The selective updating of working memory: a predictive coding account

Yen Yu

Wellcome Trust Centre for Neuroimaging
Institute of Neurology

A thesis submitted for the degree of Doctor of Philosophy
University College London
August 2014

Primary supervisor: Professor Karl J. Friston
Secondary supervisor: Dr. William D. Penny

Declaration

I, Yen Yu, 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.

Abstract

Goal-relevant information maintained in working memory is remarkably robust and resistant to distractions. However, our nervous system is endowed with exceptional flexibility; therefore such information can be updated almost effortlessly. A scenario – not uncommon in our daily life – is that selective maintaining and updating information can be achieved concurrently. This is an intriguing example of how our brain balances stability and flexibility, when organising its knowledge. A possibility – one may draw upon to understand this capacity – is that working memory is represented as beliefs, or its probability densities, which are updated in a context-sensitive manner. This means one could treat working memory in the same way as perception – i.e., memories are based on inferring the cause of sensations, except that the time scale ranges from an instant to prolonged anticipation. In this setting, working memory is susceptible to prior information encoded in the brain's model of its world. This thesis aimed to establish an interpretation of working memory processing that rests on the (generalised) predictive coding framework, or hierarchical inference in the brain. Specifically, the main question it asked was how anticipation modulates working memory updating (or maintenance). A novel working memory updating task was designed in this regard. Blood-oxygen-level dependent (BOLD) imaging, machine learning, and dynamic causal modelling (DCM) were applied to identify the neural correlates of anticipation and the violation of anticipation, as well as the causal structure generating these neural correlates. Anticipation induced neural activity in the dopaminergic midbrain and the striatum. Whereas, the fronto-parietal and cingulo-operculum network were implicated when an anticipated update was omitted, and the midbrain, occipital cortices, and cerebellum when an update was unexpected. DCM revealed that anticipation is a modulation of backward connections, whilst the associated surprise is mediated by forward and local recurrent modulations. Two mutually antagonistic pathways were differentially modulated under anticipatory flexibility and stability, respectively. The overall results indicate that working memory may as well follow the cortical message-passing scheme that enables hierarchical inference.

Acknowledgements

I owe thanks to too many during my four years abroad.

First, I would like to express my sincere gratitude to my supervisors, Karl Friston, for his constant patience, encouragement, and scientific insights that shaped my path; and Will Penny, for his friendship, mentoring, and time spent with me on problem solving. My warmest thanks go to current and former members of the FIL, in particular, Marcia Bennett, for her kindness and assistance of all sorts; Thomas FitzGerald, for his inputs to my study and manuscripts; Marta Garrido, for her friendship and advice; Dimitris Pinotsis, for his help on the SHC model; and Klaas Stephan, for his constructive advice while the work of this thesis is still at a very early stage. I would also like to thank Jan, Alfonso, and Eric for their help in collecting data. Without these people, I could not have finished this thesis – it is my privilege to work with you.

I am grateful for the enormous joy and support from my friends in Taiwan and London: Professor Yu-Te Wu, for his timely and thoughtful letters; my flatmate, Sabrina Hsu, for helping me with chores while I was writing up; Kami Hsu and Willy Tseng, for their invitations, meals, and wonderful banters; Tsu-Jui Cheng, for our 'group therapy'; Tomohiro Ishizu and Sungho Tak, for their warm friendship; and Carlton Chu, for his guidance.

Finally, I must express my deepest gratitude to my families, my life and all that is involved is meaningless without you.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Overarching theme and research questions

Uncertainties and random fluctuations are intrinsic to any physical system, irrespective of whether they are biological or non-biological. In evolutionary terms, any biological entity may not exist without the ability to maintain homeostasis within a certain range of uncertainty. To maintain homeostasis, one must resolve the mutual antagonism between stability and flexibility. These can be expressed on many levels: from instantaneous (e.g., a motor reflex to scalding) to anticipatory (e.g., calculating the altitude at which to deploy a parachute, or buying in shares). The better an organism can represent the causal structure of the environment, in its information processing infrastructure, the better its ability to infer the hidden states of the world and their trajectories, and to implement contingencies given anticipated fluctuations in the hidden states.

Working memory emerges when a causal relationship between an organism's internal states and external states is inferred with high fidelity (Postle, 2006). It endows us with the ability to generate a piece of information upon an environmental cue and to continue to retain the information after the cue is extinguished. The idea of working memory embodies the retention of information because the organism that employs working memory is predisposed to the guidance of such information in its course of action. In other words, the organism believes *a priori* the possessing of specific information will be of prospective advantage.

While working memory entails beliefs about stable environmental states that their realisation in the immediate future, an equally potent aspect of working

memory is concerned with the updating of beliefs by exchange with another – i.e., updating of working memory (Miller & Cohen, 2001). Maintaining and updating working memory therefore speak to the conflicting demands of stable (precise) and flexible (uncertain) belief. As such, representing anticipatory fluctuations in working memory is equivalent to inferring environmental volatility.

The overarching theme of this thesis thus rests on anticipatory fluctuations and updating in working memory, and is developed through addressing research questions built on a number of assumptions. First, it treats working memory updating as a manifestation of cognitive flexibility; in other words, an update entails set-switching. Probabilistic evaluations of a set (in an anticipatory manner) are referred to as an anticipatory set. Secondly, dopaminergic innervations are a candidate neuromodulator that nuances the balance between stability and flexibility of working memory. The function of dopamine is characterised by tonic and phasic modes of discharge, and action through different receptor subtypes. Thirdly, the frontoparietal network, the basal ganglia, and the sensory cortices support working memory function. Finally, working memory processing may also follow the principle of hierarchical inference (or generalised predictive coding).

In brief, research questions pertain to:

Question 1: What are the behavioural relevance and the neural correlates of anticipatory set; is dopamine critical to working memory, does predicting an update implicate dopaminergic responses?

Question 2: Invalid anticipatory set is followed by improbable (or surprising) updating or maintenance of working memory, if these represent prediction error

responses then they require exogenous and/or endogenous drives for error processing; to what extent are they dissociable in terms of neural responses?

Question 3: If anticipatory set (in working memory) reflects prediction and surprise reflects prediction error, do they follow the principle of hierarchical inference in the brain by providing appropriate forward/backward influences?

In the following sections, I give a brief review of the development of working memory as a psychological construct and its biological relevance, followed by a particular focus on the neural mechanisms of working memory updating, citing the notion of 'central executive'. The neuromodulation of dopamine is then introduced and linked to recent findings in working memory. Also, the influence of dopaminergic modulations in the development of neurocomputational models of working memory is reviewed. Next, the seminal work on the prefrontal cortex-basal ganglia working memory (PBWM) model is introduced to show how the basal ganglia may enable working memory updating. Finally, I focus on the increasingly popular notion that the brain employs hierarchical inference, which may be intrinsically related to working memory processes.

1.2. A brief history of working memory

The development of working memory as a psychological construct can be traced back to the Jamesian conceptualisation (James, 1890), in which the distinction between a temporary *primary memory* and a more stabilised *secondary memory* was proposed. The theory did not aggregate much interest until the 1950s when Donald Hebb (1949) postulated two separate memory systems, short-term memory (STM) and long-term memory (LTM), which were later corroborated on empirical grounds (J. Brown, 1958; L. R. Peterson & Peterson, 1959). This had stirred up debates

(during the 1960s) as to whether a two-process memory system was necessary. Critics citing proactive interference theory challenged the view of memory trace decay, on which STM was largely theorised. Specifically, STM suffers from trace decay if its content was not rehearsed. The debate was not settled until the double dissociation based on patient studies was established in the 1970s. Patients with bilateral damage to the temporal lobe and hippocampus were shown to have reduced capacity for LTM performance, but performance in STM tasks was comparable to normal controls (Baddeley & Warrington, 1970). Shallice and Warrington (1970) demonstrated another class of patients who showed completely opposite deficits, with normal LTM but impaired STM performance. This was about the time when models of multi-process memory systems started to gain favour. Among them, perhaps the most influential one was the modal model, proposed by Atkinson and Shiffrin (1968). The modal model holds a subsystem of unitary short-term store, acting as working memory, which is of limited capacity and capable of manipulating information. In addition, the short-term store is solely responsible for encoding and retrieving LTM.

The major problem with the modal model was that it requires information to be held in STM in order to enter and formulate LTM. The model thus predicts that patients having STM dysfunction may not have functional LTM. However, the work by Shallice and Warrington (1970) established clearly the opposite case. The inconsistency inherent in the modal model advocates alternative perspectives, that STM may not serve as a general purpose working memory – as STM patients are often found functional in life. This motivated Alan Baddeley and Graham Hitch's seminal working memory model in the 1970s, from which the ensuing review originates.

Their study of the memory system took on a dual-task approach. One component of the task involved repeatedly reciting a sequence of random digits to prevent the subjects from articulatory rehearsal. The reciting was assumed to take up short-term storage capacity as the sequence size progressed. They based their assumption on that by the time all existing models agreed that the immediate serial recall task depends on STM, which is limited in capacity. The other component was a verbal reasoning task (Hitch & Baddeley, 1976). They found, with the increase of concurrent digit load, the mean reasoning time increase but no effect on accuracy. Similar results were reported using the task involving comprehension or long-term learning (Baddeley, 1986), invalidating the assumption of unitary short-term storage, which was then replaced with a multi-component working memory (Figure 1.1). Simply put, it was inferred that the STM process and the reasoning process are subserved by different component processes of working memory.

