 DCMs testing the hypothesis that the putame modulates connectivity between the sensory and motor cortices, designed to establish the optimal endogenous connectivity. The dotted lines are the connections that are included in addition to the most parsimonious model m_1. m_4 was the optimal model (see main text).

Table 4.2. BMS with regard to endogenous connectivity between PPA, FFA and PMd

	Dirichlet parameters α	**Exceedance probability ϕ**
m_1	1.79	0.00
m_2	5.79	0.15
m_3	1.81	0.00
m_4	9.62	0.84

Thus, the optimal model was a model with reciprocal connections between PMd, FFA and PPA (see **Figure 4.6A**). In this model, the parameter estimates that

described gating effects of putamen activity on visuo-motor connections, were consistently positive across subjects (PPA→PMd: $d = 0.01 \pm 0.003$ (mean ± SE), $p = 0.010$; FFA→PMd: $d = 0.011 \pm 0.004$, $p = 0.017$). Therefore, in accordance with our hypothesis, prediction error related activity in the putamen significantly modulated the strength of visuo-motor connections.

However, because putamen and PMd showed similar surprise–related activations (cf. **Figure 4.4**), it was necessary to demonstrate the specificity of our model and exclude the possibility that, instead of putamen activity gating visuo-motor connections, the PMd might be gating visuo-putamen connections. Therefore a final crucial model comparison was made to verify the directionality of the putamen influence. In this model (m_7) the role of the putamen and the PMd were swapped, such that PMd activity modulated the connection between FFA/PPA and the putamen (**Figure 4.6C**). BMS showed that this reversed model was clearly inferior to the original model; the exceedance probability that the data were more likely to have been generated by the original model rather than by the reversed model, was 99% (**Figure 4.6D**). Finally, **Table 4.3** we report a final comparison of all models at once, showing once more that m_4 is the optimal model.

Table 4.3. BMS among all tested DCMs

	Dirichlet parameters α	**Exceedance probability ϕ**
m_1	1.55	0.02
m_2	3.60	0.16
m_3	1.57	0.02
m_4	4.82	0.36
m_5	2.83	0.08
m_6	4.12	0.23
m_7	3.50	0.14

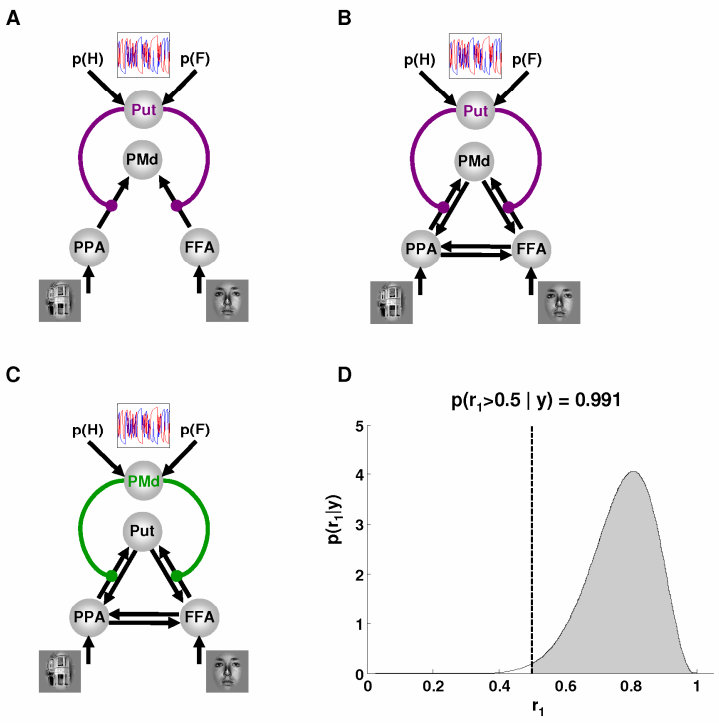

Figure 4.6. DCMs testing the respective roles of putamen and PMd. **A)** A basic DCM (cf m_1, Figure 4.5) for investigating modulation of visuo-motor connections by prediction error related activity in the putamen. **B)** The optimal DCM (cf m_4, Figure 4.5), resulting from a systematic model search procedure, included full connectivity between the PMd, PPA and FFA. Activity in the putamen significantly enhanced the connections from the PPA/FFA to the premotor cortex: p = 0.010 and p = 0.017 for the modulation of the PPA→PMd and FFA→ PMd, respectively. **C)** Alternative DCM in which the roles of the putamen and the PMd were swapped (cf m_7, Table 4.3). **D)** The Dirichlet density describing the probability of model m_4 (panel B) relative to the alternative model m_7 (Panel C), given the measured fMRI data across the group. The shaded area represents the exceedance probability of m_4 being a more likely model than m_7. This exceedance probability of Φ_1 = 99.1% was strongly favouring m_4 as a more likely model than m_7.

4.4 Discussion

In this study, we used an associative learning paradigm in which auditory cues differentially predicted subsequent visual stimuli (faces or houses) to which subjects made a speeded response. We ensured that on any given trial the *a priori* probability of a face (or house) occurring was always 50%. Thus, any expectations about the visual stimulus were entirely dependent on the auditory cue. Critically, the predictive strengths of cues were unknown and varied over time, requiring subjects to continuously update their estimates of cue-stimulus associations and thereby maximising demands on changes in network connectivity via synaptic plasticity. We modeled this dynamic inference process using a hierarchical Bayesian observer that inferred the associations from the observed cue-outcome combinations, taking into account the volatility of the environment (Behrens et al., 2007). These trial-by-trial probability estimates were subsequently used as predictor variables in the analysis of both behavioural and fMRI data. Behaviourally, speed and accuracy of motor responses significantly increased with trial-by-trial predictability of visual stimuli (**Figure 4.3**). Analysis of the fMRI data showed that FFA and PPA reflected prediction errors that were specific for their preferred stimulus (**Figure 4.4A,B**). In contrast, both the putamen and dorsal premotor cortex represented stimulus-independent prediction errors in that their activity increased the more surprising the current visual stimulus was regardless of its type (**Figure 4.4E,F**). Comparing a series of nonlinear DCMs by Bayesian model selection, we found that the activity in dorsal premotor cortex was best explained by a model in which prediction error related activity in the putamen enhanced the strength of connections from FFA and PPA to premotor cortex by a non-linear gating mechanism (**Figure 4.5**).

Two recent studies have shown that during learning of stimulus probabilities visual areas show increased responses to unexpected visual outcomes (den Ouden et al., 2009;Summerfield and Koechlin, 2008). Both studies, however, only used a single, and relatively unspecific, stimulus type (squares and gabor patches, respectively). It thus remained unclear whether this represented a general, stimulus-independent or a stimulus-specific surprise response. Moreover, the probabilities used remained stationary throughout both studies. An additional limitation of the previous study (den Ouden et al., 2009;Summerfield and Koechlin, 2008) was that it investigated incidental learning of stimulus associations and could thus not provide direct

behavioural evidence for learning. All of the above limitations were avoided by the design of the current study.

Our present results show a double dissociation among face- and house-specific areas that represents a stimulus-specific surprise response (**Figure 4.4A-D**). While FFA responses to faces increased with the magnitude of prediction error, i.e. the more surprising the occurrence of a face was, its responses to houses were unaffected by prediction error. In PPA, responses to houses increased with the magnitude of prediction error whereas responses to faces even decreased with prediction error. In both cases, this stimulus × probability interaction was significant.

In contrast to the visual areas, the bilateral putamen, left dorsal premotor cortex, right intraparietal sulcus and superior parietal gyrus showed a stimulus-independent prediction error response (**Figure 4.4F**). That is, whenever an unexpected stimulus was presented, independently of whether this was a face or house, the activity in these areas increased. The parietal activations are co-extensive with the dorsal visual stream and play an important role in attentional reorientation (Corbetta and Shulman, 2002). Their increased activity in response to surprising stimuli may therefore reflect increased attention to the unexpected visual stimuli. In contrast, the surprise-related activity in the premotor cortex is more likely to reflect the updating of the motor plan that becomes necessary when the prediction evoked by the auditory cue turns out to be wrong (Mars et al., 2007;Nakayama et al., 2008). Finally, prediction error responses in the putamen (**Figure 4.4E**) have been reported by numerous previous studies, and for very different types of learning. This suggests that the putamen is generally sensitive to violations of learned contingencies, whether these contingencies signal reward (Jensen et al., 2007;McClure et al., 2003;Menon et al., 2007;O'Doherty et al., 2004;O'Doherty et al., 2003;Seymour et al., 2004), guide decision making (Corlett et al., 2004), or predict target stimuli (as in the current study), and even when these contingencies are not behaviourally relevant at all (den Ouden et al., 2009).

The above considerations imply that the increase of premotor activity for surprising visual outcomes could at least partially be due to a re-weighting of stimulus-specific visual inputs that is controlled by the degree of prediction error encoded by activity in the putamen. In other words, the strength of connections from FFA and PPA to

premotor cortex, which provide information about the appropriateness of the planned action, might change from trial to trial, depending on the mismatch between predicted and observed visual outcome that is signalled by the putamen. To address this hypothesis, we used a recently developed nonlinear DCM (Stephan et al., 2008), which allowed us to model how connections from FFA and PPA to premotor cortex were modulated or gated by ongoing activity in the putamen. Anatomically, there are indirect projections from the putamen to the premotor cortex via the thalamus (Alexander and Crutcher, 1990;Schultz, 2000) which could mediate this gating process. To demonstrate the directionality of this mechanism, we compared this type of model to a control model in which the role of the putamen and premotor cortex were reversed, i.e. the connections from visual areas to the putamen were now modulated by premotor activity (**Figure 4.6C**). Bayesian model selection showed that the original model was clearly superior to the alternative one (**Figure 4.6D**).

Previous neurophysiological and neuroimaging investigations of associative learning have focused on identifying region-specific prediction error responses, e.g. in the ventral tegmental area (D'Ardenne et al., 2008;Yacubian et al., 2006) or the striatum (Corlett et al., 2004;Jensen et al., 2007;McClure et al., 2003;Menon et al., 2007;O'Doherty et al., 2004;O'Doherty et al., 2003;Schultz and Dickinson, 2000;Seymour et al., 2004;Tobler et al., 2006), but have not investigated effects of prediction errors on connectivity. An exception was the precursor to the present study (den Ouden et al., 2009). As in the present study, this previous work used an audio-visual associative learning paradigm and found a prediction error response in the putamen bilaterally and in visual cortex. However, the DCM in this previous study only described an anatomically uninformed influence of prediction errors per se on connectivity, but did not specify their source, because the required nonlinear models where not yet established at the time of analysis.

To our knowledge, the present study is the first to demonstrate that trial-by-trial prediction error related activity in a specific region (here the putamen) controls the plasticity of connections among other regions. This is in accordance with several theoretical concepts which have proposed that learning should be implemented neurophysiologically by prediction error dependent synaptic plasticity (Friston, 2005a;McLaren et al., 1989;Schultz and Dickinson, 2000). Deploying synaptic plasticity depending on the magnitude of prediction error is an intuitively sensible

mechanism: the larger the prediction error, the greater the need for changing one's predictions and hence to reorganise the neuronal system producing these predictions. The model in the present study represents prediction error dependent learning as a nonlinear (second order) interaction between outputs from FFA/PPA and putamen that target the dorsal premotor cortex. Several neurobiological mechanisms for this type of plasticity have been suggested by invasive recordings studies, including nonlinear dendritic integration of inputs due to voltage-dependent ion channels or activation of dendritic calcium conductances by back-propagating action potentials (for details and references, see (Stephan et al., 2008)).

In summary, the present study has used a combination of fMRI, computational learning models and DCM to demonstrate that learning-induced synaptic plasticity in the human brain during a simple audio-visual association learning task can be characterized in terms of prediction error dependent changes in effective connectivity. Such approaches may become useful for model-based inference about neurophysiological processes that cannot usually be studied non-invasively but are of clinical importance, such as synaptic plasticity and its regulation by neuromodulatory transmitters (Stephan et al., 2006). An important future step will be to combine model-based approaches as the present one with pharmacological designs that manipulate prediction error dependent changes in plasticity.

Chapter 5

Amygdala Modulates Cortico-Striatal Connections During Fear Acquisition

Abstract

This DCM study is based on a dataset which has previously been analysed using conventional statistical parametric mapping by Petrovic et al. (Petrovic et al., 2008). In the original study the authors focussed on the role of the fusiform gyrus and the amygdala in processing of learned affective values for faces. In the current study the acquired fMRI data were reanalysed, focussing on the role of the amygdala in CS+ processing. In this reanalysis, using DCM and BMS, we compared different putative mechanisms of amygdala involvement in learning the CS+US association. Specifically, we investigated amygdala-dependent gating of corticostriatal connections during processing of CS+ stimuli.

5.1 Introduction

A wealth of research in both animals and humans have identified a critical role of the amygdala in Pavlovian fear learning, in which an affectively neutral conditioned stimulus (CS) is presented with an aversive unconditioned stimulus (US), such as an electric shock (e.g. see(LeDoux, 2003;Maren, 2001)). After a number of paired presentations, a CS alone elicits a fear response, such as freezing or increased sweating. Amygdala damage in both humans and animals results in severely impaired fear conditioning (Bechara et al., 1995;Blair et al., 2005;LaBar et al., 1995). For example, LaBar et al showed reduced conditioned skin conductance responses to a CS associated with a loud noise burst in unilateral temporal lobectomy patients with temporal lobe epilepsy patients (LaBar et al., 1995).

The exact details of the physiology of fear learning mechanisms in the amygdala are yet to be elucidated, but it is generally agreed that sensory information from the cortex and thalamus is received by the basolateral part of the amygdala (Delgado et al., 2006). The lateral part especially is considered as the 'gatekeeper' to the amygdala (LeDoux, 2007). The underlying mechanism of the CS+ induced activity might be as follows: CS-US convergence induces synaptic plasticity in lateral amygdala such that after conditioning CS information is conveyed more effectively by the lateral amygdala, via intra-amygdala connections, to elicit activation in the central amygdala. The central amygdala is normally only activated by behaviourally relevant USs, as it interfaces with the motor system and prefrontal areas, controlling the expression of conditioned behavioural and autonomic fear responses (LeDoux, 2007).