Figure 1.1 Baddeley's multi-component working memory model. A tripartite system of working memory conceptualised by Alan Baddeley and Graham Hitch in 1974. The model includes two 'slave' systems, the phonological loop and the visuospatial sketchpad, and one central executive. As the names suggest, the two slave systems are in close connection with domain-specific sensory interface, whereas the central executive has a domain-general role, which processes the 'process of sensations'. The central executive is thus strongly associated with attention and the Supervisory Attentional System proposed by Norman and Shallice in 1983.

Baddeley's multi-component working memory model consists of two subsystems, the phonological loop and the visuospatial sketchpad, both of which are

left out in the course of this thesis; as well as a supra-ordinate *central executive* (Baddeley, 1992).

The idea of the central executive, otherwise known as executive control or executive function (Baddeley, 2007) is, at times, referred to as 'general purpose control mechanisms' (Miyake et al., 2000). Fractionation of executive function was needed to put the theory to test. Initially, Baddeley refers to the Supervisory Attentional System (SAS) proposed by Norman and Shallice (1983) as a candidate model for central executive. There are two levels of behavioural control according to SAS, one underpins habits, a collection of nearly automatic, effortless mental states; the other refers to the mechanisms to overcome such automaticity, the SAS *per se*. Notably, the SAS was largely conceptualised based on observations during which frontal lobe patients control their behaviour. Two seemingly paradoxical outcomes came into focus: the patients either showed rigidity, perseverating with the same pattern, unable to switch action, or they were extremely susceptible to perceived stimuli, showing great distractibility. This means of concept of central executive was not only to incorporate the psychological construct of attention, it also implicates, as a constituent component of working memory, variability in representational power, i.e., the control that balances robustness against adaptiveness (Miller & Cohen, 2001). Moreover, the fontal lobe, according to the patient study, seems to provide the functional architecture for executive function. Indeed, Baddeley (1986) coined the term *dysexecutive function*, referring central executive impairments as a neurological disorder which follows frontal lobe damage. The initial working memory model, which assumed the central executive is capable of attentional focus, storage, and decision making, was later revised to include four component processes – namely,

the capacity to focus attention, to divide attention, to switch attention, and to bridge working memory and long-term memory (Baddeley, 2012; 2007).

Around the same time that Baddeley proposed the multi-component working memory model, another line of research in monkeys demonstrated a remarkable observation. Joaquin Fuster (Fuster, 1973; Fuster & Alexander, 1973) and other researchers showed that the monkey prefrontal cortex exhibits sustained activity at the single neuron level throughout the delay phase of a delayed-response task. The implication of this and other similar findings corroborates the notion of 'reverberation' as a mechanism of a stimulus-induced transient memory (Hebb, 1949). Most importantly, the finding speaks to the correspondence between a psychological construct and a physiological phenomenon. The integration of neurobiological and psychological concepts was further advanced by Goldman-Rakic with the finding that the where/what organisation of the visual system may also apply to visual working memory (Compte, Brunel, Goldman-Rakic, & Wang, 2000; Goldman-Rakic, 1995). In particular, Goldman-Rakic and colleagues used monkey electrophysiology to show not only single neurons exhibit where/what selectivity but also that the neurons can be classified into cue-sensitive, delay-sensitive, and probe-sensitive groups (Goldman-Rakic, 1995; 1996) – as if there were 'memory fields' for processes and a topographic map of the sensorimotor cortex. This suggests that prefrontal neuronal activity is associated with sensory signals in a way that the PFC neurons 'remember' the channel from which the sensory signal originates. In addition, the prefrontal neurons are able to link such 'memory' to guide behaviours that are set apart temporally from sensory signals. This forms a neural code of an 'episode' that seems to be an ad hoc integration of

information across time (Kahneman & Treisman, 1984; Kahneman, Treisman, & Gibbs, 1992).

From Fuster to Goldman-Rakic, it has been established that the prefrontal cortex is central to working memory and may be a crucial neural substrate for executive function: Miller and Cohen (Miller & Cohen, 2001) proposed a model that describes the prefrontal cortex as providing biasing signals for enabling sensorimotor 'channels' in the association cortex. However, what and how information is represented by the prefrontal cortex remains elusive (Sreenivasan, Curtis, & D'Esposito, 2014). Recently, major working memory research has turned to ask how working memory information is embedded in long-term memory (N. Cowan, 2008; Lewis-Peacock & Postle, 2008; McElree, 2001; Oberauer, 2003), the neural circuitry implementing the flexible biasing mechanism (M. J. Frank, Loughry, & O'Reilly, 2001; Hazy, Frank, & O'Reilly, 2006; McNab & Klingberg, 2008), and the variable precision in representational power (Bays & Husain, 2008; N. Cowan, 2005; Gorgoraptis, Catalao, Bays, & Husain, 2011; Ma, Husain, & Bays, 2014).

1.3. Updating working memory

Updating of working memory representations refers to the process of "monitoring and coding incoming information for relevance to the task at hand and then appropriately revising the items held in working memory by replacing old, no longer relevant information, with newer, more relevant information" (Miyake et al., 2000). The idea of updating as a component of working memory emerges with attempts to fractionate the central executive. In cognitive psychology, Morris and Jones (1990) conducted the running memory task and demonstrated that updating

memory affects performance independently of the effect of irrelevant speech and suppression. Under Baddeley's working memory model (Baddeley, 1992), the system engaged in speech effects is the phonological loop. The authors therefore suggested that the updating process places demands on a superordinate system, i.e., the central executive. Using latent variable analysis, Miyake et al. (Miyake et al., 2000) tested three sets of cognitive tasks, each is considered to tap one of the proposed executive functions: set-shifting, updating, and inhibition. Their findings indicated that the three proposed executive functions are co-dependent, yet clearly distinguishable.

It is conceivable that updating is intertwined with the other two executive functions proposed by Miyake et al. (2000), in ways that cannot be addressed by the work of Morris et al. (1990). The running memory task was severely limited by the fact that it is characterised by distinct recency, but not primacy, effects. It has been argued that if the subjects were to process and hold all information on-line, then one might expect both primacy and recency effects as one would in a standard serial recall paradigm (Bunting, Cowan, & Saults, 2006; Palladino & Jarrold, 2008). In other words, the subjects performing the task may have employed a passive strategy, instead of actively engaging in the memory updating per se. Palladino et al. (2001) avoided these shortcomings and devised an updating task that was not subject to recency effects nor temporal criterion. They inferred, based on a measure of intrusion error rates, that the ability to inhibit irrelevant information in working memory is a critical variable to determine updating performance that underlies successful encoding of new goal-relevant information.

Another useful notion of working memory updating – with regard to executive function – is the distinction between *binding updating* and *content updating* (Artuso

& Palladino, 2011). The idea of binding updating stems from Kahneman's 'object file' (Kahneman et al., 1992; Kahneman & Treisman, 1984), in which a memory object is an ad hoc integration of different pieces of information (see also Postle, 2006; Sreenivasan et al., 2014) that are bound into a singular entity. Once the entity is formed, it may stay intact whilst its constituent information undergoes partial modification. Binding updating is therefore closely related to selective updating (Kessler & Meiran, 2008; Murty et al., 2011; Nee & Brown, 2013), whereas content updating is related to total updating (Kessler & Meiran, 2008). Generally, an updating cost can be identified when comparing reaction time measures between updating and non-updating trials (Kessler & Meiran, 2006). This (object switching) cost was first reported by Garavan (1998) and was interpreted as shifting the focus of attention to an object not currently attended, citing the hypothesis of embedded working memory (McElree:1998kw ; see also pp. 117 of Baddeley, 2007; i.e., selective attention to active long-term memory representation; N. Cowan, 1988). These findings confer working memory updating with a component of attentional control and a characteristic of set-switching. Compared with (content) updating cost, human subjects suffered from greater (binding) costs when engaging in binding updating (Artuso & Palladino, 2011). More recently, evidence has emerged regarding the possibility that binding updating involves processes beyond simple re-encoding, whilst content updating is more likely to entail re-encoding alone (Artuso & Palladino, 2014). Artuso and Palladino's (2014) finding marks a departure from the common conception that working memory updating resembles encoding (R. C. O'Reilly & Frank, 2006).

Binding and content updating may be treated as component processes in working memory updating. Other work also employed similar tasks to decompose the

updating process. In particular, Ecker et al. (Ecker, Lewandowsky, Oberauer, & Chee, 2010) proposed that an updating operation may involve a combination of three component processes: retrieval (access), transformation, and substitution. Retrieval, or more precisely, to access already-retrieved information, involves focusing on previously unattended information; transformation refers to incorporating new information; substitution is similar to the construct of content updating mentioned earlier. Using latent variable analysis, the authors concluded that the three component processes appear to be independent and, of note, that individual differences in working memory capacity are a strong predictor to retrieval and transformation, but not substitution, performance.