In humans, a number of studies have shown differential responses in the amygdala to stimuli that have been associated with an aversive outcome (CS+), compared to stimuli that do not predict an aversive stimulus (CS-). However, while all studies show overall increased activity in the amygdala to the CS+ compared to the CS- (e.g. (Buchel et al., 1998;LaBar et al., 1998;Marschner et al., 2008;Tabbert et al., 2005)), the precise timecourse of these differential responses is variable. Several studies have reported an initial increase in the response to CS+ stimuli in the amygdala, followed by a later decrease (Buchel et al., 1998;LaBar et al., 1998;Marschner et al., 2008) whereas some studies show a differentiation between CS+ and CS- during the late acquisition phase, such as a study by Tabbert et al. where differentiation between CS+ and CS− within the amygdala was observed during a late acquisition phase (Tabbert et al., 2005). One explanation for these inconsistent results could be that the speed of learning differs between these different paradigms.

Learning speed will depend on several features of the conditioning paradigm, including the similarity of the CS+ and CS-, the aversiveness of the US, and the reinforcement schedule. Easily differentiated CSs with deterministic reinforcement schedules are likely to induce very fast learning, and might show differentiation within the first couple of trials (e.g. (Marschner et al., 2008)). A complementary explanation could be that two processes evolve simultaneously in different parts of the amygdala that cannot be resolved at the spatial resolution of human fMRI. Results in favour of this suggestion come from rodent studies, in which single cell

recordings in the dorsal subnucleus of the lateral amygdala in rats during fear conditioning showed differential responses in two distinct cell populations (Repa et al., 2001). Cells in the dorsal tip of lateral amygdala, exhibited short-latency responses (<20 ms) that were only transiently changed. Cells in the more ventral part of the lateral amygdala, had longer latency responses, but maintained enhanced responding throughout training. This sustained response could reflect the fear/anxiety induced by the CS+, whereas as Marschner et al. suggest, the transient signal is reminiscent of prediction error response (Marschner et al., 2008): The amygdala encodes sensory contingencies to rapidly learn CS–US associations, so that when the CS is first paired with the US, this surprising stimulus elicits a large response, but when the CS-US association is subsequently learned, little response is elicited and the response decreases. However, to fully test for such a prediction error response, one would have to test not only the learned response to a presented US, but also to its absence. In the study by Marschner et al. this was not possible because the CS+ was always followed by a US.

The study presented here is a reanalysis of a previously published dataset (Petrovic et al., 2008). We used a refined model for the initial SPM analysis, obtaining results that go beyond those reported by Petrovic et al. and support the prediction error hypothesis proposed by Marschner et al. (Marschner et al., 2008). In a subsequent step, we used DCM to investigate the mechanisms of how fear learning in the amygdala can influence corticostriatal processing of conditioned stimuli. It has long been thought that the amygdala guides and initiates motor responses to affective stimuli, with the ventral striatum playing a pivotal role as the interface between the extended amygdala and motor systems coordinating responses to conditioned stimuli (e.g. (Haber et al., 1995;Mogenson et al., 1980)). More specifically, it has been suggested that output from the basal amygdala to the striatum controls actions in response to conditioned stimuli ((LeDoux, 2007). In the present study, we therefore compared a set of nonlinear DCMs embodying different mechanisms how the amygdala might mediate the processing of sensory information in the striatum and prefrontal cortex.

5.2 Methods & Statistical analysis

5.2.1 Experimental Design – fMRI

The dataset used in the current study was previously analysed by Petrovic et al. who used a conventional SPM analysis to investigate differential CS processing for social stimuli(Petrovic et al., 2008). While being scanned using fMRI, the participants were subjected to a fear conditioning paradigm where two visually presented faces were associated with aversive electric shocks in a 50% reinforcement schedule. Two further faces were presented but never associated with a shock (i.e. 0% contingency reinforcement schedule). For each of the CS types, one of the faces directly looked into the camera, whereas for the other stimulus the gaze was averted (see **Figure 5.1**). Since Petrovic et al. did not find very strong behavioural or fMRI effects for gaze direction, this factor was neglected in the present reanalysis. The basic design thus had a 2x2 factorial structure, the factors being CS type (CS+: 50% reinforcement; CS-: 0% reinforcement) and trial outcome (US present vs. absent). However, due to the use of a 0% reinforcement schedule there are, by definition, no CS- US trials; therefore, the design included only 3 different trial types. Additionally, we focused on learning effects, distinguishing between trials in the first vs. the second half of the experiment. This resulted in 6 trial types overall.

The subjects were instructed for each presented face to decide as quickly as possible whether the face was in the centre or offset (by 5 mm) to the left or right of the visual field. On any given trial, the face appeared for 990 ms, with a stimulus-onset asynchrony (SOA) that was jittered between 10.8 and 14.4 seconds for each trial (see **Figure 5.1**). For the CS+US trials, the electric shock was delivered at the end of the face presentation, for a duration of 1 ms at 80% of the most painful sensation imaginable as determined by a visual analogue scale (for details see (Petrovic et al., 2008)). Each subject completed 30 trials of each of the four CSs, being exposed to a total of 30 shocks. The total scanning duration was 24 min.

Although the majority of the subjects (n = 20/27) could not correctly identify which faces were associated with the shocks in a post-experimental interview, both skin conductance responses (SCR) and explicit ratings showed that the cue-shock association had been learned; in the second half of the experiment the skin

conductance response (SCR) to (unpaired) CS+ stimuli was larger than to CS- stimuli, and subjects rated CS- faces as more likeable, and CS+ faces as less likeable, after the experiment than before. Furthermore, using a reinforcement learning model to specify regressors, analysis of the fMRI data showed increasing activity in both amygdala and fusiform gyrus in response to the CS+ stimuli (Petrovic et al., 2008).

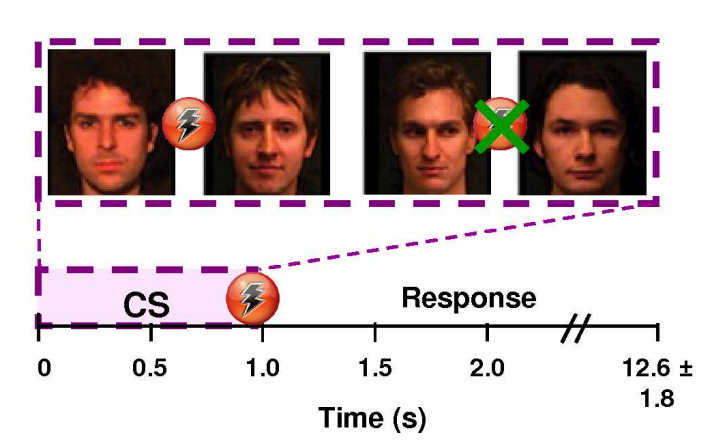

Figure 5.1. Timeseries of a single trial. The CS stimuli were presented for 990 ms, and in case of the CS+ stimuli, these were followed in 50% of the cases by a painful electric shock at the end of the CS presentation. Subjects performed an offset detection task on the CS stimuli.

5.2.2 Subjects

27 healthy male subjects (aged 18-36 years) with no history of neurological or psychiatric disorder were included in this study. Written informed consent was obtained from all volunteers prior to the study, which was approved by the National Hospital for Neurology and Neurosurgery Ethics Committee.

5.2.3 fMRI Data Acquisition

A 1.5 Tesla Siemens Sonata MRI scanner (Siemens, Erlangen, Germany) was used to acquire T1-weighted fast-field echo structural images and multi-slice T2*-weighted echo-planar volumes with blood oxygenation level dependent (BOLD) contrast (TR = 3.96 s, TE = 50 ms). For each subject, 360 scans were acquired in one continuous session. The first four volumes of each session were discarded to allow for T1

equilibrium effects. We used a 30° tilted orbitofrontal sequence (Deichmann et al., 2002) with a flip angle of 90° covering the whole brain in 44 slices.

5.2.4 fMRI Data Analysis

fMRI data were analysed using the statistical software package SPM5 (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm). The 356 images from each subject were realigned to correct for head movements, corrected for movement-by-distortion interactions (Andersson et al., 2001), spatially normalized to the Montreal Neurological Institute (MNI) template brain and smoothed spatially with a 3-dimensional Gaussian kernel of 8 mm full width half maximum. The data were then modelled voxel-wise, using a GLM that included regressors for six experimental trial types (described below) consisting of trains of delta functions convolved with the canonical hemodynamic response function. The data were high-pass filtered (cut-off 128 seconds) to remove low-frequency signal drifts, and a first-order autoregressive model was used to remove serial correlations (Friston et al., 2002a). In distinction to the previous analysis (Petrovic et al., 2008), we used a more precise temporal model of the stimulus onsets, and a more appropriate microtime bin for defining regressors (minimising the overall timing error across slices). Contrast images of parameter estimates encoding effects of interest were created for each subject and entered separately into voxel-wise one-sample *t*-tests (df = 26), to implement a second-level random effects analysis. We report regions that survive cluster-level correction for multiple comparisons (family-wise error, FWE) across the whole brain at $P<0.05$, or for predefined regions of interest (small volume correction, SVC) at $P<0.05$. These regions of interest included the amygdala, striatum and the prefrontal cortex.

5.2.5 SPM contrasts

In order to assess these learning effects at a neurophysiological level, the three different trial types (see **Table 5.1**) were split between the first and second half of the experiment, following previous fear conditioning studies, (Buchel et al., 1998;LaBar et al., 1998), resulting in 6 regressors.

The contrasts used in the current analysis are described in **Table 5.2**. In order to assess the main effect of ‘face’ stimulation, the regressors for CS+USC were not

used in order not to contaminate our results with shock events. Because 'face' stimuli are known to activate the fusiform gyrus (e.g. see **Chapter 4**), we will refer to the face responsive part of the fusiform gyrus as the fusiform face area (FFA) for simplicity. One should keep in mind, however, that the nonspecific nature of the contrast ('face' vs 'fixation') prevents any strong claims about the degree of specificity for face stimuli exhibited by the identified area.

The main effect of pain contrast was restricted to CS+ trials, as CS- stimuli were never paired with a painful stimulus. The crucial contrast to test for learning effects was the change in response to the presence or absence of shocks over time. Since we were specifically interested in the roles of the striatum, amygdala and prefrontal cortex in fear learning, we performed an additional restricted search in these areas, using anatomical masks generated from the Anatomy Toolbox (amygdala, (Eickhoff et al., 2005)) and the PickAtlas toolbox (prefrontal cortex mask included the inferior, middle and superior frontal gyrus, and striatum, (Maldjian et al., 2003)). The Anatomy toolbox is a probabilistic cytoarchitectonic atlas based on histological investigation of a group of post mortem brains, whereas the PickAtlas is based on topographical landmarks alone. Thus, the Anatomy toolbox was used preferably.

Table 5.1. Design and stimulus frequency. Note that the design is not entirely factorial, because by definition there are no CS- US trials, resulting in 3 different trial types.

	shock (US)	**no shock**
CS+	30	30
CS-	0	60

Table 5.2. Contrast definnitions.

	early			**late**		
	CS+	**CS+US**	**CS-**	**CS+**	**CS+US**	**CS-**
main effect of 'face'	1	0	1	1	0	1
main effect of 'pain'	-1	1	0	-1	1	0
pain x time (+)	-1	1	0	1	-1	0
pain x time (-)	-1	1	0	1	-1	0

5.2.6 DCM

5.2.6.1 DCM specification

As described in the results section below and as shown in **Figure 5.2**, the SPM analysis demonstrated that amygdala, striatum and prefrontal cortex showed a time x shock interaction, such that the response to an unpaired CS+ over time increased, whereas the response to a CS+ paired with a shock decreased. Based on these SPM results, a set of alternative nonlinear DCMs were constructed that could all potentially account for the interactions observed in these areas. All DCMs additionally included the FFA as an input region that was driven by the face stimuli; the sensory effects of shocks entered the system via the amygdala (see **Figure 5.3**). These driving inputs were modelled as individual events. The direct input into the amygdala represents a lumped influence via three possible pathways since the basolateral amygdala receives noxious information from the insula, the thalamus and the parabrachial nucleus (Shi and Davis, 1999).

In order to reduce computational complexity, the DCM analysis proceeded in two steps. The first step was to determine the most likely mechanism, in terms of shock-induced modulation of connection strengths, for explaining the shock x time interaction in the amygdala. This was done using a reduced 3-area model that did not include the PFC. The second step was to investigate how shock x time interactions in striatum and PFC could be best explained in terms of nonlinear gating of connections to the striatum and prefrontal cortex by amygdala activity. This hierarchical approach was necessary for computational reasons: testing all relevant variants of the full 4-area DCM would have taken a very long time.

In the first step, three 3-area DCMs that all could explain the shock x time interaction in the amygdala, were constructed and fitted to the data (see **Figure 5.3A**). In the first model, the time modulation for both the paired and unpaired CS+ trials affected the self-connection of the amygdala; this model reflects learning that was occurring within the amygdala, such that over time the response to paired CS+ would be dampened and to unpaired CS+ stimuli it would be enhanced. In a second model, both trial types were allowed to modulate the connection from the FFA to the amygdala. This model reflects how transfer of the CS+ input to the amygdala changed over time, depending on whether it was paired with a shock or not. Finally,

in a third model the effects of the paired and unpaired CS+ were modelled separately: the unpaired CS+ modulated the FFA→amygdala connection, whereas the paired CS+ affected the intrinsic self connections of the amygdala. Note that the sensory effects of shocks entered the system via the amygdala; therefore the opposite arrangement (i.e. the paired CS+ modulating the FFA→amygdala connection) was not a sensible alternative.