Taken together, working memory updating may be a multi-component process and interact with other aspects of executive functions – although the interdependency between component processes and executive function, or whether such fractionation creates redundancy, remains to be determined. Nevertheless, these notions on binding and executive functions speak to the remarkably flexible functionality of otherwise stable working memory, which seems to sit comfortably with known functional anatomy. Notably, the prefrontal cortex holds two hallmarks – that its damage causes both perseveration (inadequate updating) and increased distractibility (inappropriate updating; Miller & Cohen, 2001). But with normal function, the prefrontal cortex provides biasing signals to the association and sensorimotor cortex, which may be the neural basis for mediating binding (Treisman & Gelade, 1980). In addition, the basal ganglia may be capable of implementing inhibition (see Aron, 2007 for review) and set-switching (e.g., Hikosaka & Isoda, 2010). Critically, both the prefrontal cortex and the basal ganglia are densely innervated by dopaminergic

inputs. Ample evidence has indicated that dopaminergic modulations are a potent factor that nuances flexibility and stability in working memory.

Working memory updating has been an active area of neuroimaging study since the late 1990s (see Salmon et al., 1996). Early reports focused on the localisation of updating-related neural activity. Wager and Smith (2003) summarised findings from 60 neuroimaging studies of working memory, reporting peak activations in a meta-analysis. Their analysis showed that regions consistently activated during updating tasks (e.g., *n*-back) are found in Brodmann's area (BA) 6, 8, 9, and 7, and are predominantly right lateralised. Later, in a meta-analysis specifically focused on the *n*-back paradigm, Owen et al. (2005) demonstrated that working memory updating involves bilateral cortical activation mainly in the fronto-parietal network, frontal pole, anterior cingulate cortex, insular, thalamus, and cerebellum. None of these two works had addressed the involvement of the basal ganglia in working memory updating.

Our understanding about the functional role of the parietal cortex in memory updating is perhaps as limited as that of the prefrontal cortex. However, in a study that followed the paradigm in Miyake et al. (Miyake et al., 2000), Collette and colleagues (2005) provided evidence that updating, as well as other executive components, recruits the intraparietal sulcus (BA 40). The authors suggested that the common activation in the intraparietal sulcus reflects selective attention to relevant stimuli and suppression of irrelevant information. Their conclusion was in line with the finding in Vogel et al. (2005) and in McNab et al. (2008).

Since the influential working memory model proposed by Frank et al. (Hazy et al., 2006), the functional role of the basal ganglia in working memory has been an active topic. One notable work is by McNab and Klingberg (2008). In this work, the

authors demonstrated that the 'filtering set' activity in the basal ganglia predicted individual working memory capacity and was inversely related to the activity in the parietal cortex. The filtering set activity was induced by predictive cues about whether or not distractors are present in the upcoming stimuli. Their findings corroborate the notion that the basal ganglia controls information represented in working memory. Consistent with this notion, Murty et al. (2011) showed that the meso-cortico-striatal activity was specifically modulated by selective updating of working memory (see also Podell et al., 2012). Few studies have reported the midbrain activity in association with working memory updating. Recently, in a careful set-up with cardiac-gating imaging sequences and transcranial magnetic stimulation, D'Ardenne et al. (2012) showed that the dopaminergic midbrain is indeed activated during memory updating and implicates phasic dopamine discharge (D'Ardenne, McClure, Nystrom, & Cohen, 2008).

Overall, working memory updating may involve multiple psychological constructs and relate to different aspects of executive function. The binding notion seems to fit with the functional role of the prefrontal cortex, the basal ganglia and dopamine modulation.

1.4. Neuromodulations

1.4.1. Anatomy of dopamine

Dopamine has a profound influence in the functioning of multiple aspects of cognition, including working memory. Dopaminergic projection neurons – i.e., neurons that use dopamine as a primary neurotransmitter – are found primarily in the

substantia nigra pars compacta (SNc) of the basal ganglia and the ventral tegmental area (VTA) in the midbrain. The other known dopamine-secreting areas include several nuclei in the hypothalamus and subthalamus (Iversen, 2010). The functional significance of dopamine can be seen from the diversity of projection pathways it gives rise to. Notably, the mesocortical and nigrostriatal pathways are associated with some key areas that provide the neural basis of working memory. The mesocortical pathway arises from the VTA and targets the cerebral cortex, with a convergence in the prefrontal cortex, especially the dorsolateral prefrontal cortex (Iversen, 2010); the nigrostriatal pathway, on the other hand, targets the striatum from the SNc. The nigrostriatal pathway targets specifically the matrix compartment of the striatum, where striatal medium spiny neurons (MSNs) form the direct and indirect pathways with other nuclei of the basal ganglia. The matrix compartment also receives cortical afferents, principally from superficial layer V.

Dopamine binds to two receptor types that are categorised pharmacologically based on their ligand recognition properties and effects on cAMP production: the D1- and D2-family. In general, these receptors are located postsynaptically, especially for the D1-family receptors (Levey et al., 1993). The D2-family, by contrast, may be found postsynaptically on dopaminergic targets or, to a greater extent than the D1-family, on dopamine neurons as presynaptic autoreceptors. This indicates a difference in the localisation of the dopamine receptors. Levey et al. (1993) used antibodies raised to specifically bind dopamine receptor subtypes and demonstrated that the D1 and D2 receptors are differentially enriched in the striatal patch and matrix compartments. Additionally, both receptor types may regulate neurotransmitter release by presenting themselves in the axonal terminals. Sesack, Aoki, and Pickel (1994) claimed consistently that the localisation of D2-family

receptors subserve auto-regulation at the level of dendritic spines in the midbrain and at the presynaptic axonal terminals in the striatum. Importantly, the autoreceptors underpin an excitability fine-tuning that governs the pattern of firing discharge of the dopaminergic neurons, which, in turn, regulates the extracellular dopamine concentration and postsynaptic reactivity (Mercuri et al., 1997).

The two receptor types not only differ in synaptic localisation, they are also shown to distribute differentially across cortical and subcortical regions, as well as across laminae. In a monkey study (see Goldman-Rakic, Lidow, Smiley, & Williams, 1992 for dopamine resemblance between human and other primates), Lidow et al. (1991) used autoradiography with a D2 antagonist ([3H]raclorpride) and concluded that low-density D2 distribution was detected for frontal, parietal, and occipital lobes, with a preferentially high concentration in cortical layer V. This result was compared with that of D1-sepcific binding, which revealed that the density for the D1 receptor is over 10 to 20-fold higher than that for the D2 receptor. Also, compared with the D2 laminar preference, the D1 receptors were observed primarily in supragranular layers and infragranular layers. Both receptors show a rostro-caudal decrease in density, suggesting a gradient of functional significance. In spite of the disproportionate receptor density, the D2 receptor may play a greater role in the human basal ganglia than in other brain regions. Camps et al. (1989) used autoradiographic techniques with the administration of radioactive D2 antagonist in human *post mortem* brain tissue. The result revealed the highest D2 densities in the caudate, putamen, olfactory tubercle, and SNc. Although the D1 receptor is still the dominant subtype, the dominance is around a tenth as compared with the D1 density in other brain regions, namely, the D1/D2 concentration ratio is at 2 - 3 in the basal ganglia, contrasting with the ratio of 10 - 20 in the neocortex (Lidow et al., 1991).

1.4.2. Dopamine modulates neuronal excitability

A bewildering aspect of dopamine lies in the fact that it is neither an excitatory nor an inhibitory neurotransmitter, unlike other neurotransmitters that work on ionotropic receptors, such as glutamate and GABA. The dopamine receptor families belong to the G protein coupled receptor class, a major role of this receptor class is by affecting a secondary messenger system, which increases or decreases intracellular level of cAMP. The net influence of dopamine is therefore dependent upon the receptor subtype with which it interacts, as well as the reaction of the postsynaptic cell to the cAMP. Generally speaking, the effect of dopamine is the regulation of excitability as a summation of multiple factors (Iversen, 2010).

Prefrontal neuronal excitability may be modulated by a postsynaptic dopamine-glutamate interaction via the D1 receptor. In a rodent study, Wang and O'Donnell (2001) reported that a synergism exists between NMDA and D1 receptor activation, which led to increased spike numbers with decreased latency. Multiple pathways were suggested to mediate the synergism by observing its removal through the administration of protein kinase A (PKA) inhibitors and Ca^{2+} chelator. The same results were extended by Tseng and O'Donnell (2004) in which the role of D2 receptor activation was characterised in light of D1-NMDA synergism. The excitatory effect of NMDA in the prefrontal cortex was attenuated by D2 agonists. The D2-induced NMDA attenuation was, however, removed by $GABA_A$ antagonists, suggesting a mediation of GABAergic interneurons. Overall, prefrontal pyramidal cell excitability is modulated by D1 and D2 receptors in opposite ways (Trantham-Davidson, Neely, Lavin, & Seamans, 2004).