In the second step, the 3-area model that was identified as optimal in the first step was extended to include the PFC as fourth region and was systematically varied along two dimensions. The main question of interest was to test systematically for modulatory (gating) influences on the FFA → Striatum and Striatum→ PFC connections that depended on amygdala activity (see **Figure 5.3B**). Either of these gating connections alone could potentially explain the observed shock x time interaction in the striatum and prefrontal cortex, via the reciprocal connections between these areas. Secondly, although the observed SPM results could in principle be explained without any connections from the prefrontal cortex and the striatum to the amygdala, there is evidence for such anatomical connections (e.g. (Haber and Fudge, 1997)) , and therefore we also tested whether inclusion of these connections improved the model. In summary, the 4-area models varied across two dimensions: (i) gating influences by the amygdala and (ii) prefrontal and striatal connections to the amygdala (see **Figure 5.3B)**.

5.2.6.2 Choice of areas and time series extraction

As the exact locations of activation maxima varied over subjects, we ensured the comparability of our models across subjects by using combined anatomical-functional constraints in selecting the subject-specific time series (cf. (Stephan et al., 2007c)). As a summary time-series, we computed the first eigenvector across all supra-threshold voxels within a radius of 4 mm around the chosen local maximum for the left FFA, left amygdala, right striatum and left dorsolateral prefrontal cortex (DLPFC). For the amygdala, we thresholded the subject-specific SPMs at $P<0.05$ (uncorrected) for the shock x time interaction and determined the local maximum within a mask combining the main effect of pain (at the same threshold) with an anatomical mask of the amygdala generated from the probabilistic cytoarchitectonic atlas in MNI space (Eickhoff et al., 2005). For the striatum and PFC, the subject

specific SPMs for the shock x time contrast were also thresholded at $P<0.05$. For the striatum the individual maximum within an anatomical mask (putamen, (Maldjian et al., 2003)) was determined, and for the PFC the maximum within 8 mm from the group maximum.

The left rather than the right prefrontal activation was included in the DCM based on the pattern of parameter estimates of the interaction; whereas the left PFC showed the same interaction pattern as the amygdala and striatum, the interaction in the right PFC was driven by the changing response to paired CS+ trials, and could be due to habituation to the pain response alone (see parameter estimates in **Figure 5.2**). Finally, for the FFA, the maximum within 8 mm of the group maximum from our previous study ([-45 -54 -21], see **Chapter 4**) was chosen. The reason to use the maximum from this previous study was the nonspecific nature of the 'face' contrast in this study; in the previous study a more specific (face>house) contrast was used. **Figure 5.2** shows the parameter estimates across subjects from the extracted areas. Time series could be extracted for all four areas in 16 out of 27 subjects. In the remaining subjects, one or more of the areas could not be defined due to the lack of a significant interaction that met the anatomical and functional criteria described above (amygdala: 6; FFA: 0; PFC: 3; striatum: 7). These subjects were excluded from the DCM analysis.

5.2.6.3 Model comparison

The optimal models for each of the two sets of DCMs (3- and 4-area model, see **Figure 5.3**) were determined using BMS. In brief, the negative free energy (F) approximation to the model evidence for each subject and each model were used to estimate the exceedance probability and Dirichlet parameters α. Given the factorial nature of the tested model space for the 4-area models, we were able to use model space partitioning in order to test for the effects of (i) varying the modulatory effects of the amygdala and (ii) including or excluding the prefrontal/striatal connections to the amygdala respectively.

5.2.6.4 Group level inference on parameters

Because in one subject the parameter estimates deviated by more than 3 standard deviations from the rest of the group, normality assumptions were violated, rendering

standard parametric statistical tests inappropriate. Therefore, to test whether the coupling parameters were consistently different from zero across subjects, we applied a nonparametric test (Wilcoxon's signed-rank test) and report Bonferroni corrected *p*-values.

5.3 Results

5.3.1 SPM results

As expected, whenever a face was presented, compared to baseline fixation, activity in the primary visual cortex and the fusiform face area increased (see **Table 5.3**). Furthermore, the main effect of pain showed widespread increases in activity in a collection of brain areas known as the 'pain matrix', including the insula and amygdala bilaterally, anterior cingulate cortex (ACC), brainstem, primary and secondary somatosensory cortex (S1 and S2) on the right (stimulation was on the left). The critical shock x time interaction contrast was significant in the amygdala, prefrontal cortex and the ventral striatum (putamen) such that over time responses to the unpaired CS+ increased but to the paired CS+ decreased (**Table 5.3** and **Figure 5.2**). Although the resolution of our fMRI procedure precludes any definite conclusions, the activation in the amygdala was located medially and might have been situated in the central nucleus.

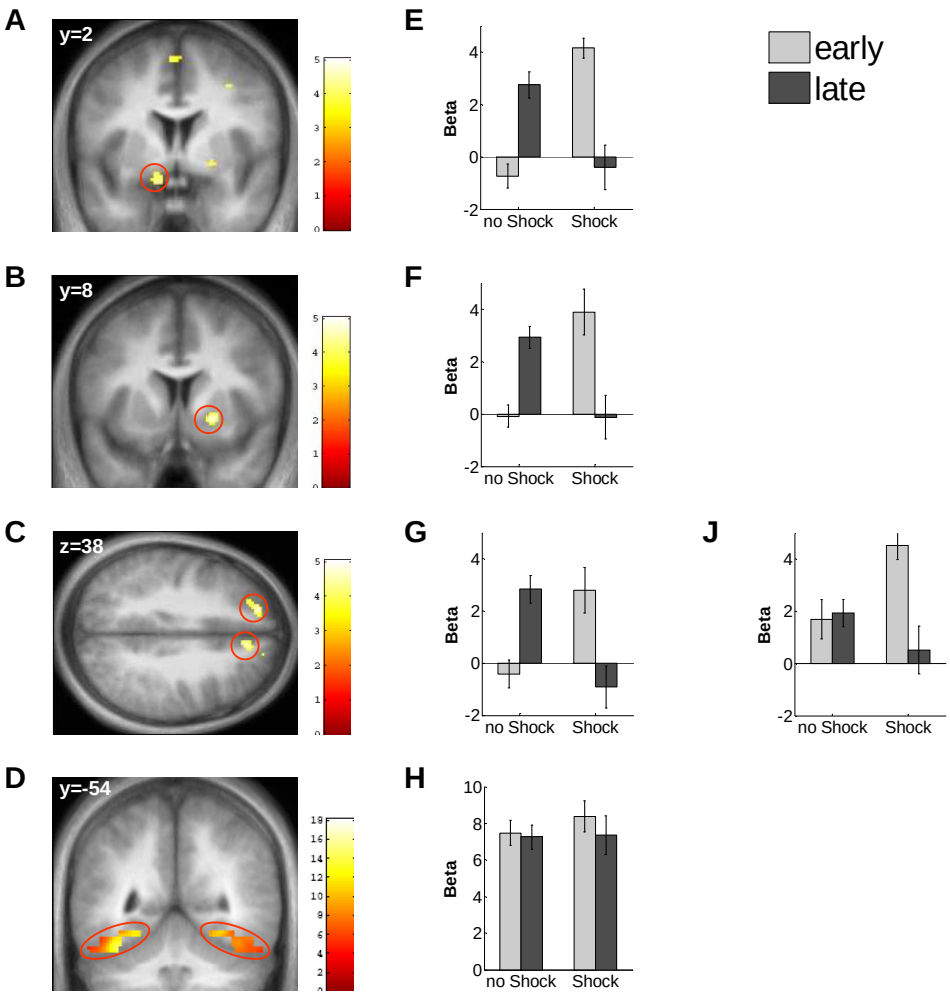

Figure 5.2. Main and time effects for face and shock stimuli. Panels **A-C** show the SPM for the pain x time interaction at the group level, with sections showing the amygdala, striatum and prefrontal cortex; for illustrative purposes, the SPM is thresholded at p <0.001 (uncorrected, df = 26), and displayed on a section of the averaged anatomical scan. Panel **D** shows the main effect of face presentation within the fusiform gyrus anatomical mask, at the same threshold. Panels **E-H** show the associated parameter estimates from the individual local maxima (as determined following the functional and anatomical constraints described in the main text) across the 16 individuals who were included in the DCM analysis. Note that these parameter plots are only displayed for illustrative purposes and are not used for further inference tests. The amygdala, striatum and left DLPFC (**E-G**) all show the same pattern of interaction: over time the response to a paired CS+ decreases, whereas the response to the CS+ alone increases. (**J**) shows the interaction contrast in the right

medial PFC; here the interaction is driven by the decreasing response to the painful stimulus, while the response to the unpaired CS+ stays the same.

Table 5.3. MNI coordinates and Z-values for significantly activated regions.

	MNI coords.				
Foci of activation	X	y	z	Z score	Cluster size
Main effect of 'face'					
L occipital cortex*†	-16	-104	6	7.82	
R occipital cortex*†	-16	-104	6	7.66	
L fusiform gyrus*†	-26	-84	-18	Inf	
L FFA – DCM**	-45	-54	-24	5.60	
R fusiform gyrus*†	34	-66	-12	7.67	
R FFA – DCM**	45	-57	-24	5.43	
Main effect of 'pain'					
L insula*†	40	-14	16	6.74	
R insula*†	-38	-16	14	5.91	
L amygdala*†	-34	2	-22	6.27	
R amygdala*†	30	2	-22	4.58	
L thalamus*†	-8	-4	6	5.42	
R thalamus*†	8	-2	10	4.54	
R S1*†	34	-30	66	5.25	
ACC*†	2	26	24	4.87	
Interaction 'pain x time (+)'					
L amygdala**	-12	2	-16	3.83	32
L dorsolateral PFC **	-18	46	38	4.02	62
R dorsomedial PFC**	10	38	36	4.03	103
R Putamen / Ventral striatum**	20	8	-6	4.16	48
Interaction 'pain x time (-)'					
No activations above threshold					

*significant at $P<0.05$ (FWE whole-brain cluster-level corrected)

** significant at $p<0.05$ (SVC)

† Given the rather unspecific nature of these contrasts, the activations are all part of one large cluster. These activations are all significant at cluster level as well as for height level whole brain correction, and cluster sizes are therefore not reported.

5.3.2 DCM results

Due to the computational demands of nonlinear DCMs, especially with increasing numbers of areas, an initial set of 3-area DCMs was fitted to determine which connection should be modulated to optimally model the pain x time interaction observed in the amygdala (see **Figure 5.3A**). BMS showed that the optimal model was m_2 where the time effect for both shock and non-shock trials affected the connection from the FFA to the amygdala (see **Table 5.4**). In this model, the parameter estimates describing the modulatory influence of time on the connection from the FFA to amygdala was consistently negative for the shock trials (median[8] b = -0.19, P_{corr} = 0.0008), and consistently positive (mean b = 0.16, P_{corr} = 0.0008) for the no-shock trials, consistent with the increasing response to no-shock trials and the decreasing response to shock trials.

In a second step, this optimal 3-area model was extended to include the left PFC and systematically varied to test for the modulatory influence of the amygdala on forward connections from the sensory to striatal and prefrontal areas. This variation was along two dimensions: (i) connections which were gated nonlinearly by amygdala activity (3 options) and (ii) existence vs. absence of backward connections from PFC and striatum to amygdala (2 options). Given this 2x3 factorial model space for the 4-area DCM (see **Figure 5.3B**), model space partitioning could be applied to test separately whether there was convincing evidence for (i) modulation by the amygdala of the FFA→ STR and STR → PFC connections and (ii) endogenous connectivity from the prefrontal cortex and striatum to the amygdala.

With respect to the former, there was clear evidence for modulation of both connections by amygdala activity: the exceedance probability that models including both modulatory influences (m_3 and m_6) had a higher probability of having generated the group data set than models including just a modulatory influence on the FFA→ STR connection (m_1 and m_4) or on the STR → PFC connections (m_2 and m_5), was 80% (see **Table 5.5**).

However, concerning backward connections to the amygdala, the evidence was less clear. Models without these connections (m_{1-3}) fared marginally better than models that did include them (m_{4-6}), but the difference was small; the exceedance probability

[8] Because we used a nonparametric inference method, we report the median rather than the mean.

that the data were more likely to have been generated by a model without the amygdala connections was 0.53, versus 0.47 for models that did include these connections (see **Table 5.5**). Given this lack of differentiability between the two model classes, one would prefer the more parsimonious model class, i.e. without backward connections (cf. Occam's razor).

From this analysis based on model space partitioning it is apparent that model m_3 (both modulatory influences, no backward connections) should be used for inference about the model parameters. This was corroborated by BMS amongst all models treated individually: here, model m_3 had the highest exceedance probability of all six models (xp = 0.42, see **Table 5.5**)[9].

In model m_3 the parameter estimates describing gating effects of amygdala activity on ascending connections were consistently positive across subjects (FFA→striatum: mean d = 0.23, P_{corr} = 0.0018; striatum→PFC: mean d = 0.013, P_{corr} = 0.019). Furthermore, the time dependent modulatory effects of the paired and unpaired CS+ trials were consistently different from zero, reproducing the results of the reduced 3-area model (unpaired CS+: mean b = -0.14, P_{corr} = 0.0018, paired CS+: mean b = 0.10, P_{corr} = 0.0018).

[9] It is possible that this lack of evidence to distinguish between these two sets of models is due to the fact that one of the two connections does increase model fit, but the other one doesn't. Thus to fully test for evidence for the presence of either of these two fixed connections, one needs to extend the model space and add an additional 6 models, which include either the STR→AMY connection or the PFC→AMY connection. We did indeed run these models, and the results stay the same; there is no good evidence in favour of including either of these connections. However, for reasons of brevity these results are not included here.

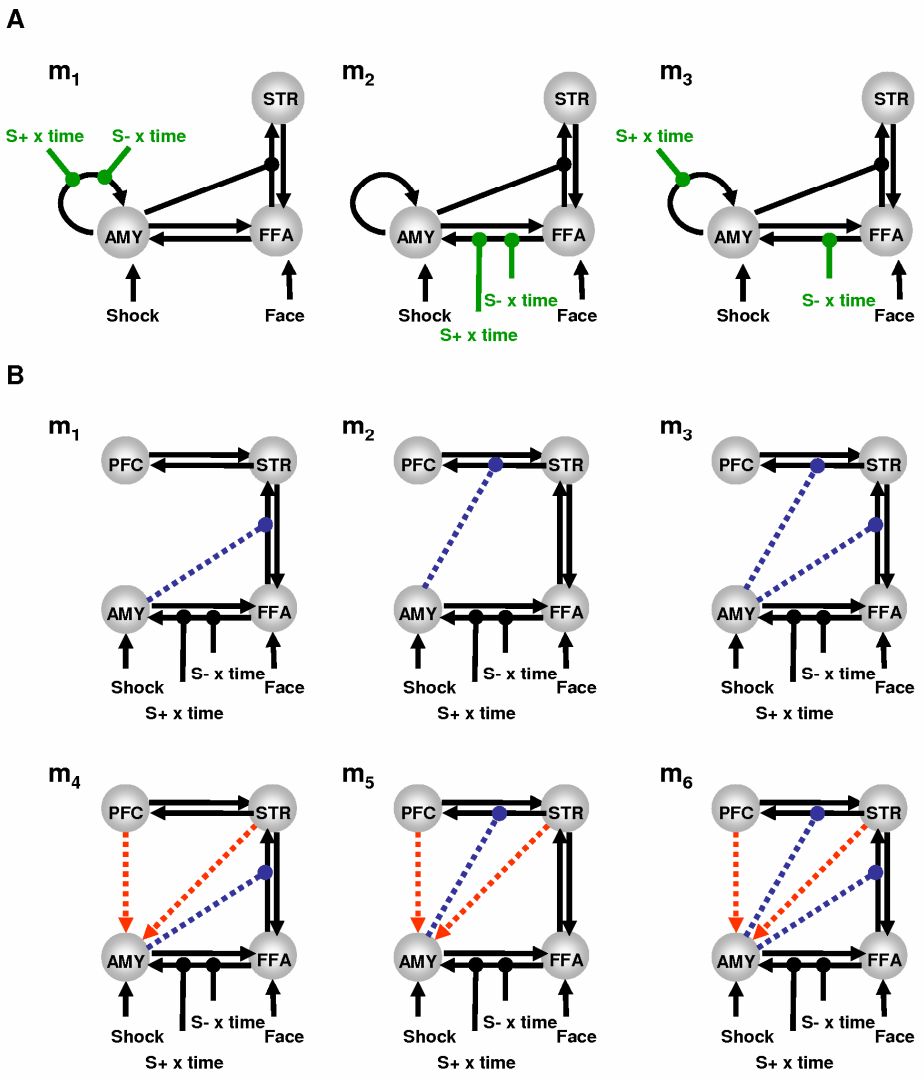

Figure 5.3. 3- and 4-area nonlinear DCMs to model the shock x time interaction. **(A)** Shows the reduced (3-area) model to determine the site of the modulatory learning effect. In m_1 the time x pain interaction modulated the self-connection of the amygdala. In the m_2, shock and non-shock trial x time interaction modulated the FFA→amygdala connection. In m_3, the effect of the shock and non-shock trials was separated, such that the self-connection was affected specifically by the interaction of shock x time.

(B) Shows the full (4-area) DCM testing, in a factorial fashion, firstly for the presence of a gating influence of the amygdala on FFA→striatum connection (m_1 and m_4), the striatum→PFC connection (m_2 and m_5), or both of these (m_3 and m_6), and

secondly for the presence of backward connections from the prefrontal cortex and striatum (absent in (m_{1-3} and present in m_{4-6}).

Table 5.4. BMS results for the 3 area model.

	Dirichlet parameters α	Exceedance probability *xp*
m_1	2.66	0.00
m_2	15.60	1.00
m_3	1.74	0.00

Table 5.5. BMS results for 4-area model. Models varied with regard to the presence or absence of backwards connections to the amygdala, and the modulatory influence of the amygdala on forward connections.

		Amygdala activity modulating:			
		FFA→STR	**STR → PFC**	**FFA→ STR STR→ PFC**	**total**
Backwards connections to the amygdala	**absent**	m_1 α = 3.19 xp = 0.086	m_2 α = 2.53 xp = 0.043	m_3 α = 5.46 xp = 0.415	m_{1-3} α = 11.2 xp = 0.53
	present	m_4 α = 3.27 xp = 0.092	m_5 α = 2.58 xp = 0.046	m_6 α = 4.97 xp = 0.317	m_{4-6} α = 10.8 xp =0.47
	Total	$m_{1,4}$ α = 6.5 xp = 0.14	$m_{2,5}$ α = 5.1 xp = 0.06	$m_{3,6}$ α = 10.4 xp = 0.80	

5.4 Discussion

This study was a reanalysis of a previously published fear conditioning fMRI study, in which four different face stimuli were presented to subjects (Petrovic et al., 2008). On each trial, subjects had to decide whether the presented face was off centre or not. Two of these faces (CS+) were followed by a shock with a 50% contingency, while the other two faces (CS-) were never paired with a shock. Skin conductance responses (SCR) showed that subjects slowly learned the predictive relationships between the faces and the shocks; the SCR for (unpaired) CS+ trials significantly increased from the first to the second half of the experiment compared to the CS- trials. This result suggests that learning happened rather slowly; if learning occurred within a few trials, one would not expect to find noticeable differences between the first and second half of the experiment. The fact that only 25% of the subjects could identify, on post-experimental debriefing, which faces were associated with shocks further corroborates this notion (Petrovic et al., 2008). This slow timescale of learning seems to be at odds with previous fear conditioning studies in humans, where responses to unpaired CS+ rapidly adapt (Buchel et al., 1998;LaBar et al., 1998;Marschner et al., 2008), and is likely to be due to the fact that in the current study, there were two CS+ and two CS- stimuli, that were all very similar (compare **Figure 5.1**). Furthermore, the task was unrelated to the CS stimulus in that subjects had to detect displacement of the stimulus, which would probably direct attention away from the the stimulus identity. This differs from most previous studies, where subjects did attend the stimulus identity, performing a gender discrimination task (Kalisch et al., 2006). This prolonged time course of learning allowed us to look at the differences in BOLD responses during the first and second half of the experiment to assess changing responses to paired and unpaired CS+ trials.

5.4.1 Prediction errors in the amygdala?

The amygdala is the prime anatomical substrate for fear conditioning, especially the lateral nucleus where CS and US inputs converge, inducing synaptic plasticity which changes amygdala responses to CS+ stimuli. Results from the current study support this central role of the amygdala, which was the only brain structure to show both a main effect of painful stimulation and an interaction of pain and time. The nature of

this interaction was such that while the response to unpaired CS+ increased, the response to paired CS+ trials decreased. This pattern of responses is reminiscent of the prediction error responses described in **Chapter 3;** both the surprising presence and absence of a shock elicits an increased response.

Previous studies indeed suggested that the amygdala is sensitive not only to noxious stimuli, but also to how predictable these stimuli are. In general amygdala responses to a noxious stimulus are rapidly attenuated / habituated, due to feedforward inhibition mechanisms in the amygdala itself (LeDoux, 2007). When comparing predictable versus unpredictable stimulation, most studies find that the amygdala responds more to unpredictable stimulation (but see (Carlsson et al., 2006)). In a cross-species study of mice and humans, Herry and colleagues reported that the amygdala responds more strongly to unpredictable noxious stimuli than to predictable ones, and even responds to temporal unpredictability per se, which might be aversive in itself (Herry et al., 2007). Furthermore, using fMRI, Knight and colleagues have shown that amygdala activity increased when experimental contingencies were changed during Pavlovian fear conditioning. This implies that the amygdala might be particularly important for forming new associations among stimuli with behavioural relevance (Knight et al., 2004). In the present study, we built on this previous line of research and investigated where, in a simple network model of associative learning of aversive stimuli, synaptic plasticity was most likely expressed to account for the shock x time interaction responses identified by an initial SPM analysis. We were particularly interested whether there was evidence for modulatory (gating) influences by the amygdala on cortico-striatal connections.

Model comparison of a set of three DCMs showed that the shock x time interaction was best explained by a model in which both paired and unpaired CS+ trials were allowed to modulate the FFA→ amygdala connection. In this model processing of the CS+ input to the amygdala changed over time, depending on whether it was paired with a shock or not. These modelling results from healthy volunteers are nicely consistent with anatomical studies in animals and lesion studies in patients. Amaral et al. showed that in macaques the amygdala is extensively connected to the fusiform gyrus and the primary visual cortex (Amaral et al., 2003). In addition, Vuilleumier et al. reported that patients with amygdala lesions do not show the increased response to fearful faces the occipital cortex and fusiform gyrus that

healthy controls exhibit, and furthermore that the level of amygdala damage predicted the level of visual modulation (Vuilleumier et al., 2004).

5.4.2 Amygdala influences CS+ processing in the cortico-striatal circuit

The SPM results indicated a central role of the amygdala in fear conditioning; although two other areas, the striatum and the PFC, also showed the time x shock interaction, only the amygdala showed an additional main effect of pain. Comparing a set of 4-area nonlinear DCMs we showed that the observed shock x time interaction in the striatum and PFC could be modelled by a gating influence of the amygdala on the connectiona from the FFA to the putamen and from the putamen to the PFC, respectively.

As described in the introduction, the amygdala is well known to mediate the effects of conditioned reinforcers on behaviour. It has been suggested that the underlying mechanism is a modulation of cortico-striatal circuits by amygdala activity (LeDoux, 2007;Mogenson et al., 1980). The striatum, especially its ventral part, is a site of convergence for amygdala and prefrontal projections (Haber and Fudge, 1997). Such connections would allow the amygdala to initiate the motor response to affective stimuli and affect subcortical habit memories putatively stored in striatal circuits. The ventral striatum has long been considered to be an interface between cortical areas involved in processing the emotional valence of stimuli and cortical areas mediating motor responses to those stimuli (e.g. (Haber et al., 1995;Mogenson et al., 1980)).

Finally, given the extensive anatomical connections from the prefrontal cortex and striatum to the amygdala (Haber et al., 1995;Haber and Fudge, 1997), we tested whether there was any evidence that these connections played a functional role in the fear conditioning paradigm used in the current study. Our model comparison approach did not provide such evidence. This might be explained by previous observations that the (ventral) PFC seems to play a role mostly in fear extinction rather than fear acquisition (Sotres-Bayon et al., 2009).

5.4.3 Limitations and future directions

The amygdala, striatum and PFC showed a pattern of responses that was similar to the pattern of prediction error responses to non-affective CS+ stimuli observed in a previous study (**Chapter 3**): unexpected shocks elicited a larger response as than unexpected shocks, as did the unexpected absence of a shock. However, there are possible other interpretations of these findings. For example, it is possible that two separate processes together explain the observed response pattern. The increased response to the unpaired stimulus could reflect the increased (negative) affective value that the CS+ has acquired as it becomes associated with the shock (e.g. (Friston et al., 1994;LeDoux, 2007;Morris and Dolan, 2004)), whereas the decreasing response to the paired CS+ could be due to habituation to the shock itself. These two processes would probably take place in different subnuclei of the amygdala (LeDoux, 2007). However, because of the fast stimulus presentation (**Figure 5.1**) and the long duration of the BOLD response, it is difficult to temporally separate the response to the CS+ and to the (presence or absence of) the shocks, nor can we, because of the limited spatial resolution of standard fMRI methods, distinguish between processes in the different subnuclei of the amygdala.

There are a number of different approaches that could shed light on these questions. One could use electrophysiological recordings which have a much higher temporal resolution. However, scalp-based recording methods such as MEG or EEG do not allow one to measure activity in subcortical structures, including the amygdala, with sufficient signal-to-noise ratio. A better option might be to use an adapted paradigm, in which the CS+ US association probability is changing over time. This would allow one to separate habituations responses, which are likely in the form a of a linear decay function, from a prediction error like response, which would be proportional to the current association strength. Furthermore, one could use very high resolution fMRI optimised for the amygdala to image the processes taking part in the different subnuclei, for example responses to the CS+ and the US.

In conclusion, despite the fact that we remain somewhat agnostic to the exact interpretation of the time dependent responses in the amygdala, striatum and PFC in this fear conditioning paradigm, the results from this study support a role for the amygdala in influencing CS+ processing in cortico-striatal pathways. The functional

role of such a mechanism may reside in providing striatal and cortical regions with information about the emotional valence of the CS+US association. In other words, the modulatory influence exerted by the amygdala on cortico-striatal connections could represent the mechanism by which the amygdala mediates motor responses to affective stimuli, including habit formation through striatal circuits, and emotional colouring of the fear experience in the prefrontal cortex.

Chapter 6

General Conclusions

6.1 Contributions

The aim of this thesis was to establish experimental models which characterise synaptic plasticity in terms of connectivity changes between neural populations in a range of associative learning tasks. Specifically, we wanted to investigate the role of prediction errors in mediating this plasticity. In **Chapter 1** we discussed animal work suggesting that changes in connectivity underlie learning (Genoux and Montgomery, 2007;Gu, 2002;Ji et al., 2005;Morris, 1989;Tye et al., 2008). Furthermore, prediction errors, or surprising events, are thought to signal the need for updating beliefs; they thus play a central role for associative learning in animals and humans (cf. **Section 1.1.1**). Indeed, surprise appears to be at the heart of not only to reward-based learning, but any form of (associative) learning (**Section 1.1.2**). Taken together, this suggests that surprising outcomes could drive the modulation of connection strengths, i.e. synaptic plasticity, during associative learning. Although a large number of animal electrophysiology and human fMRI studies have shown surprising outcomes to elicit responses in the striatum and in sensory areas, to our knowledge, the notion that surprise dependent changes of connectivity mediate learning has not been investigated empirically means before.

In this thesis, I employed standard and Bayesian associative learning models (see **Section 1.2)** to estimate the surprise engendered by observed events, and combined these with plausible physiological models of connectivity (**Chapter 2**) to investigate surprise dependent modulation of connections during associative learning.