In the striatum, dopamine modulation also controls intrinsic excitability and glutamatergic signalling, although the effect depends on the receptor subtype expressed on the striatal MSNs (D1 MSN and D2 MSN). A classical model (Albin, Young, & Penney, 1989) outlines one aspect of how dopamine shapes striatal activity. This model has been elaborated with new findings (Nicola, Surmeier, & Malenka, 2000; Redgrave et al., 2010). The classical view states, in principle, that the D1 receptors excite the striatonigral ('direct') pathway, whereas the D2 receptors inhibit the striataopallidal ('indirect') pathway. This means different receptor subtypes are segregated, whilst a smaller subpopulation of MSNs coexpress both subtypes (D. J. Surmeier, Song, & Yan, 1996). Through D1 receptor signalling, the MSNs may approach a more depolarised state known as an *up* state under sustained glutamatergic stimulation. Whereas, such signalling with transient or uncoordinated glutamate release may not form an up state (Nicola et al., 2000). A similar physiological consequence of D1 dopamine was also observed in the deep layer pyramidal neurons of the prefrontal cortex exhibiting bistability (Lavin & Grace, 2001). By contrast, D2 signalling exerts an opposite effect, which inhibits presynaptic release of glutamate, thereby diminishing D2 MSNs stimulation (Bamford et al., 2004).

1.4.3. Tonic and phasic modes

Dopamine neurons are known to fire in two distinct modes that affect extracellular concentration and pre-/post-synaptic receptor binding, one is characterised by a low-frequency (around 4 - 5 Hz for primates), spike-independent *tonic* mode and the other by a short-latency (70 - 100 ms), short-duration (around 200 ms), burst of neuronal activity called the *phasic* mode. The phasic mode is also referred to as spike-dependent activity, in which packets of action potentials (20 - 80

Hz) at a hundred-millisecond scale are separated by a longer electrical silence (Iversen, 2010). These phasic spiking activities are in response to salient, unexpected events that are attributed to prediction error signals (Redgrave, Gurney, Gurney, & Reynolds, 2008; Schultz & Dickinson, 2000). The terminal release of tonic dopamine in the striatum is suggested to involve prefrontal glutamatergic afferents (Grace, 1991). A primary form of this excitatory regulation takes place in the dopaminergic VTA cell bodies (Karreman & Moghaddam, 1996). The tonic and phasic modes of dopamine are one determinant of extracellular dopamine concentration. The other factor affecting the concentration concerns the cellular metabolism that governs dopamine re-uptake. Dopamine concentration is speculated to modulate higher cognitive performance, including working memory. As mentioned earlier, the anatomy of dopamine receptor varies across the rostro-caudal axis, as well as across receptor subtypes. Studies have shown that dopamine metabolism also exhibits regional diversity. This may have a pronounced effect of dopamine acting on different receptor subtypes, as the D1-family and the D2-family receptors have quite distinct dopamine-binding affinities. Taken together, the likelihood of dopamine receptor activation is a function of the dopamine affinity of receptor subtypes and the concentration to which the receptors are exposed.

Specifically, the D2 receptor has higher dopamine affinity than that of the D1 receptor (Cools & D'Esposito, 2011; Schultz, 2007). For a resting, unengaged animal, the tonic mode is able to maintain an extracellular concentration of a few nanomolar, which is sufficient to provide tonic D2 activation (Richfield, Penney, & Young, 1989). Whereas, the D1 receptor is not activated unless a higher concentration – over 100 nanomoles produced by phasic bursts – is provided (Richfield, Young, & Penney, 1987).

1.4.4. Homeostasis hypothesis

A hypothesis relevant to the interaction between the independently regulated tonic and phasic dopamine in the striatum was proposed by Grace (1991). The hypothesis states that the extent to which dopamine may express its spike-dependent influence depends on synaptic homeostasis. The homeostasis hypothesis rests on two premises. Firstly, the behaviourally relevant phasic dopamine release in the synaptic space is subject to fast, low-affinity/high-capacity re-uptake systems (Iversen, 1973), such that homeostatic responses are not triggered. Secondly, the (prefrontal) glutamate-mediated terminal release of spike-independent dopamine is at background concentrations and unaffected by the low-affinity re-uptake system. Therefore, changes in tonic dopamine release would contribute to extracellular dopamine concentration, thereby triggering homeostatic regulations via inhibitory D2 autoreceptors. On stimulating the D2 autoreceptors, the dopamine release due to fast spikes is down-regulated (but see Benoit-Marand, Borrelli, & Gonon, 2001 for spike-dependent autoregulation). In other words, the amplitude of the phasic responses is set by the cellular responsiveness shaped by the tonic dopamine release (Grace, 1991).

The hypothesis above has stimulated another supplement hypothesis along the same line of reasoning but ascribed to the prefrontal cortex with distinctive dopamine elimination routes, as compared with that of the striatum (Bilder, Volavka, Lachman, & Grace, 2004). The distinction between the prefrontal cortex and the striatum in dopamine re-uptake mechanisms is characterised by insignificant involvement of the dopamine transporter (DAT) and monoamine oxidase (MAO) in the prefrontal cortex (Lewis et al., 2001; Sesack, Hawrylak, Matus, Guido, & Levey, 1998). In the striatum, both the DAT and MAO are responsible for fast re-uptake of

phasic dopamine in the synaptic space. Instead, the catechol-O-methyltransferase (COMT) takes the principal role in the elimination of extracellular dopamine in the prefrontal cortex. However, COMT is generally found in extrasynaptic space, leaving a longer distance for the dopamine to travel before the re-uptake. This probably underlies the higher background dopamine in the prefrontal cortex than that in the striatum (Moghaddam & Bunney, 1993; Sharp, Zetterström, & Ungerstedt, 1986). As a consequence, Bilder's hypothesis predicts that the greater extent of dopamine diffusion increases the likelihood of extrasynaptic D1 receptor stimulation (Smiley, Levey, Ciliax, & Goldman-Rakic, 1994). This would in turn enable the glutamate-mediated release of tonic dopamine in the striatum, thereby reciprocally reducing the postsynaptic responsiveness to phasic dopamine. Another complementary possibility is that the background prefrontal dopamine may attenuate cellular excitability via D2 stimulation (Tseng & O'Donnell, 2004), affecting downstream glutamatergic corticostriatal projections. Taken together, elevated tonic dopamine level in the prefrontal cortex may have a net effect of reducing coherent input to the striatal D1 MSNs. These hypotheses have an important implication as they speak to the underlying *stability* of cortical activation states. Critically, both prefrontal D1 or D2 stimulations by background dopamine may contribute to the tonic enabling of the striatal indirect pathway and thus promote flexible set-switching or working memory updating (J. D. Cohen, Braver, & O'Reilly, 1996; M. J. Frank et al., 2001; Durstewitz, 2008; but see Stelzel, Fiebach, Cools, Tafazoli, & D'Esposito, 2013).

1.4.5. The Val^{158}Met polymorphism

An interesting aspect of dopamine function is the COMT Val^{158}Met polymorphism. This COMT genotype entails a methionine (Met)/valine (Val) substitution at codon 158 of the COMT gene (Lachman et al., 1996). Individuals exhibiting homozygosity for the Met allele are associated with a three- to four-fold reduction in COMT activity than that of Val homozygotes (Weinshilboum, Otterness, & Szumlanski, 1999). The Met homozygotes therefore have relatively higher baseline dopamine than the Val homozygotes in the prefrontal cortex, whilst the heterozygotes demonstrate an intermediate level of baseline dopamine. With the polymorphic phenomenon, some predictions concerning dopamine-related performance are made: that individuals with the Met-allele may exhibit superior cognitive flexibility and working memory performance than the Val variants (Bellander et al., 2014; Cools & D'Esposito, 2011). In particular, the improvement in Met allele performance is often characterised by task-related reduction in regional BOLD responses (Mier, Kirsch, & Meyer-Lindenberg, 2010). This is comparable to the rCBF reduction as a result of catecholamine agonist-induced working memory improvement (e.g., Mehta et al., 2000; dorsolateral prefrontal cortex and parietal cortex). The task-related BOLD reduction is also associated with faster reaction times without accuracy trade-off (Mattay et al., 2003; Tan et al., 2007). Tan et al. (2007) demonstrated trend speed-up for the Met allele (Table 1. of Tan et al., 2007) but a marked task-related reduction in the fronto-parietal network throughout a series of working memory updating and manipulation tasks. The task-related reduction may be a consequence of $GABA_A$-mediated inhibition via extracellular D2 stimulation. Of note, the magnitude of task-related reduction can be further emphasised by disrupting dopamine re-uptake mechanisms using amphetamine (Mattay et al.,

2003). The effect, compared with that of placebo administration, was most pronounced with the Val homozygotes with increment of updating difficulty (n-back). Critically, the use of drug resulted in the magnitude of regional responses in the Val group closely resembled that of the Met group. By contrast, the Met homozygotes, albeit with trend task-related BOLD reduction, showed a marked BOLD increase and impaired performance with amphetamine, indicating over-abundant dopamine under the 'inverted-U' model (Cools & Robbins, 2004).

1.4.6. Dopamine dose-dependency, terminal synthesis, and working memory capacity

An 'inverted-U' relationship seems to hold between cognitive performance and baseline dopamine level. This dose-performance model predicts that excessive dopamine is as detrimental as insufficient dopamine. Williams and Goldman-Rakic (1995) determined that a dose-dependency exists for D1 receptor function, in which monkeys performing an ODR (oculomotor delayed-response) task showed that prefrontal neurons ('memory field') encoding target location – throughout the delay period – were selectively enhanced by a low, but not high, dose of D1 antagonist. A follow-up study by Vijayraghavan (2007) showed that the low-level D1 receptor agonism enhanced tuning in the memory field by suppressing neuronal responses in the non-target fields.