In **Chapter 3**, I used a carefully balanced design with auditory cues predicting visual outcomes to investigate whether previously described responses in the visual cortex were driven by predictions or prediction errors. Critically, learning was shown to occur at a neurophysiological level, even though the audiovisual associations were

irrelevant to the behavioural task and outside the subjects' awareness. We observed prediction error dependent responses in the (peripheral) primary visual cortex and the ventral striatum, as predicted by an RW model of associative learning. This response could be explained in terms of changes in connectivity from auditory to visual cortex, where the connections were modulated by the prediction errors.

In **Chapter 4** we extended the paradigm from the previous study such that the learned associations were now task-relevant. In this study we employed a hierarchical Bayesian ideal observer model that could capture changing audiovisual associations. Again both sensory areas and the striatum showed prediction error dependent responses; in the sensory areas, this response was specific to the presented visual stimulus, whereas the striatal responses reflected prediction errors per se. In parallel to these striatal responses we observed prediction error responses in the motor planning areas. Using a nonlinear DCM, we showed, for the first time, that these prediction error responses in motor planning areas could be explained by a modulation of sensory-motor connections by the prediction error dependent output of the striatum.

In **Chapter 5** we reanalysed a pre-existing fMRI dataset to investigate prediction error like responses in the amygdala during fear conditioning. Here we showed that prediction errors modulate amygdala processing of sensory input, and furthermore that amygdala activity modulates cortico-striatal connections as neutral cues become associated with noxious outcomes.

Model selection to decide between different DCMs relied on Bayesian model comparison methods as described in **Chapters 3-5**. In **Chapter 3** the selection of the best model was based on the group Bayes factor, which was calculated by multiplying the individual Bayes factors for each subject. However, this fixed effects approach does not take into account random variations in optimal model structure across subjects. Therefore, in **Chapters 4-5** we used a newly developed second level Bayesian random effects analysis which accounts for such random effects.

6.2 Limitations

In addition to the limitations of the specific designs and paradigms discussed in the results chapters (**Chapter 3-5**), what follows are some general considerations on the

use and usefulness of dynamic causal models. DCM for fMRI is a state-space model that explains observed BOLD responses in specific regions of the brain in terms of changes in effective connectivity between these areas. Crucially, this connectivity can be modulated by external inputs, or, in case of the nonlinear DCM extension, by outputs from other areas.

6.2.1 Effective connections are not anatomical connections

One common misconception about DCMs is that the presence of an effective connection between two areas equates to a single synapse at the anatomical level. Instead, an effective connection may also be a summary of multisynaptic connectivity between two areas. In other words, we remain agnostic with respect to the precise anatomical nature of the connection. What effective connectivity does reflect is a causal influence of one area on another. For example, the visual to premotor connections described in **Chapter 4** are unlikely to be monosynaptic connections; yet we showed a directed causal influence from the sensory to the premotor areas.

6.2.2 Interpreting causality

The course temporal resolution of the fMRI as well as the smoothness of the BOLD response itself do not allow for interpretations about temporal causality in cortical network models of fMRI data. This often leads to the question how one can then make a claim about causal influences between areas in DCM for fMRI. This is explained by the fact that the shape of the modelled BOLD responses differ when areas receive direct external inputs versus input from another area. Direct inputs with elicit a sharp peak and then rapid decline, whereas in downstream areas the response rises and falls more slowly. Thus, causality in DCM for fMRI is determined by the shape of the modelled BOLD response.

6.2.3 Exploring and defining model space

DCM is a method for hypothesis testing and has very limited use as an explorative tool. Completing a full search of all possible models, or hypotheses, within a given set of nodes and inputs is simply too computationally demanding as soon as one deals with models with more than two areas. Consider for example a nonlinear DCM

with 4 areas, as employed in **Chapter 4-5**. Currently, each model of that sort takes roughly 30 minutes to run on a standard PC. Let us simplify the model space and assume that we know where the driving and modulatory inputs enter, so that all we have to do is explore the endogenous connection part of model space. There are 12 potential endogenous connections, so systematically testing for all possible combinations of endogenous connections would result in $2^{12} = 4096$ models, i.e. over 2000 hours or 85 days of processing time per subject, still disregarding the modulatory inputs and connections. Thus, one has to make principled decisions as to which models constitute a sensible set of models to test, constraining model space using biological and theoretical constraints.

Even if one could exhaustively search model space, and use a partitioning approach (**Chapter 2**) to investigate the contribution of different connections and modulatory inputs, questions might always arise as to whether the right areas have been included in the model. Because one cannot compare models that relate to different datasets, it is not possible, in the context of DCM for fMRI, to compare models that include different nodes. However, it is important to keep in mind here again that DCM is a hypothesis-driven method, set up to test very specific mechanistic hypotheses about interactions between different areas in the brain. Thus, rather than trying to build a model of the entire brain, only areas that are thought to be involved in the process regarding the underlying hypothesis, should be included.

Finally, it is important to keep in mind (cf. **Chapter 1**, **Section 1.3**) that there is no single 'right' model of the world that can describe the world in all its facets; there are only better or worse approximations to particular aspects of reality.

6.3 Future Research

6.3.1 MEG to resolve temporal resolution

Due to the course temporal resolution of fMRI as well as the smoothness of the BOLD response in combination with the rapid stimulus designs used in the work described in this thesis, it was not possible to investigate the within trial temporal evolution of prediction responses evoked by the cues, and prediction error responses evoked by the outcomes. Magnetoencephalography (MEG) would provide an

excellent tool to evaluate within-trial predictions and learning. These data could then be combined with TD learning models or extensions of Bayesian models that model within trial timing effects, in combination with DCM for M/EEG (Kiebel et al., 2008).

6.3.2 Pharmacology

As was described in the introduction, brain connectivity speaks to three key issues: synaptic strength, changes in synaptic strength (plasticity), and modulation of this plasticity. Synaptic plasticity is likely to underlie the changes in effective connectivity during associative learning as demonstrated in the work presented in this thesis. Having established experimental models of connectivity in two very simple, non-reward based associative learning paradigms in **Chapters 3-4**, these could now be repeated using pharmacological manipulations, to investigate the role of different neurotransmitters. The most obvious candidate to start with would be ACh receptor agonists and antagonists. ACh is one of the most important modulators of synaptic plasticity in the context of associative learning in the perceptual domain; in humans in has been shown to affect perceptual learning effects such as the MMN (Baldeweg et al., 2006) and repetition priming (Thiel et al., 2002c), as well as associative fear learning using auditory cues (Thiel et al., 2002b;Thiel et al., 2002a). DCM would be an ideal tool to investigate the modulatory effects of these neurotransmitters on effective connectivity in humans.

6.3.3 Associative learning, connectivity & schizophrenia

Connectivity is the basis of physiological neural information processing and may be central to the pathophysiology of various neurological and psychiatric diseases, most notably schizophrenia (for detailed reviews see (Friston, 2005b;Stephan et al., 2006)). The mechanistic models of connectivity underlying the associative learning paradigms discussed in **Chapter 3-4** were deliberately designed to be suitable for assessing changes in connectivity in patients with schizophrenia. The behavioural tasks are extremely simple, such that patients could easily perform them, and yet they evoke consistent changes in connection strengths. In the future, using simple physiological models of this sort in combination with formal theoretical learning

models may help to obtain a mechanistic understanding of abnormalities of synaptic plasticity in schizophrenia.

Publications and Other Work During the PhD

Chapter 3 has been published in Cerebral Cortex (den Ouden et al., 2009). A paper based on Chapter 4 has been submitted to Neuron, and Chapter 5 is in preparation for submission to Journal of Neuroscience. Furthermore, I contributed to the development of the nonlinear DCM, discussed in Chapter 2, which has been published in Neuroimage (Stephan et al., 2008). In relation to the Bayesian learning model presented in Chapter 4, I collaborated on theoretical work considering the brain as a Bayesian observer of the environment with Jean Daunizeau, which is currently submitted to PlosOne. During my PhD I also contributed to patient and fMRI studies investigating aberrant salience in schizophrenia patients in collaboration with Jonathan Roiser. One of these studies has been published in Psychological Medicine (Roiser et al., 2009), and a second one is currently submitted to Neuroimage. Finally, during my PhD I published two papers based on work prior to my PhD (den Ouden et al., 2005a; den Ouden et al., 2005b), and co-authored one further paper (Blakemore et al., 2007).

Published

den Ouden,H.E., Friston,K.J., Daw,N.D., McIntosh,A.R., and Stephan,K.E. (2009). A Dual Role for Prediction Error in Associative Learning. Cereb. Cortex 1175-1185.

Roiser,J.P., Stephan,K.E., **den Ouden,H.E.**, Barnes,T.R., Friston,K.J., and Joyce,E.M. (2009). Do patients with schizophrenia exhibit aberrant salience? Psychol. Med. *39*, 199-209.

Stephan,K.E., Kasper,L., Harrison,L.M., Daunizeau,J., **den Ouden,H.E.**, Breakspear,M., and Friston,K.J. (2008). Nonlinear dynamic causal models for fMRI. Neuroimage. *42*, 649-662.

Blakemore,S.J., **den,Ouden,.H.E.**, Choudhury,S., and Frith,C. (2007). Adolescent development of the neural circuitry for thinking about intentions. Soc. Cogn Affect. Neurosci. *2*, 130-139.

den Ouden,H.E., Frith,U., Frith,C., and Blakemore,S.J. (2005a). Thinking about intentions. Neuroimage. *28*, 787-796.

den Ouden,H.E., van EE, R., and de Haan,E.H. (2005b). Colour helps to solve the binocular matching problem. J. Physiol *567*, 665-671.

In submission

den Ouden, H. E., Daunizeau J, Roiser, JP, Friston, K. J., Stephan, K. E. 2009, Striatal prediction error drives cortical connectivity changes during associative learning.

den Ouden, H. E., Petrovic, P., Dolan RJ, Friston, KJ, Stephan, K.E. The amygdala modulates cortico-striatal connections during fear acquisition.

Daunizeau, J., **den Ouden H.E.**, Pessiglione M., Stephan K.E., Kiebel S.J. & Friston K.J. Observing the Observer: Meta-Bayesian models of learning and decision-making.

Roiser, J. P., Stephan, K. E., **den Ouden, H. E.**, Friston, K. J., & Joyce, E. M. Adaptive and aberrant reward signals in the human brain.

Appendix A

Prediction vs. Prediction Error in the Rescorla-Wagner (RW) Model

Here we show that predictions and prediction errors computed by the RW model are linearly related under mean-correction. In fact, one is identical to the negative of the other. This linear dependence between predictions and prediction errors is problematic for GLM analyses since it precludes separate testing for the contributions of prediction errors and predictions to the dependent variable. Note that whenever there is any experimental factor other than the learning process itself, it is necessary to model the interaction among these factors and learning, and this requires mean-correction of the vectors involved before computing their Hadamard product (cf. (Friston et al., 1997). In SPM, these interaction terms are known as "parametric modulation".

At trial *t*, the prediction error PE_t is the difference between the predicted outcome ϕ_t and the actual outcome λ_t:

$$PE_t = \lambda_t - \phi_t \tag{A.1}$$

The prediction (error) at trial *t* is the sum of the mean-corrected prediction (error) and the mean:

$$PE_t = PE_{corr,t} + \overline{PE} \tag{A.2}$$

$$\phi_t = \phi_{corr,t} + \bar{\phi} \tag{A.3}$$

For typical reinforcement schemes, the outcome λ_t takes on the values 1 (unconditioned stimulus is present) or 0 (unconditioned stimulus is absent). For both trial types, the mean-corrected prediction error is exactly the negative of the mean-corrected prediction, as we will show below:

For $\lambda t = 1$

$$PE_t = 1 - \phi_t \tag{A.4}$$

$$\overline{PE} = 1 - \bar{\phi} \tag{A.5}$$

$$PE_t + \phi_t = 1 \tag{A.6}$$

Substituting **Equations A.2,3,5** into **Equation A.6** gives:

$$PE_{corr,t} = -\phi_{corr,t} \tag{A.7}$$

Similarly, for $\lambda_t = 0$,

$$\overline{PE} = -\bar{\phi} \tag{A.8}$$

$$PE_t = -\phi_t \tag{A.9}$$

Substituting **Equations A.2,3,8** into **Equation A.9** gives:

$$PE_{corr,t} = -\phi_{corr,t} \tag{A.10}$$

This shows that independent of the outcome λ_t, the meancorrected prediction error is always the negative of the meancorrected prediction.

Appendix B

Bayesian Volatility-Based Associative Learning Model

We start with the premise that subjects represent or infer the causes of their sensory inputs and optimise their behaviour on the basis of this inference. From a Bayesian perspective, the brain is an *observer* of its own sensory signals. This means subjects invert some forward or generative model of sensory inputs to represent the unobserved (hidden) causes of that input.

Any learning then relies strongly on the subject's model of the world (the perceptual model), which can have dramatic effects on both predicted behaviour (Kording et al., 2007;Trepel et al., 2005) and modelled neurophysiological signals (Pessiglione et al., 2007;Tom et al., 2007).

In what follows, we describe the volatility-based perceptual model used in this study to estimate the volatility and probabilities of the observed events. This model subsumes the set of probabilistic assumptions the brain encoded in order to represent the causes of paired audio-visual stimuli.

The perceptual model generates sensory input u (*e.g.*, experimental stimuli) from hidden causes, x (*e.g.*, experimental factors or environmental states) and can be expressed in terms of a likelihood model $p(u|x)$ and prior beliefs $p(x)$. The states of the world x are unknown to the subject but might be under experimental control. In our example, u is a series of cue-outcome pairs, presented to the observer and x encodes an experimentally controlled cue-outcome association that is hidden from the subject. The prior belief itself is decomposed into a hierarchy of conditional probability density functions, as will be described bellow.

Let u_t be the outcome at trial *t* be a multinomial random variate such that:

$$p\left(u_t \left| u_t^c, r_t \right.\right) = Mult\left(u_t \left| r_t \right.\right) = \prod_{i=1}^{n} \left(r_t^i\right)^{u_t^i}, \quad \text{(B.1)}$$

where $\left(r_t^i\right)^{i=1,\ldots,n}$ is a *n*x1 vector of probabilities describing completely the distribution of the *n* possible outcomes.