The results were attributed to endogenous dopaminergic tone, where the D1 blockade unmasked cells with excessive tone (G. V. Williams & Goldman-Rakic, 1995), whereas adequate D1 stimulation suppressed noisy, spontaneous neuronal activity (Vijayraghavan et al., 2007). Zahrt (1997) also demonstrated a similar D1 dose-dependency in rodents using D1 receptor agonists. Intriguingly, dose-

dependency of performance changes with the specific task being performed (Phillips, Ahn, & Floresco, 2004). More specifically, the task difficulty contributes to how effective the receptor manipulation would be. In a rodent study based on a delayed version of radial maze task, the task difficulty was determined by the length of delay before the memory test, Floresco and Phillips (2001) demonstrated that the prefrontal administration of D1 agonists significantly improved the proportion of correct memory retrieval in the extended-delay group. The performance improved with the increase of D1 agonist dose. By contrast, the group taking the easy task exhibited a dose-dependent increase in error rate. Under the inverted-U model, the findings suggest a right shift in the dose-performance function for the difficult task, in which the drug administration corresponds to the rising segment of the difficult curve but to the descending segment of the easy curve.

Apart from the dose-dependency and the dose/task interaction on performance, the inverted-U model has an intrinsic level of variability: the drug efficacy seems to vary across individuals. Granon et al. (2000) referred this variability to the dependence on individual baseline performance. That is, undrugged poor performers (rodents) received more behavioural enhancement under the influence of D1 agonist than the good performers. On the contrary, the administration of D1 antagonist impaired the good performers but not the poor performers. It is conceivable that the good performers may have optimal level of baseline dopamine for the specific task, such that they would show little or even adverse effects under D1 stimulation. Instead, the poor performers may gain from the D1 stimulation because their basal dopamine level may be at the far left of the inverted-U curve. The opposite case was also found for D1 antagonism, in which only the good performers suffered from the drug effect (Granon et al., 2000).

Similar findings were also obtained for humans, with the individual initial conditions being determined by the measure of working memory capacity (Kimberg, D'Esposito, & Farah, 1997). Kimberg et al. (1997) manipulated the placebo/D2 agonism (via bromocriptine) effect on participants performing a Wisconsin Card Sorting task (WCST) and demonstrated that the drug eliminated the differences in performance between high-span and low-span individuals, in which the high-span individuals had better performance under the placebo treatment. In other words, the D2 agonism enhanced performance of individuals with low span size, whilst impaired that with high span size. A corroborating finding also reported beneficial effects of methylphenidate (that blocks dopamine and norepinephrine transporters) administration in subjects with lower working memory capacity (Mehta et al., 2000).

Given the dose/performance dependency changes as an inverted-U function, with reference to the initial condition, it is therefore possible to associate individual working memory capacity with basal dopamine level. Cools et al. (Cools, Gibbs, Miyakawa, Jagust, & D'Esposito, 2008) performed their study in this regard. Basal dopamine level was determined by the terminal synthesis in the striatum using PET imaging. The authors used radioactive tracers that track decarboxylase activity as an index of presynaptic dopamine synthesis capacity. They detected a positive, age-corrected, correlation between left caudate synthesis capacity and working memory capacity; trend correlations were also found in the rest of the striatum. However, current evidence is limited to the association between listening span and dopamine synthesis and cannot be extended to other span tests yet (Cools et al., 2008).

1.4.7. Computational theoretical models

Theoretical models incorporating known biophysical properties of neuronal systems often allow one to gain insights into the mechanistic principles of the system in question. Indeed, numerous models have been attempted and refined to understand the perplexing behaviour of dopamine in modulating working memory function (Dreher & Burnod, 2002; Durstewitz & Seamans, 2008; Durstewitz, Kelc, & Güntürkün, 1999; Durstewitz, Seamans, & Sejnowski, 2000). Next, we turn to these models that concern working memory process, prefrontal cortex, and dopamine modulation.

A working hypothesis of the prefrontal cortex function – with regard to working memory – is that the prefrontal cortex serves to protect goal-related information against interference during the delay period. This is associated with the control of representational stability and is considered as one of the hallmarks of the prefrontal cortex (Miller & Cohen, 2001). In terms of theoretical modelling, the mechanisms underlying prefrontal representational stability are characterised by several, though not mutually exclusive, computational principles (Durstewitz et al., 2000): (1) recurrent excitation within cell assemblies; (2) asymmetrical feedforward/feedback connectivity constituting a 'synfire chain'; (3) maintenance of membrane conductance through cellular bistability; and (4) discrete and continuous attractor states.

In Durstewitz et al. (1999), the functional role of dopamine was conceived in two aspects: (1) stimulating firing of the delay-sensitive prefrontal neurons; (2) preempting presynaptic inputs that encode goal-irrelevant information. The model did not, however, make explicit distinctions about the contribution of receptor subtypes or the temporal dynamics of dopamine discharge. Instead, model

parameters were designed to reflect synaptic or voltage-gated membrane conductance in prefrontal pyramidal neurons. Specifically, the underlying mechanisms related to (D1) dopamine-NMDA synergism that enhances a persistent shift in inward Na^+ currents, which may increase evoked excitability of pyramidal cells, thereby increasing lateral inhibition (Goldman-Rakic, 1995). Additionally, spontaneous afferent glutamatergic stimulation was prevented by modulating high-threshold Ca^{2+} and slowly inactivating K^+ currents. These spontaneous afferents, presumably goal-irrelevant, elicit distal excitatory postsynaptic potentials (EPSPs) that are mediated by dendritic Ca^{2+} currents. Dopamine may attenuate distal EPSPs, increasing influence of proximal EPSPs (Yang & Seamans, 1996). In other words, dopamine induced the effect that neurons received more current influx proximally than distally. Finally, GABAergic interneuron activity is enhanced, probably via D2 receptors (Trantham-Davidson et al., 2004; Tseng & O'Donnell, 2004), manner. Taken together, the model by Durstewitz et al. (Durstewitz et al., 1999) showed that dopamine may have a positive influence in stabilising prefrontal neural representations. This is achieved via several plausible dopamine modulations: (1) the increase of Na^+ traffic; (2) the reduction of K+ efflux; (3) the decoupling between distal and proximal pyramidal neurons; and (4) the reduction of dendritic Ca2+ conductance. The model also predicted the disruptive effect under supranormal dopamine levels, which showed appropriate neuronal excitations and strong inhibitory feedbacks.

Durstewitz and Seamans (2008) provided a more comprehensive review on models regarding to the functional implications of dopamine in the prefrontal cortex by taking the receptor-specific contribution into account. Using attractor network models, they proposed the existence of two discrete dynamic regimes, one associated

with D1 dominance and the other with D2 dominance. The D1-dominated state is characterised by a steep energy landscape, in which working memory-related attractor states staying in an energy 'well' are robustly maintained, whereas the D2-dominated state corresponds to a flat landscape, which may be beneficial for flexible switching amongst representational states. A previously proposed model (Dreher & Burnod, 2002) also appealed to the notion that the dopamine thresholding of prefrontal afferents is state-dependent.

1.5. Prefrontal cortex-basal ganglia working memory (PBWM)

The classical view of basal ganglia function is that it enables motor control and action selection through extensive connections with behaviour effector systems (Mink, 1996). With the advance of understanding the parallel organisation of the basal ganglia, as well as the connectional anatomy, it is now conceivable that the basal ganglia subserve not only motor functions but also complex cognitions (Alexander, DeLong, & Strick, 1986; Bar-Gad & Bergman, 2001; Desrochers & Badre, 2012; Draganski et al., 2008; Redgrave et al., 2010). Amongst the advances, the link between the basal ganglia, the prefrontal cortex, and working memory function is reviewed here. Especially, an influential computational model proposed by Frank, Hazy, O'Reilly, and colleagues (M. J. Frank et al., 2001; Hazy, Frank, & O'Reilly, 2007; R. C. O'Reilly, 2006; R. C. O'Reilly & Frank, 2006) is brought to focus here – in light of its implication in working memory updating. The model gives an account of the control of information access into working memory (McNab & Klingberg, 2008) based on theories of artificial neural networks, the 'long short-term

memory' (Hochreiter & Schmidhuber, 1997), and a biologically realistic temporal difference (TD, or actor-critic) architecture.

1.5.1. The neuroanatomy of the basal ganglia

The basal ganglia are comprised of the striatum (caudate and putamen), the nucleus accumbens, the subthalamic nucleus (STN), and the globus pallidus, which includes the internal (GPi) and external (GPe) segments. The basal ganglia receive (glutamatergic) inputs from virtually all areas of the neocortex, with a specific laminar origin of layer V, and in some cases layer III. The basal ganglia send outputs via the GPi and the substantia nigra pars reticular (SNr) that target thalamic nuclei, which eventually reach cortical layer IV. The intralaminar thalamic nuclei also project back to the striatum. The corticostriatal projections are unique as far as a single MSN is concerned (Zheng & Wilson, 2002). This is because for the dendritic field of a single MSN, some 2800 MSNs are also present, whereas a single corticostriatal axon traversing this field has on average 40 boutons and makes one or a few contacts with a single MSN. This makes finding two striatal neurons that share a common cortical input rather unlikely. Therefore, the activation of a MSN depends on multiple convergent cortical neurons – a distinct feature that implicates the foundation of information processing in the basal ganglia (R. L. Cowan & Wilson, 1994).

The targets of the corticostriatal afferents are the striatal medium spiny neurons, a type of GABAergic projection neurons that accounts for 95% of the striatal neurons. Two critical basal ganglia circuits are provided by the striatal medium spiny neurons, the 'direct' and 'indirect' pathways. The direct pathway is named for its

direct input to the output nuclei in the GPi and the SNr. Whereas, the indirect pathway follows two waypoints sequentially, the GPe and the STN, before the output nuclei. The output neurons in the GPi are GABAergic and exhibit a relatively high level of tonic activity; in other words, the GPi tonically inhibits their thalamic targets. These inhibitory effects may be removed by the excitatory glutamatergic inputs from the neocortex via the direct pathway. As a consequence, activation of the striatal MSNs will inhibit the output neurons of GPi/SNr, thereby *disinhibiting* the thalamic targets. The indirect pathway is, however, more complex: the striatal output neurons first target the GABAergic GPe neurons, thereby disinhibiting the STN output, allowing the glutamatergic neurons of the STN to activate the GPi/SNr neurons, and finally result in enhanced inhibition in the thalamic neurons. Although the exact mechanisms regulating the activation of the direct and indirect pathways are complex and require extended text to explain, their concerted role provides antagonistic/counterbalanced regulation of basal ganglia output.

Unlike the neocortex, the striatum lacks distinct cytoarchitectonic organisation, which means laminar structures are not present. However, it is well established that the striatum, along with the rest of the nuclei, maintain a topographic relationship with the neocortex (Draganski et al., 2008). For example, projections from the prefrontal cortex converge in the rostral part of the caudate nucleus, putamen, and globus pallidus. Likewise, the caudal part of the caudate, putamen, and globus pallidus receive inputs from the motor cortex.

The axons of striatal medium spiny neurons exhibit an asymmetric degree of recurrent collaterals with respect to the dopamine receptor subtypes they express. As noted by Taverna et al. (2008), unidirectional MSN-to-MSN synapses are formed between D1 receptor-expressing MSNs, as well as between D2 receptor-expressing

MSNs. D2 receptor-expressing MSNs also form synapses with D1 receptor-expressing MSNs, but the reverse case is scarce. In other words, direct pathway MSNs tend to innervate MSNs of the same pathway, whereas indirect pathway MSNs innervate both MSN types equally. Additionally, this receptor-dependent MSN coupling seems to be disrupted in Parkinson's disease models, suggesting a functional role in behavioural switching or possibly cognitive flexibility, although the exact mechanism remains unknown (Kreitzer, 2008). More recently, Lalchandani et al. (2013) suggested, in an *in vitro* study, that the efficacy of MSN collaterals may be regulated by dopamine, in which D2 agonist administration resulted in increased synaptic $GABA_A$ clusters and $GABA_A$ release sites that led to a greater synaptic efficacy. Further studies are still needed to understand how collateral inhibition may enable integration of information in the striatum.

1.5.2. The 1-2-AX continuous performance task (1-2-AX CPT)

The 1-2-AX CPT (Figure 1.2) was devised by Braver and Cohen (Braver & Cohen, 2000) based on a simpler version (Barch et al., 1997) and is used as a model task to demonstrate the behaviour of the PBWM introduced in the ensuing section. The task involves the presentation of a fixed set of stimuli (1, 2, A, X, and Y). X is the target to which the subject should respond if it follows an A, and the most recent number seen is 1. Alternatively, Y may be the target if the preceding stimulus is B, and the most recent number observed is 2. The task therefore entails both subgoals and goals: the subgoals are defined by the number stimuli that induce, at task level, the maintenance of the contingency within which the goal – A-X or B-Y sequence – is dealt. A hierarchy of functional demands in working memory is instantiated here,

including the encoding of relevant stimuli, the active maintenance of task and stimulus information in the presence of distractors, and the contingent updating of the A-X/B-Y sequence.

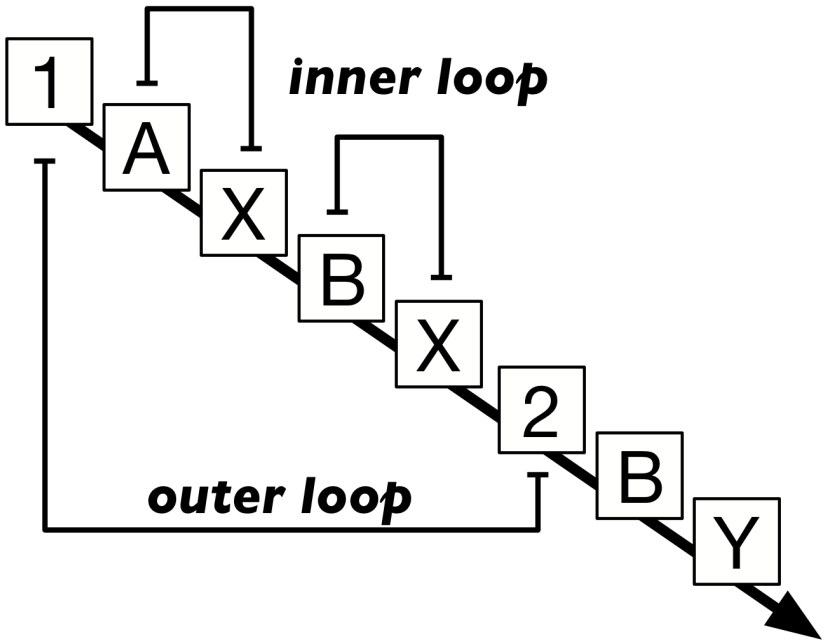

Figure 1.2 The 1-2-AX continuous performance task (1-2-AX CPT). This widely applied cognitive task encapsulates two types of stimuli that require working memory to be engaged in a hierarchical manner. The task performance is context-sensitive, with the context being induced by the number stimuli (1 or 2). Unless a different number stimulus is encountered, the context is maintained and the context-dependent cue-response pairing is exercised. For example, a subject may only respond to the 'X' stimulus immediately following an 'A' if the current context is '1'. Alternatively, the target to which one makes responses is 'Y' when it is following a 'B' under the '2' context.

1.5.3. Computational model

The PBWM model (OReilly:2006gy ; also see M. J. Frank et al., 2001) employs a series of mechanistically plausible considerations (R. C. O'Reilly, 1998) concerning the interactions between the prefrontal cortex, the basal ganglia, and the dopaminergic midbrain which together enable working memory. It is envisaged that

the model must be able to learn what information to keep over time and implement contingencies with respect to the delayed cue-outcome relationship. The model assumes working memory representations are maintained in prefrontal cortical activity, whilst the basal ganglia subserve a dynamic gating mechanism via disinhibitory neural pathways that determine what information can be represented by the prefrontal cortex. The gating mechanism is made adaptive by means of reinforcement learning, which reflects potent dopaminergic neuromodulation in the basal ganglia. Overall, the model enables computations associated with three key functional demands of working memory mentioned earlier: rapid updating, robust maintenance, and selective updating.

The notion of gating by disinhibition is straightforward and makes direct link with the functional anatomy of the basal ganglia described previously (see also Figure 1.3). Firstly, the prerequisite to enable any prefrontal representation is to follow the thalamic disinhibition by activating the direct ('Go') pathway, thereby toggling the prefrontal cellular bistability (Camperi & Wang, 1998). Following this, active maintenance of information is then achieved by excitatory feedback circuit and recurrent excitatory inhibition that are intrinsic to the prefrontal cortex (Goldman-Rakic, 1995; 1996). At this stage, the indirect ('NoGo') pathway of the basal ganglia is enabled – to prevent erroneous encoding/updating in the maintained representations. Finally, selective updating (the case in which 1 is maintained but A or X may be updated, for example) proceeds in light of the connectional parallelism of the basal ganglia anatomy and the corticostriatal sparsity (Alexander et al., 1986; Zheng & Wilson, 2002). O'Reilly and colleagues viewed such structures as 'stripes' that can be modelled as a single unit of the prefrontal/basal ganglia system. The

remaining issue with the PBWM model pertains to the question of how the basal ganglia learn when to fire a Go or a NoGo signal.