This forms the likelihood of our generative model. Note that from there on, we will consider that each of the cues u_t^c is associated with its own likelihood, and consequently, its own generative model. This means that everything we state below is conditional on the given cue. As a consequence, the Bayesian inversion of such a set of generative models is also conditional on each cue, and has to be replicated for all different cues.

This vector of cue-outcome association probabilities follows *a priori* the following Dirichlet distribution:

$$p\left(r_t \left| r_{t-1}, v_t \right.\right) = Dir\left(r_t \left| a_t \right.\right) = \frac{\Gamma\left(a_t^0\right)}{\prod_{i=1}^{n} \Gamma\left(a_t^i\right)} \prod_{i=1}^{n} \left(r_t^i\right)^{a_t^i - 1} \quad \text{(B.2)}$$

This transition density is actually a martingale, i.e. it is a first order Markov process whose current first order moment is equal to its previous realization:

$$\left\langle r_t \right\rangle = r_{t-1}. \quad \text{(B.3)}$$

Furthermore, the precision of the transition from r_{t-1} to r_t is parameterized by a scalar quantity v_t, which measures the volatility of the environment:

$$\sum_{i=1}^{n} a_t^i = \exp\left(-v_t\right) + 1 \quad \text{(B.4)}$$

The volatility itself is assumed to vary over time as a martingale, and the above parameterization makes a simple AR(1) model possible:

$$p\left(v_t \middle| v_{t-1}, K\right) = N\left(v_t \middle| v_{t-1}, K\right) = \frac{1}{\sqrt{2\pi K}} \exp\left(-\frac{1}{2K}\left(v_t - v_{t-1}\right)^2\right), \quad \text{(B.5)}$$

where K is the prior variance of the volatility, i.e. the volatility's volatility.

The prior on K itself is supposed to be non-informative, i.e.:

$$p(K) \propto 1.$$

To summarize, the generative model assumes the following cascade of events (illustrated in the graph in **Figure B.1**):

1- A value for the volatility variance K is randomly drawn from its prior pdf $p(K)$.

2- This value determines the transition pdf of the volatility. Then, a first value v_1 is randomly drawn from $p\left(v_t \middle| v_{t-1}, K\right)$.

3- Knowing the volatility v_1 then allow us to derive the transition density for r_1. Then, a first value for the cue-outcome association probability is drawn from $p\left(r_t \middle| r_{t-1}, v_t\right)$.

4- This finally defines the likelihood of the outcome itself: the first outcome u_1 is then drawn randomly from $p\left(u_t \middle| u_t^c, r_t\right)$.

5- The steps 3, 4 and 5 are repeated in time, giving rise to three time series for the volatility v_t, the cue-outcome association probability r_t and the observed outcomes u_t.

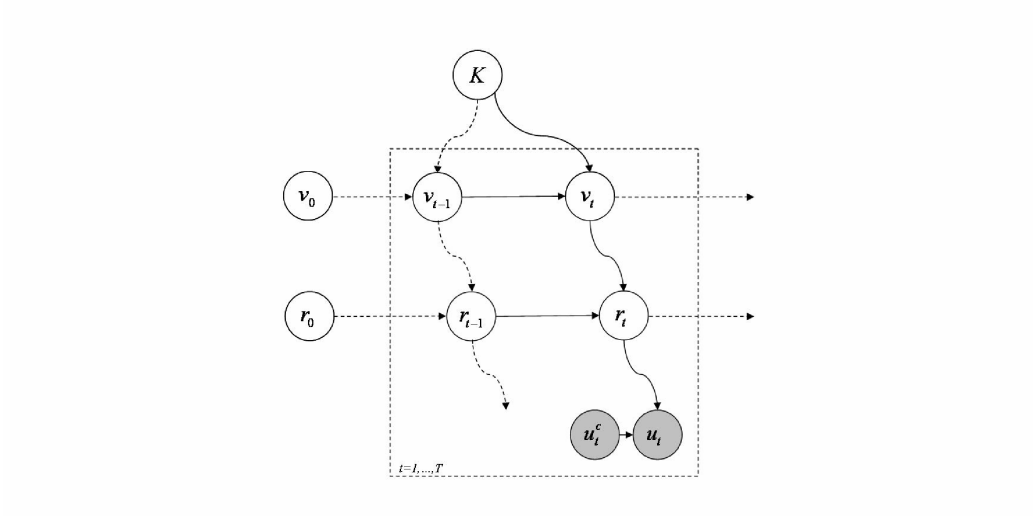

Figure B.1. Graph illustration of the volatility model. u_t= observed outcome at trial t; r_t = cue-outcome association probability; v_t= volatility; K = variance of the volatility.

Then, the model assumes that the observer actually updates its posterior belief 'on the fly', in the light of incoming data, in a Kalman filter-like manner. The joint posterior pdf over the full set of unknown variables, namely $x = \{K, v, r\}$, then follows the following prediction and update steps:

$$\text{prediction: } p\left(r_t, v_t, K \middle| u_{1:t-1}\right) = \iint p\left(r_t \middle| r_{t-1}, v_{t-1}\right) p\left(v_t \middle| v_{t-1}, K\right) p\left(r_{t-1}, v_{t-1}, K \middle| u_{1:t-1}\right) dr_{t-1} dv_{t-1}$$

$$\text{update: } p\left(r_t, v_t, K \middle| u_{1:t}\right) = \frac{p\left(r_t, v_t, K \middle| u_{1:t-1}\right) p\left(u_t \middle| r_t\right)}{\iiint p\left(r_t, v_t, K \middle| u_{1:t-1}\right) p\left(u_t \middle| r_t\right) dr_t dv_t dK}$$

These two steps are iterated as long as new data are measured and, after each cue-outcome observation, yield estimates of both the current cue-outcome association probability r_t and the environmental volatility v_t, as well as an estimator of the static volatility's variance K, given all previously observed data. The trajectory of these estimates as a function of time (trial t) then served as predictors for behavioural data (response speeds) and neuroimaging data (BOLD responses in SPM and DCM analyses).

Abbreviations

Brain Areas and Neural Properties

5HT	Serotonin
A1	Primary auditory cortex
ACC	Anterior cingulate cortex
ACh	Acethylcholine
AMPA	a-amino-3-hydroxyl-5-methyl-4- isoxazole-propionate
AMY	Amygdala
DA	Dopamine
DLPFC	Dorsolateral prefrontal cortex
FFA	Fusiform face area
LTP/D	Long term potentiation/depression
NE	Norepinephrine
NMDA	N-methyl-D-aspartate
PFC	Prefrontal cortex
PPA	Parahippocampal place area
S1	Primary somatosensory area
S2	Secondary somatosensory area
STR	Striatum
V1	Primary visual cortex
VTA	Ventral tegmental area

Methodological Terminology

AIC	Akaike information criterion
AR(1)	First order autoregressive moving-average model
BF	Bayes factor
BIC	Bayesian information criterion
BMS	Bayesian model selection
BOLD	Blood oxygen level dependent
DCM	Dynamic causal modelling
df	Degrees of freedom
EEG	Electroencephalography
EM	Expectation-Maximization
EPI	Echo-planar imaging
F	Free energy
fMRI	Functional magnetic resonance imaging
FWE	Family wise error
GLM	General linear model
ME	Main effect
MEG	Magnetoencephalography
MNI	Montreal neurological institute
MRI	Magnetic resonance imaging
PDF	Probability density function
PET	Positron emission tomography
ROI	Region of interest
RT	Reaction time

RS	Response speed
SD	Standard Deviation
SEM	Structural equation modelling
SPM	Statistical parametric mapping
SVC	Small volume correction
SCR	Skin conductance response

Theoretical Terminology

CR	Conditioned response
CS	Conditioned stimulus
MMN	Mismatch negativity
nlDCM	NonLinear dynamic causal model
RW	Rescorla Wagner
UR	Unconditioned response
US	Unconditioned stimulus
TD learning	Temporal difference learning
TO cue	Trial onset cue
PE	Prediction error

Bibliography

Abler,B., Walter,H., Erk,S., Kammerer,H., and Spitzer,M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. Neuroimage. *31*, 790-795.

Akaike,H. (1974). A new look at the statistical model identification. IEEE Trans. Automatic Control *19*, 716-723.

Akatsuka,K., Wasaka,T., Nakata,H., Kida,T., and Kakigi,R. (2007). The effect of stimulus probability on the somatosensory mismatch field. Exp. Brain Res. *181*, 607-614.

Alexander,G.E., and Crutcher,M.D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci. *13*, 266-271.

Amaral,D.G., Behniea,H., and Kelly,J.L. (2003). Topographic organization of projections from the amygdala to the visual cortex in the macaque monkey. Neuroscience *118*, 1099-1120.

Andersson,J.L., Hutton,C., Ashburner,J., Turner,R., and Friston,K. (2001). Modeling geometric deformations in EPI time series. Neuroimage. *13*, 903-919.

Aron,A.R., Shohamy,D., Clark,J., Myers,C., Gluck,M.A., and Poldrack,R.A. (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. J. Neurophysiol. *92*, 1144-1152.

Ashburner,J., and Friston,K.J. (2005). Unified segmentation. Neuroimage. *26*, 839-851.

Baddeley,A., and Della,S.S. (1996). Working memory and executive control. Philos. Trans. R. Soc. Lond B Biol. Sci. *351*, 1397-1403.

Baier,B., Kleinschmidt,A., and Muller,N.G. (2006). Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. J. Neurosci. *26*, 12260-12265.

Baldeweg,T. (2006). Repetition effects to sounds: evidence for predictive coding in the auditory system. Trends Cogn Sci. *10*, 93-94.

Baldeweg,T., Wong,D., and Stephan,K.E. (2006). Nicotinic modulation of human auditory sensory memory: Evidence from mismatch negativity potentials. Int. J. Psychophysiol. *59*, 49-58.

Bayer,H.M., and Glimcher,P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129-141.

Bays,P.M., Flanagan,J.R., and Wolpert,D.M. (2006). Attenuation of self-generated tactile sensations is predictive, not postdictive. PLoS. Biol. *4*, e28.

Bays,P.M., and Wolpert,D.M. (2007). Computational principles of sensorimotor control that minimize uncertainty and variability. J. Physiol *578*, 387-396.

Bechara,A., Tranel,D., Damasio,H., Adolphs,R., Rockland,C., and Damasio,A.R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. Science *269*, 1115-1118.

Beck,J.M., Ma,W.J., Kiani,R., Hanks,T., Churchland,A.K., Roitman,J., Shadlen,M.N., Latham,P.E., and Pouget,A. (2008). Probabilistic population codes for Bayesian decision making. Neuron *60*, 1142-1152.

Behrens,T.E., Woolrich,M.W., Walton,M.E., and Rushworth,M.F. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214-1221.

Bestmann,S., Harrison,L.M., Blankenburg,F., Mars,R.B., Haggard,P., Friston,K.J., and Rothwell,J.C. (2008). Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. Curr. Biol. *18*, 775-780.

Blair,H.T., Sotres-Bayon,F., Moita,M.A., and LeDoux,J.E. (2005). The lateral amygdala processes the value of conditioned and unconditioned aversive stimuli. Neuroscience *133*, 561-569.

Blakemore,S.J., Wolpert,D.M., and Frith,C.D. (1998). Central cancellation of self-produced tickle sensation. Nat. Neurosci. *1*, 635-640.

Bray,S., and O'Doherty,J. (2007). Neural coding of reward-prediction error signals during classical conditioning with attractive faces. J. Neurophysiol. *97*, 3036-3045.

Brocher,S., Artola,A., and Singer,W. (1992). Agonists of cholinergic and noradrenergic receptors facilitate synergistically the induction of long-term potentiation in slices of rat visual cortex. Brain Res. *573*, 27-36.

Brodersen,K.H., Penny,W.D., Harrison,L.M., Daunizeau,J., Ruff,C.C., Duzel,E., Friston,K.J., and Stephan,K.E. (2008). Integrated Bayesian models of learning and decision making for saccadic eye movements. Neural Netw. *21*, 1247-1260.

Brogden,W.J. (1939). Sensory Preconditioning. Journal of Experimental Psychology *25*, 323-332.

Büchel,C., Coull,J.T., and Friston,K.J. (1999). The predictive value of changes in effective connectivity for human learning. Science *283*, 1538-1541.

Buchel,C., Morris,J., Dolan,R.J., and Friston,K.J. (1998). Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron *20*, 947-957.

Carlsson,K., Andersson,J., Petrovic,P., Petersson,K.M., Ohman,A., and Ingvar,M. (2006). Predictability modulates the affective and sensory-discriminative neural processing of pain. Neuroimage. *32*, 1804-1814.

Chen,C.C., Kiebel,S.J., and Friston,K.J. (2008). Dynamic causal modelling of induced responses. Neuroimage. *41*, 1293-1312.

Corbetta,M., and Shulman,G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. *3*, 201-215.

Corlett,P.R., Aitken,M.R., Dickinson,A., Shanks,D.R., Honey,G.D., Honey,R.A., Robbins,T.W., Bullmore,E.T., and Fletcher,P.C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. Neuron *44*, 877-888.

Courville,A.C., Daw,N.D., and Touretzky,D.S. (2006). Bayesian theories of conditioning in a changing world. Trends Cogn Sci. *10*, 294-300.

Crammond,D.J., and Kalaska,J.F. (2000). Prior information in motor and premotor cortex: activity during the delay period and effect on pre-movement activity. J. Neurophysiol. *84*, 986-1005.

Csepe,V., Karmos,G., and Molnar,M. (1987). Evoked potential correlates of stimulus deviance during wakefulness and sleep in cat--animal model of mismatch negativity. Electroencephalogr. Clin. Neurophysiol. *66*, 571-578.

D'Ardenne,K., McClure,S.M., Nystrom,L.E., and Cohen,J.D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science *319*, 1264-1267.

David,O., Kiebel,S.J., Harrison,L.M., Mattout,J., Kilner,J.M., and Friston,K.J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. Neuroimage *30*, 1255-1272.

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

Deichmann,R., Josephs,O., Hutton,C., Corfield,D.R., and Turner,R. (2002). Compensation of susceptibility-induced BOLD sensitivity losses in echo-planar fMRI imaging. Neuroimage. *15*, 120-135.

Delgado,M.R., Olsson,A., and Phelps,E.A. (2006). Extending animal models of fear conditioning to humans. Biol. Psychol. *73*, 39-48.

den Ouden,H.E., Friston,K.J., Daw,N.D., McIntosh,A.R., and Stephan,K.E. (2009). A Dual Role for Prediction Error in Associative Learning. Cereb. Cortex 1175-1185.

Eickhoff,S.B., Stephan,K.E., Mohlberg,H., Grefkes,C., Fink,G.R., Amunts,K., and Zilles,K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage *25*, 1325-1335.

Ekman,P., and Friesen,W. (1976). Pictures of facial affect (Palo Alto, CA: Consulting Psychologists Press).

Falchier,A., Clavagnier,S., Barone,P., and Kennedy,H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. J. Neurosci. *22*, 5749-5759.

Fiorillo,C.D., Tobler,P.N., and Schultz,W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898-1902.

Fletcher,P.C., Anderson,J.M., Shanks,D.R., Honey,R., Carpenter,T.A., Donovan,T., Papadakis,N., and Bullmore,E.T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nat. Neurosci. *4*, 1043-1048.

Friston,K. (2003). Learning and inference in the brain. Neural Netw. *16*, 1325-1352.

Friston,K. (2005a). A theory of cortical responses. Philos. Trans. R. Soc. Lond B Biol. Sci. *360*, 815-836.

Friston,K. (2005b). Disconnection and cognitive dysmetria in schizophrenia. Am. J. Psychiatry *162*, 429-432.

Friston,K., Kilner,J., and Harrison,L. (2006). A free energy principle for the brain. J. Physiol Paris *100*, 70-87.

Friston,K., Mattout,J., Trujillo-Barreto,N., Ashburner,J., and Penny,W. (2007). Variational free energy and the Laplace approximation. Neuroimage. *34*, 220-234.

Friston,K.J., Büchel,C., Fink,G.R., Morris,J., Rolls,E., and Dolan,R.J. (1997). Psychophysiological and modulatory interactions in neuroimaging. Neuroimage. *6*, 218-229.

Friston,K.J., Frith,C.D., and Frackowiak,R.S. (1993a). Principal component analysis learning algorithms: a neurobiological analysis. Proc. Biol. Sci. *254*, 47-54.

Friston,K.J., Frith,C.D., Liddle,P.F., and Frackowiak,R.S. (1993b). Functional connectivity: the principal-component analysis of large (PET) data sets. J. Cereb. Blood Flow Metab *13*, 5-14.

Friston,K.J., Glaser,D.E., Henson,R.N., Kiebel,S., Phillips,C., and Ashburner,J. (2002a). Classical and Bayesian inference in neuroimaging: applications. Neuroimage. *16*, 484-512.

Friston,K.J., Harrison,L., and Penny,W. (2003). Dynamic causal modelling. Neuroimage. *19*, 1273-1302.

Friston,K.J., Penny,W., Phillips,C., Kiebel,S., Hinton,G., and Ashburner,J. (2002b). Classical and Bayesian inference in neuroimaging: theory. Neuroimage. *16*, 465-483.

Friston,K.J., and Stephan,K.E. (2007). Free-energy and the brain. Synthese *159*, 417-458.

Friston,K.J., Stephan,K.E., Lund,T.E., Morcom,A., and Kiebel,S. (2005). Mixed-effects and fMRI studies. Neuroimage. *24*, 244-252.

Friston,K.J., Tononi,G., Reeke,G.N., Jr., Sporns,O., and Edelman,G.M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience *59*, 229-243.

Garrido,M.I., Kilner,J.M., Kiebel,S.J., and Friston,K.J. (2009a). Dynamic causal modelling of the response to frequency deviants. J. Neurophysiol.

Garrido,M.I., Kilner,J.M., Kiebel,S.J., Stephan,K.E., and Friston,K.J. (2007). Dynamic causal modelling of evoked potentials: a reproducibility study. Neuroimage *36*, 571-580.

Garrido,M.I., Kilner,J.M., Stephan,K.E., and Friston,K.J. (2009b). The mismatch negativity: A review of underlying mechanisms. Clin. Neurophysiol.

Genoux,D., and Montgomery,J.M. (2007). Glutamate receptor plasticity at excitatory synapses in the brain. Clin. Exp. Pharmacol. Physiol *34*, 1058-1063.

Gewirtz,J.C., and Davis,M. (2000). Using pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. Learn. Mem. *7*, 257-266.

Gläscher,J., and Büchel,C. (2005). Formal learning theory dissociates brain regions with different temporal integration. Neuron *47*, 295-306.

Grol,M.J., Majdandzic,J., Stephan,K.E., Verhagen,L., Dijkerman,H.C., Bekkering,H., Verstraten,F.A., and Toni,I. (2007). Parieto-frontal connectivity during visually guided grasping. J. Neurosci. *27*, 11877-11887.

Gu,Q. (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. Neuroscience *111*, 815-835.

Haber,S.N., and Fudge,J.L. (1997). The interface between dopamine neurons and the amygdala: implications for schizophrenia. Schizophr. Bull. *23*, 471-482.

Haber,S.N., Kunishio,K., Mizobuchi,M., and Lynd-Balta,E. (1995). The orbital and medial prefrontal circuit through the primate basal ganglia. J. Neurosci. *15*, 4851-4867.

Hare,T.A., O'Doherty,J., Camerer,C.F., Schultz,W., and Rangel,A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J. Neurosci. *28*, 5623-5630.

Hebb,D.O. (1949). The organisation of behaviour (New York: John Wiley).

Herry,C., Bach,D.R., Esposito,F., Di Salle,F., Perrig,W.J., Scheffler,K., Luthi,A., and Seifritz,E. (2007). Processing of temporal unpredictability in human and animal amygdala. J. Neurosci. *27*, 5958-5966.

Hutton,C., Deichmann,R., Turner,R., and Anderson,J.M. (2004). Combined correction for geometric distortion and its interaction with head motion in fMRI. Proceedings of ISMRM 12, Kyoto, Japan.

Hutton,C., Bork,A., Josephs,O., Deichmann,R., Ashburner,J., and Turner,R. (2002). Image distortion correction in fMRI: A quantitative evaluation. Neuroimage. *16*, 217-240.

Itti,L., and Baldi,P. (2005). Bayesian surprise attracts human attention. NIPS.

Jensen,J., Smith,A.J., Willeit,M., Crawley,A.P., Mikulis,D.J., Vitcu,I., and Kapur,S. (2007). Separate brain regions code for salience vs. valence during reward prediction in humans. Human Brain Mapping *28*, 294-302.

Ji,W., Suga,N., and Gao,E. (2005). Effects of agonists and antagonists of NMDA and ACh receptors on plasticity of bat auditory system elicited by fear conditioning. J. Neurophysiol. *94*, 1199-1211.

Jiao,H., Zhang,L., Gao,F., Lou,D., Zhang,J., and Xu,M. (2007). Dopamine D(1) and D(3) receptors oppositely regulate. J. Neurochem. *103*, 840-848.

Joel,D., and Weiskopf,N. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organisation of the striatum. Neuroscience *96(3)*, 451-457.

Kalisch,R., Korenfeld,E., Stephan,K.E., Weiskopf,N., Seymour,B., and Dolan,R.J. (2006). Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. J. Neurosci. *26*, 9503-9511.

Kamin LJ (1969). Selective association and conditioning. In Fundamental issues in instrumental learning, N.J. Mackintosh, and Honig W.K., eds. (Halifax, Canada: Dalhousie Univ. Press), pp. 42-64.

Kiebel,S.J., Garrido,M.I., Moran,R.J., and Friston,K.J. (2008). Dynamic causal modelling for EEG and MEG. Cogn Neurodyn. *2*, 121-136.

Kirkwood,A., Rozas,C., Kirkwood,J., Perez,F., and Bear,M.F. (1999). Modulation of long-term synaptic depression in visual cortex by acetylcholine and norepinephrine. J. Neurosci. *19*, 1599-1609.

Knight,D.C., Smith,C.N., Cheng,D.T., Stein,E.A., and Helmstetter,F.J. (2004). Amygdala and hippocampal activity during acquisition and extinction of human fear conditioning. Cogn Affect. Behav. Neurosci. *4*, 317-325.

Kording,K.P., Beierholm,U., Ma,W.J., Quartz,S., Tenenbaum,J.B., and Shams,L. (2007). Causal inference in multisensory perception. PLoS. ONE. *2*, e943.

LaBar,K.S., Gatenby,J.C., Gore,J.C., LeDoux,J.E., and Phelps,E.A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron *20*, 937-945.

LaBar,K.S., LeDoux,J.E., Spencer,D.D., and Phelps,E.A. (1995). Impaired fear conditioning following unilateral temporal lobectomy in humans. J. Neurosci. *15*, 6846-6855.

LeDoux,J. (2003). The emotional brain, fear, and the amygdala. Cell Mol. Neurobiol. *23*, 727-738.

LeDoux,J. (2007). The amygdala. Curr. Biol. *17*, R868-R874.

Leh,S.E., Ptito,A., Chakravarty,M.M., and Strafella,A.P. (2007). Fronto-striatal connections in the human brain: a probabilistic diffusion tractography study. Neurosci. Lett. *419*, 113-118.

Lewis,D.A. (1991). Distribution of choline acetyltransferase-immunoreactive axons in monkey prefrontal cortex. Neuroscience *40(2)*, 363-374.

Ljungberg,T., Apicella,P., and Schultz,W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. J. Neurophysiol. *67*, 145-163.

Ma,X., and Suga,N. (2005). Long-term cortical plasticity evoked by electric stimulation and acetylcholine applied to the auditory cortex. Proc. Natl. Acad. Sci. U. S. A *102*, 9335-9340.

Maldjian,J.A., Laurienti,P.J., Kraft,R.A., and Burdette,J.H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage. *19*, 1233-1239.

Maren,S. (2001). Neurobiology of Pavlovian fear conditioning. Annu. Rev. Neurosci. *24*, 897-931.

Mars,R.B., Piekema,C., Coles,M.G., Hulstijn,W., and Toni,I. (2007). On the programming and reprogramming of actions. Cereb. Cortex *17*, 2972-2979.

Marschner,A., Kalisch,R., Vervliet,B., Vansteenwegen,D., and Buchel,C. (2008). Dissociable roles for the hippocampus and the amygdala in human cued versus context fear conditioning. J. Neurosci. *28*, 9030-9036.

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

McIntosh,A.R., Cabeza,R.E., and Lobaugh,N.J. (1998). Analysis of neural interactions explains the activation of occipital cortex by an auditory stimulus. J. Neurophysiol. *80*, 2790-2796.

McIntosh,A.R., and Gonzales-Lima,F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. Human Brain Mapping *2*, 2-22.

McLaren,I.P. (1989). The computational unit as an assembly of neurones: an implementation of an error correcting learning algorithm. In The computing neuron, Durbin R, Miall C, and Mitchison G, eds. (Amsterdam: Addison-Wesley), pp. 160-178.

McLaren,I.P., Kaye,H., and Mackintosh,N.J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In Parallel distributed processing: Implications for psychology and neurobiology, R.G.M. Morris, ed. (Oxford: Clarendon Press), pp. 102-120.

Menon,M., Jensen,J., Vitcu,I., Graff-Guerrero,A., Crawley,A., Smith,M.A., and Kapur,S. (2007). Temporal Difference Modeling of the Blood-Oxygen Level Dependent Response During Aversive Conditioning in Humans: Effects of Dopaminergic Modulation. Biol. Psychiatry *62*, 765-772.

Metherate,R., and Hsieh,C.Y. (2003). Regulation of glutamate synapses by nicotinic acetylcholine receptors in auditory cortex. Neurobiol. Learn. Mem. *80*, 285-290.

Mirenowicz,J., and Schultz,W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. J. Neurophysiol. *72*, 1024-1027.

Mirenowicz,J., and Schultz,W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. Nature *379*, 449-451.

Mogenson,G.J., Jones,D.L., and Yim,C.Y. (1980). From motivation to action: functional interface between the limbic system and the motor system. Prog. Neurobiol. *14*, 69-97.

Montague,P.R., and Berns,G.S. (2002). Neural economics and the biological substrates of valuation. Neuron *36*, 265-284.

Morris,J.S., and Dolan,R.J. (2004). Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. Neuroimage. *22*, 372-380.

Morris,J.S., Friston,K.J., and Dolan,R.J. (1998). Experience-dependent modulation of tonotopic neural responses in human auditory cortex. Proc. Biol. Sci. *265*, 649-657.

Morris,R.G. (1989). Synaptic plasticity and learning: selective impairment of learning rats and blockade of long-term potentiation in vivo by the N-methyl-D-aspartate receptor antagonist AP5. J. Neurosci. *9*, 3040-3057.

Murray,S.O., Kersten,D., Olshausen,B.A., Schrater,P., and Woods,D.L. (2002). Shape perception reduces activity in human primary visual cortex. Proc. Natl. Acad. Sci. U. S. A *99*, 15164-15169.

Naatanen,R., Gaillard,A.W., and Mantysalo,S. (1978). Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst) *42*, 313-329.

Naatanen,R., Tervaniemi,M., Sussman,E., Paavilainen,P., and Winkler,I. (2001). "Primitive intelligence" in the auditory cortex. Trends Neurosci. *24*, 283-288.

Nakahara,H., Itoh,H., Kawagoe,R., Takikawa,Y., and Hikosaka,O. (2004). Dopamine neurons can represent context-dependent prediction error. Neuron *41*, 269-280.

Nakayama,Y., Yamagata,T., Tanji,J., and Hoshi,E. (2008). Transformation of a virtual action plan into a motor plan in the premotor cortex. J. Neurosci. *28*, 10287-10297.

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

O'Doherty,J.P., Dayan,P., Friston,K., Critchley,H., and Dolan,R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329-337.

Pagnoni,G., Zink,C.F., Montague,P.R., and Berns,G.S. (2002). Activity in human ventral striatum locked to errors of reward prediction. Nat. Neurosci. *5*, 97-98.