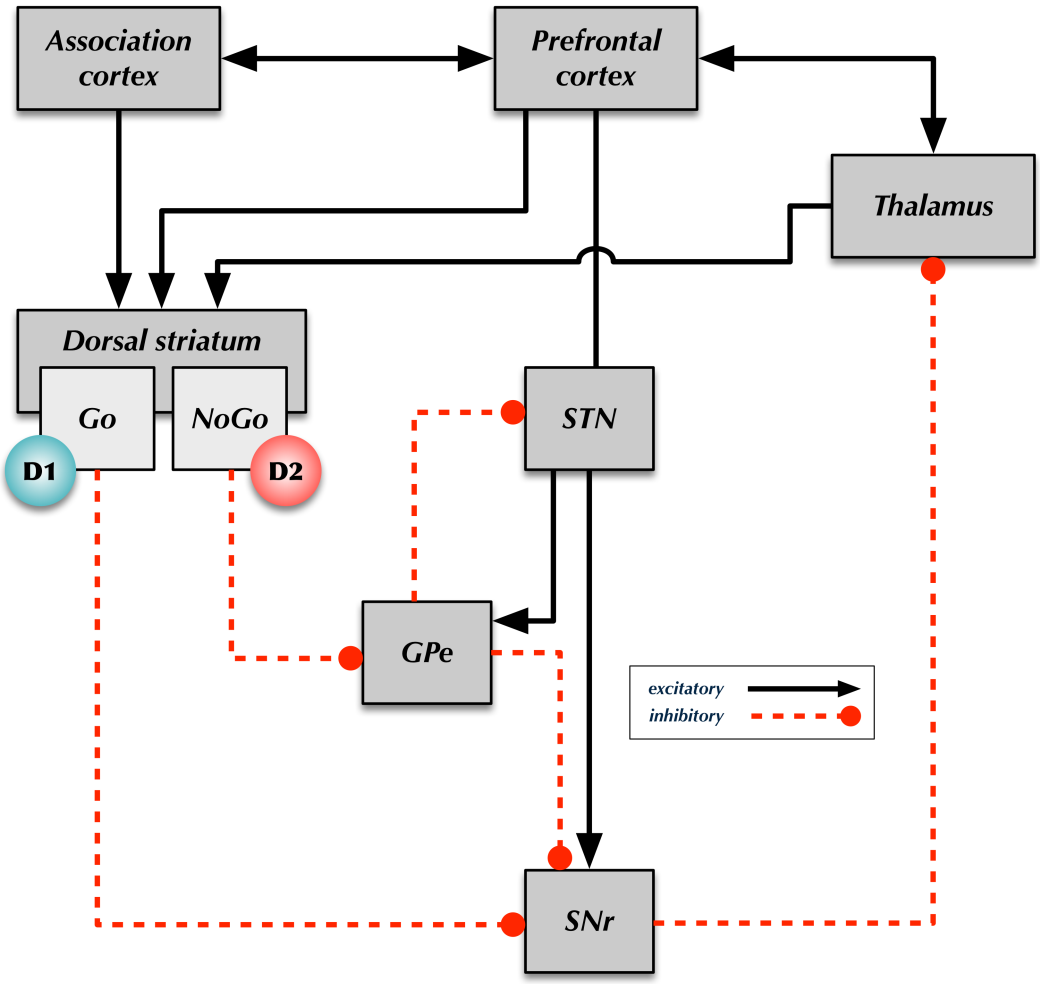

Figure 1.3 Parallel loops in the cortico-striato-thalamo-cortical pathway.

1.5.4. Empirical support

The PBWM model is especially successful in predicting the involvement of the dopaminergic midbrain while working memory is being updated (D'Ardenne et al., 2012; Marklund et al., 2009; Murty et al., 2011). The gating notion is also supported by studies in which the basal ganglia is implicated when distracting information was

present and had to be distinguished from task-relevant information to guarantee task performance (Baier et al., 2010; McNab & Klingberg, 2008); or when information from one domain-specific sensory channel is selected over the other in a context-sensitive manner (van Schouwenburg, Ouden, & Cools, 2010).

Murty et al. (2011) compared neural responses during distinct working memory demands: maintenance, overwriting (total updating), and selective updating. They demonstrated that the dopaminergic midbrain, SN/VTA, and the bilateral caudate were significantly recruited. Correlation analyses also showed a functional connectivity between the striatum and the midbrain, but not with the prefrontal cortex. Interestingly, during the overwriting phase, no midbrain activation was detected, whilst the striatum was reliably deactivated, suggesting a tonic enabling of the 'NoGo' pathway (Hazy et al., 2006).

D'Ardenne et al. (2012) employed the A-X-CPT and demonstrated that the right dorsolateral prefrontal cortex was activated while subjects underwent context updating, i.e., the presentation of A in the A-X pair. Furthermore, they delivered single-pulse TMS to the context-sensitive region stated and found a context-dependent, time-dependent disruption in context updating, suggesting a critical role of the right dorsolateral prefrontal cortex in context encoding. The time at which the TMS delivery is effective was at 150 ms after the contextual cue was presented. Additionally, the context updating was associated with dopamine-related phasic BOLD responses in the SN/VTA (D'Ardenne et al., 2008). However, the authors have yet to demonstrate whether the dopaminergic signal is associated with the basal ganglia in implementing the gating mechanism, altogether the signal is significantly correlated with the right dorsolateral prefrontal responses in a context-dependent manner.

In unmedicated Parkinson's disease patients, Marklund et al. (2009) used fMRI to isolated transient from sustained brain activity during an N-back working memory updating task (Donaldson, 2004). They found a consistent under-recruitment in the caudate nuclei of patients relative to the normal controls. In the control group, updating was associated with a transient BOLD profile that can be specifically related to the phasic gating signal underlying working memory updating (Marklund et al., 2009).

An analysis based on effective connectivity and Bayesian model comparisons also revealed that the basal ganglia served as a 'gain control' between the prefrontal cortex and domain-specific association cortices. van Schouwenburg et al. (2010) used non-linear dynamic causal models (Stephan et al., 2008) and showed that the basal ganglia mediate cognitive flexibility, i.e., set-switching, in response to behaviour-relevant changes in the environment. Specifically, it is the top-down connectivity that was gated by the basal ganglia activity, suggesting a downstream attentional control in the association cortex.

Recently, an fMRI study showed that the activity in the basal ganglia not during working memory processing but during the instruction phase – in which the subjects were cued to ignore certain stimuli whilst scanning through them – predicts individual working memory capacity, as well as the task performance (McNab & Klingberg, 2008). This activity is referred to as 'filtering set' (McNab & Klingberg, 2008). A crucial implication from the observation of filtering set activity is that the gating function of the basal ganglia is selective and may be implemented in an anticipatory or preparatory sense. The study also served as to establish that there is a convergence between perceptual set and working memory mediated by the basal ganglia. Indeed, the parietal activity associated with non-specific sensory storage

(Vogel et al., 2005) decreases with the amplitude of the basal ganglia (globus pallidus) filtering set activity (McNab & Klingberg, 2008).

The notion that the basal ganglia gate or 'filter' information entering working is further supported by a lesion study (Baier et al., 2010). In an experimental setting similar to that of McNab et al. (2008), Baier et al. (2010) showed that lesions of the (left) putamen specifically caused the subjects to perform unreliably when distractors were presented together with the target stimulus. The result was confirmed by relating behavioural variables (e.g., the filtering ability as measured by differences in accuracy between distraction and distraction-free conditions) to the anatomical location of the lesions using VLBM (voxel-wise lesion-behaviour brain mapping; Rorden, Karnath, & Bonilha, 2007). Additionally, working memory performance was impaired in a load-dependent manner in patients with prefrontal cortex lesions, suggesting a key role for the prefrontal cortex in actively maintaining goal-relevant representations. Of note, the VLBM associated variations in working memory capacity with lesions in the insular cortex.

1.5.5. Limitations

Although the PBWM model gives a formal, mechanistic account – of how the meso-prefronto-basal ganglia circuitry achieves adaptive representation of working memory – it is limited in terms of biological realism, with respect to the ability to capture neuronal activity generated by continuous-time dynamical systems. Such systems can be expressed at the level of neurons (e.g., Durstewitz & Seamans, 2008) or as an ensemble (e.g., Friston et al., 2012). It also lacks a probabilistic representation, whereby environmental states are represented by density functions (Friston & Friston, 2005; e.g., Koechlin & Summerfield, 2007), as opposed to a

'slot' device. In addition, it cannot distinguish between stimulus-bound and contextual representations, which speak respectively to quite distinct neurophysiological mechanisms: one relates to information encoded in synaptic activity, and the other to modulation of plasticity or synaptic gain.

1.6. The predictive brain

1.6.1. The predictive coding hypothesis

Rao and Ballard (Rao & Ballard, 1999) proposed the hierarchical predictive coding model to address 'extra-classical' receptive field phenomenon in visual cortex. It has been known for decades that for individual neurons in the primary visual cortex (in layer II and III) there is an optimally configured stimulus (e.g., the orientation of a line segment) that, when presented in the neuron's (classical) receptive field, elicits the most rigorous response (Hubel & Wiesel, 1968). The extra-classical phenomenon means that if the optimal stimulus extends beyond the neuron's receptive field, the neuronal response is suppressed. This is referred to as 'endstopping' and holds for the case that the 'classical' neuron is suppressed when the surrounding extra-classical receptive field is exposed to the stimulus with a specific property that matches the centre. The authors made a remarkable prediction that the extra-classical neurons provide predictive codes to the classical neurons, and together achieve three fundamental aspects of information processing in a neural network: (1) to encode exogenous statistical regularities, (2) to only signal deviations, and (3) to reduce redundancy (Rao & Ballard, 1999).

It naturally follows that the neural network may have a hierarchical organisation, whereby neurons of higher level may have a more general sense of the world and if such sense is sufficient to 'explain away' lower sensations, then there is of no biological value to process this information. Additionally, the idea about signalling only the deviations – those that cannot be predicted, i.e., prediction errors – is potent in the sense that such signals carry information that is not already predicted and may be of biological importance.