Pearce,J.M., and Bouton,M.E. (2001). Theories of associative learning in animals. Annu. Rev. Psychol. *52*, 111-139.

Pearce,J.M., and Hall,G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. *87*, 532-552.

Penny,W.D., Stephan,K.E., Mechelli,A., and Friston,K.J. (2004a). Comparing dynamic causal models. Neuroimage. *22*, 1157-1172.

Penny,W.D., Stephan,K.E., Mechelli,A., and Friston,K.J. (2004b). Modelling functional integration: a comparison of structural equation and dynamic causal models. Neuroimage. *23 Suppl 1*, S264-S274.

Pessiglione,M., Schmidt,L., Draganski,B., Kalisch,R., Lau,H., Dolan,R.J., and Frith,C.D. (2007). How the brain translates money into force: a neuroimaging study of subliminal motivation. Science *316*, 904-906.

Pessiglione,M., Seymour,B., Flandin,G., Dolan,R.J., and Frith,C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042-1045.

Petrovic,P., Kalisch,R., Pessiglione,M., Singer,T., and Dolan,R.J. (2008). Learning affective values for faces is expressed in amygdala and fusiform gyrus. Soc. Cogn Affect. Neurosci. *3*, 109-118.

Pincze,Z., Lakatos,P., Rajkai,C., Ulbert,I., and Karmos,G. (2002). Effect of deviant probability and interstimulus/interdeviant interval on the auditory N1 and mismatch negativity in the cat auditory cortex. Brain Res. Cogn Brain Res. *13*, 249-253.

Pitt,M.A., and Myung,I.J. (2002). When a good fit can be bad. Trends Cogn Sci. *6*, 421-425.

Raftery,A.E. (1995). Bayesian model selection in social research. In Sociological Methodology, P.V. Marsden, ed. (Cambridge, MA: pp. 111-196.

Rao,R.P., and Ballard,D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79-87.

Repa,J.C., Muller,J., Apergis,J., Desrochers,T.M., Zhou,Y., and LeDoux,J.E. (2001). Two different lateral amygdala cell populations contribute to the initiation and storage of memory. Nat. Neurosci. *4*, 724-731.

Requin,J., and Granjon,M. (1969). The effect of conditional probability of the response signal on the simple reaction time. Acta Psychol. (Amst) *31*, 129-144.

Rescorla,R.A., and Wagner,A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Classical Conditioning II: Current Research and Theory, A.H. Black, and W.F. Prokasy, eds. (New York: Appleton Century Crofts), pp. 64-99.

Rockland,K.S., and Ojima,H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. Int. J. Psychophysiol. *50*, 19-26.

Rodriguez,P.F., Aron,A.R., and Poldrack,R.A. (2006). Ventral-striatal/nucleus-accumbens sensitivity to prediction errors during classification learning. Hum. Brain Mapp. *27*, 306-313.

Romo,R., and Schultz,W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J. Neurophysiol. *63*, 592-606.

Salazar-Colocho,P., Del,R.J., and Frechilla,D. (2007). Serotonin 5-hT1A receptor activation prevents phosphorylation of NMDA receptor NR1 subunit in cerebral ischemia. J. Physiol Biochem. *63*, 203-211.

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

Schultz,W. (2000). Multiple reward signals in the brain. Nat. Rev. Neurosci. *1*, 199-207.

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

Schultz,W., and Dickinson,A. (2000). Neuronal coding of prediction errors. Annu. Rev. Neurosci. *23*, 473-500.

Schwarz,G. (1978). Estimating the dimension of a model. Ann. Stat. *6*, -421.

Seitz,A.R., Kim,D., and Watanabe,T. (2009). Rewards Evoke Learning of Unconsciously Processed Visual Stimuli in Adult Humans. Neuron *61(5)*, 700-707.

Seymour,B., Daw,N., Dayan,P., Singer,T., and Dolan,R. (2007). Differential encoding of losses and gains in the human striatum. J. Neurosci. *27*, 4826-4831.

Seymour,B., O'Doherty,J.P., Dayan,P., Koltzenburg,M., Jones,A.K., Dolan,R.J., Friston,K.J., and Frackowiak,R.S. (2004). Temporal difference models describe higher-order learning in humans. Nature *429*, 664-667.

Shanks,D.R. (1995). The Psychology of Associative Learning (Cambridge, UK: Cambridge University Press).

Shergill,S.S., Bays,P.M., Frith,C.D., and Wolpert,D.M. (2003). Two eyes for an eye: the neuroscience of force escalation. Science *301*, 187.

Shergill,S.S., Samson,G., Bays,P.M., Frith,C.D., and Wolpert,D.M. (2005). Evidence for sensory prediction deficits in schizophrenia. Am. J. Psychiatry *162*, 2384-2386.

Shi,C., and Davis,M. (1999). Pain pathways involved in fear conditioning measured with fear-potentiated startle: lesion studies. J. Neurosci. *19*, 420-430.

Sotres-Bayon,F., Diaz-Mataix,L., Bush,D.E., and LeDoux,J.E. (2009). Dissociable roles for the ventromedial prefrontal cortex and amygdala in fear extinction: NR2B contribution. Cereb. Cortex *19*, 474-482.

Stagg,C., Hindley,P., Tales,A., and Butler,S. (2004). Visual mismatch negativity: the detection of stimulus change. Neuroreport *15*, 659-663.

Stephan,K.E. (2004). On the role of general system theory for functional neuroimaging. J. Anat. *205*, 443-470.

Stephan,K.E., Baldeweg,T., and Friston,K.J. (2006). Synaptic plasticity and dysconnection in schizophrenia. Biol. Psychiatry *59*, 929-939.

Stephan,K.E., Harrison,L.M., Kiebel,S.J., David,O., Penny,W.D., and Friston,K.J. (2007a). Dynamic causal models of neural system dynamics:current state and future extensions. J. Biosci. *32*, 129-144.

Stephan,K.E., Kasper,L., Harrison,L.M., Daunizeau,J., den Ouden,H.E., Breakspear,M., and Friston,K.J. (2008). Nonlinear dynamic causal models for fMRI. Neuroimage. *42*, 649-662.

Stephan,K.E., Marshall,J.C., Penny,W.D., Friston,K.J., and Fink,G.R. (2007b). Interhemispheric integration of visual processing during task-driven lateralization. J. Neurosci. *27*, 3512-3522.

Stephan,K.E., Marshall,J.C., Penny,W.D., Friston,K.J., and Fink,G.R. (2007c). Interhemispheric integration of visual processing during task-driven lateralization. J. Neurosci. *27*, 3512-3522.

Stephan,K.E., Penny,W.D., Daunizeau,J., Moran,R.J., and Friston,K.J. (2009). Bayesian model selection for group studies. Neuroimage.

Stephan,K.E., Weiskopf,N., Drysdale,P.M., Robinson,P.A., and Friston,K.J. (2007d). Comparing hemodynamic models with DCM. Neuroimage. *38*, 387-401.

Strange,B.A., Duggins,A., Penny,W., Dolan,R.J., and Friston,K.J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw. *18*, 225-230.

Summerfield,C., Egner,T., Greene,M., Koechlin,E., Mangels,J., and Hirsch,J. (2006). Predictive codes for forthcoming perception in the frontal cortex. Science *314*, 1311-1314.

Summerfield,C., and Koechlin,E. (2008). A neural representation of prior information during perceptual inference. Neuron *59*, 336-347.

Summerfield,C., Trittschuh,E.H., Monti,J.M., Mesulam,M.M., and Egner,T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. Nat. Neurosci.

Sutton,R.S., and Barto,A.G. (1990). Time-derivative models of pavlovian reinforecement. In Learning and Computational Neuroscience: Foundations of Adaptive Networks, M. Gabriel, and J. More, eds. MIT press), pp. 497-537.

Sutton,R.S., and Barto,A.G. (1998). Reinforcement Learning: an Introduction (Cambridge, MA: MIT press).

Tabbert,K., Stark,R., Kirsch,P., and Vaitl,D. (2005). Hemodynamic responses of the amygdala, the orbitofrontal cortex and the visual cortex during a fear conditioning paradigm. Int. J. Psychophysiol. *57*, 15-23.

Takada,M., Tokuno,H., Nambu,A., and Inase,M. (1998). Corticostriatal projections from the somatic motor areas of the frontal cortex in the macaque monkey: segregation versus overlap of input zones from the primary motor cortex, the supplementary motor area, and the premotor cortex. Exp. Brain Res. *120*, 114-128.

Tanji,J., and Evarts,E.V. (1976). Anticipatory activity of motor cortex neurons in relation to direction of an intended movement. J. Neurophysiol. *39*, 1062-1068.

Thiel,C.M., Bentley,P., and Dolan,R.J. (2002a). Effects of cholinergic enhancement on conditioning-related responses in human auditory cortex. Eur. J. Neurosci. *16*, 2199-2206.

Thiel,C.M., Friston,K.J., and Dolan,R.J. (2002b). Cholinergic modulation of experience-dependent plasticity in human auditory cortex. Neuron *35*, 567-574.

Thiel,C.M., Henson,R.N., and Dolan,R.J. (2002c). Scopolamine but not lorazepam modulates face repetition priming: a psychopharmacological fMRI study. Neuropsychopharmacology *27*, 282-292.

Thorndike E.L. (1898). Animal intelligence: An experimental study of the associative processes in animals. Psychological Review, Monograph Supplement *24(8)*, 1-109.

Tobler,P.N., Fiorillo,C.D., and Schultz,W. (2005). Adaptive coding of reward value by dopamine neurons. Science *307*, 1642-1645.

Tobler,P.N., O'Doherty,J.P., Dolan,R.J., and Schultz,W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. J. Neurophysiol. *95*, 301-310.

Tobler,P.N., O'Doherty,J.P., Dolan,R.J., and Schultz,W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J. Neurophysiol. *97*, 1621-1632.

Tom,S.M., Fox,C.R., Trepel,C., and Poldrack,R.A. (2007). The neural basis of loss aversion in decision-making under risk. Science *315*, 515-518.

Trepel,C., Fox,C.R., and Poldrack,R.A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. Brain Res. Cogn Brain Res. *23*, 34-50.

Turner,D.C., Aitken,M.R., Shanks,D.R., Sahakian,B.J., Robbins,T.W., Schwarzbauer,C., and Fletcher,P.C. (2004). The role of the lateral frontal cortex in causal associative learning: exploring preventative and super-learning. Cereb. Cortex *14*, 872-880.

Tye,K.M., Stuber,G.D., de,R.B., Bonci,A., and Janak,P.H. (2008). Rapid strengthening of thalamo-amygdala synapses mediates cue-reward learning. Nature *453*, 1253-1257.

von Kriegstein,K., and Giraud,A.L. (2006). Implicit multisensory associations influence voice recognition. PLoS. Biol. *4*, e326.

Vuilleumier,P., Richardson,M.P., Armony,J.L., Driver,J., and Dolan,R.J. (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. Nat. Neurosci. *7*, 1271-1278.

Waelti,P., Dickinson,A., and Schultz,W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. Nature *412*, 43-48.

Wasserman,E.A., Elek,S.M., Chatlosh,D.L., and Baker,A.G. (1993). Rating causal relations: Role of probability in judgments of response^outcome contingency. J. Exp. Psychol. Learn. Mem. Cogn. *19*, 174-188.

Watkins,S., Shams,L., Tanaka,S., Haynes,J.D., and Rees,G. (2006). Sound alters activity in human V1 in association with illusory visual perception. Neuroimage. *31*, 1247-1256.

Weinberger,N.M. (2004). Specific long-term memory traces in primary auditory cortex. Nat. Rev. Neurosci. *5*, 279-290.

Weinberger,N.M. (2007). Associative representational plasticity in the auditory cortex: a synthesis of two disciplines. Learn. Mem. *14*, 1-16.

Whiteley,L., and Sahani,M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. J. Vis. *8*, 2-15.

Wise,S.P., and Mauritz,K.H. (1985). Set-related neuronal activity in the premotor cortex of rhesus monkeys: effects of changes in motor set. Proc. R. Soc. Lond B Biol. Sci. *223*, 331-354.

Wittmann,B.C., Bunzeck,N., Dolan,R.J., and Duzel,E. (2007). Anticipation of novelty recruits reward system and hippocampus while promoting recollection. Neuroimage. *38*, 194-202.

Wittmann,B.C., Daw,N.D., Seymour,B., and Dolan,R.J. (2008). Striatal activity underlies novelty-based choice in humans. Neuron *58*, 967-973.

Wolf,M.E., Mangiavacchi,S., and Sun,X. (2003). Mechanisms by which dopamine receptors may influence synaptic plasticity. Ann. N. Y. Acad. Sci. *1003*, 241-249.

Wolpert,D.M., Doya,K., and Kawato,M. (2003). A unifying computational framework for motor control and social interaction. Philos. Trans. R. Soc. Lond B Biol. Sci. *358*, 593-602.

Wolpert,D.M., Ghahramani,Z., and Jordan,M.I. (1995). An internal model for sensorimotor integration. Science *269*, 1880-1882.

Yacubian,J., Glascher,J., Schroeder,K., Sommer,T., Braus,D.F., and Buchel,C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J. Neurosci. *26*, 9530-9537.

Yacubian,J., Sommer,T., Schroeder,K., Glascher,J., Braus,D.F., and Buchel,C. (2007). Subregions of the ventral striatum show preferential coding of reward magnitude and probability. Neuroimage. *38*, 557-563.

Yoshida,W., Dolan,R.J., and Friston,K.J. (2008). Game theory of mind. PLoS. Comput. Biol. *4*, e1000254.

Zink,C.F., Pagnoni,G., Chappelow,J., Martin-Skurski,M., and Berns,G.S. (2006). Human striatal activation reflects degree of stimulus saliency. Neuroimage. *29*, 977-983.

Zink,C.F., Pagnoni,G., Martin,M.E., Dhamala,M., and Berns,G.S. (2003). Human striatal response to salient nonrewarding stimuli. J. Neurosci. *23*, 8092-8097.