It is also conceivable that the size of the receptive field increases as the hierarchy progresses, such that neurons at a higher level may have a receptive field of the entire visual field, or even be able to encode a 'template' of some aspects of the physical world. But what do these templates reflect in the real world? The answer to this is rooted in Helmholtzian notions that underlie perceptual inference and perceptual learning (Friston & Friston, 2005). Briefly, the process of inferring the *cause* of sensations is perceptual inference, whilst the process of capturing the interdependency between causes and sensations is perceptual learning. The 'template' is therefore a hypothesis – a 'diagram'– about how sensations are generated. This, in part, necessitates the hierarchical organisation of the neural systems (Markov & Kennedy, 2013; Mumford, 1992). The reason for this is simple: if the sensory infrastructure can recapitulate the causal structure of the environment, it suggests hierarchical structures in its environment (Dayan, Hinton, Neal, & Zemel, 1995; Friston, 2005).

1.6.2. Optimising precision and attention

Predictive coding mentioned above partly captures the modern treatment of perception, that is, in terms of hypothesis testing, the sensory signal is tested as to whether it is sampled from a distribution known *a priori*. This is equivalent to inferring the state of the world using generative models that represent hidden causes of the state. One can easily see how perception as hypothesis testing resembles statistical analyses in most scientific disciplines. The simplest example is perhaps the Student's *t*-test, in which a group difference is detected by dividing the difference in group means with the standard error, under the null distribution. Two quantities are estimated here: the observed difference as the *prediction error* and its standard error or *precision* (i.e., inverse variance). In the predictive coding framework, the information being carried forward can then be analogously regarded as precision-weighted prediction error.

But what exactly, in psychological and neurobiological terms, is precision and the process involved in estimating precision? Feldman and Friston (2010) argued that the precision of sensory signals is inherent to the environmental states that are to be inferred. This means perceptual inference – i.e., optimising inferred environmental states – entails the process of optimising the precision or uncertainty of the states – which corresponds to attention. In other words, attention is an emergent property as our brain makes hierarchical inferences. Inferring the uncertainty of the state of the world is thus an integral part of predictive processing in the brain.

The assumption that neural systems have a hierarchical architecture is important because it is formally equivalent to empirical Bayes models (Friston, 2009). This means top-down effects serve as empirical priors that constrain behaviours of the lower levels, as well as the bottom-up effects. In a very broad sense, if an organism

employing hierarchical inference abides by the Bayesian principle, the only two sorts of things that concern the organism (or its brain) are the state of the world and the uncertainty about the state. This allows a very parsimonious characterisation of a behaving organism in relation to the environment upon which it acts (e.g., Friston et al., 2012).

1.6.3. Cortical message passing

In reality, the primate brain is hierarchically organised. This observation is based on the inter-laminar connectivity across cortical macrocolumns. The lamination and columnar structure appear to be quite similar across sensory and association cortices. One model (Mumford, 1992) assumes that brain activity emerges from the convergence of feedforward and feedback information processing, and re-propagation.

Under this model, the feedforward and the feedback pathways do not interact until they meet at a common processing unit. Mumford (1992) proposed three pathways that provide the up/down streams with topographical segregation: (1) ascending pathway; (2) standard descending pathway; and (3) extra-descending pathway. The ascending pathway entails supragranular pyramidal neurons of lower level projecting to the higher cortical layer IV. The standard descending pathway refers to pyramidal axons of layer V in a higher level terminating in Layer I (containing apical dendrites of layer II/III) and VI of lower level. At a higher level, extra-descending projections from supragranular layers (II/III) may also terminate in lower level layer I and VI.

As a general rule accepted by many, including Rao (1999) and Friston (2005), lower superficial layer neurons terminate in higher layer IV, whilst higher deep layer

neurons terminate in lower layer I/VI. Additionally, it is generally accepted that the feedforward connection has a role as driving inputs, whereas the feedback connection as both driving and modulatory inputs. However, as pointed out in Markov (2013), laminar projections not conforming to the above model are reported: in addition to sending feedforward projections, layer III neurons also project backward; also, deep layer (V) neurons send forward projections. Although the proximity between macrocolumns may contribute to the heterogeneity, it limits the current generative models in terms of their generalisability. Nevertheless, on the level of population dynamics and the granularity of underlying neural architecture concerned, Mumford's proposal seems to be an adequate description (see Bastos et al., 2013).

1.6.4. Generalised predictive coding

Predictive codes are 'predictive' in the sense that the *current* sensations are being predicted. However, predictions can also be implemented in an anticipatory sense. For example, directing attention by cueing the target location before the target onset enhances perceptual decision (Feldman & Friston, 2010; Posner, 1980). Forecasting stimulus category or identity with predictive cues also modulate sustained activity of working memory circuits (Bollinger, Rubens, Zanto, & Gazzaley, 2010) or regional connectivity (Rahnev et al., 2011). These studies show that the pre-stimulus deployment of attention and working memory have a role in cross-temporal integration of perception and can thus be regarded as top-down modulation (Gazzaley & Nobre, 2012). In fact, working memory and attention share considerable neural substrates (Mayer et al., 2007) and are sometimes complementary psychological constructs (Baddeley, 2012; 2007). It can be argued that working memory is attention optimised not for exogenous percepts but for

endogenous instantiations of likely percepts (cf. N. Cowan, 2008; McElree, 2001; Oberauer, 2002) that one may bring to bear in the near future. In this sense, working memory and attention is compatible with a more generalised form of the predictive coding model, i.e., generalised predictive coding. The key proposition of the generalised predictive coding framework (Friston, 2008; Friston, Mattout, & Kilner, 2011) is that the hidden causes and states of the world are represented in terms of their generalised motion. This means that the generalised states prescribed by neuronal populations traverse through the state-space along the trajectory that encodes future states. The traversal may visit variables that are responsible for generating sensory data in a transient, metastable (Bick & Rabinovich, 2009; Friston et al., 2012) or a relatively stable manner (Amit, Fusi, & Yakovlev, 1997). In other words, if the brain's generative model includes trajectories or future (fictive) states, then working memory becomes a necessary part of predictive coding. In this context attention corresponds to the optimisation of precision or confidence in future outcomes based on recent experience.

1.7. Chapter Outline

Chapter Two gives a detailed description on the design of the experiments employed to address our research questions, the participants, the imaging procedures, and the methods. The main findings are concerned with regards to the principal experiment – the working memory updating task. A subsidiary task for determining individual working memory capacity is also documented; the task was conducted as a control for confounding factors that may interact with updating performance. Another key aspect in this chapter is related to the methodology. We outline the

methods that address our research questions. We focus on methods that play a key role through Chapter 3 to 5, including the general linear model (GLM), multivariate pattern analysis (MVPA), based on support vector machines (SVMs), and dynamic causal modelling (DCM). Concise theoretical backgrounds are provided. Detailed, topic-specific treatments are, however, described in respective chapters.

Chapter Three is the first chapter reporting empirical findings. Specifically, we show that behavioural performance was modulated by the set-outcome relationship, indicating the influences of valid versus invalid sets, which, according to our interpretation, speak to predictive nature underlying the neural implementation of the anticipatory set. The use of mixed temporal profiles enabled GLM to isolated set-related sustained activation as hypothesised. We report differential set-related activation in the striatum and the SN/VTA. These results are discussed in light of dopaminergic neuromodulation. The prediction was made that anticipatory set may implicate the release and maintenance of tonic dopamine, given its role in nuancing attractor dynamics. A proposed set-dependent 'inverted-U' function of dose and performance (based on Goldman-Rakic, Cools and others) was used to explain the set-performance correlations. This chapter concluded that the task induced a second-order 'set' in which a non-specific perceptual set is embodied.

Chapter Four is the second empirical chapter. It serves as a follow-up analysis for the issues discussed in the previous chapter. Namely, the lack of power and sensitivity of the GLM to detect surprise-related activation. The major cause was ascribed to the small number of trials. This is because irregularity and rarity are components inherent to the prediction-surprise paradigm. In addition, between-subject variability in surprise-related responses may impair the efficiency of the GLM analysis due to the fact that distributed network may react to the surprise, as

compared with more localised, stimulus-bound surprise in perceptual decisions. The solution was to make use of the variance-covariance structure of the surprise-related response, which underpin MVPA. The MVPA approach is sensitive to the *pattern* of response of all voxels considered, as opposed to interrogating the amplitude of voxel-wise response. We proposed two types of surprise response were induced by our task, omission and deviation, and hypothesised that the two surprise types corresponded to differentiable patterns of (informative) voxel extent. We documented that the omission pattern encoded more information in the fronto-parietal and cingulo-operculum network. Whereas, the deviation pattern involved the visual cortices, cerebellum, and the midbrain. Classifier weights and voxel counts were used to quantify to degree of informativeness within these regions-of-interest. Functional implications with respect to each pattern are discussed. This chapter concludes that two levels of predictions may be involved in the task, suggesting that the working memory function is subserved by an adaptive stimulus control and a set-maintenance control mechanisms.

Chapter Five proposes an integrative perspective based on the findings in the previous two chapters. Firstly, it considers an ostensible discrepancy in the main findings between Chapter 3 and 4. In Chapter 3, we reported that the sustained set activation recruited more posterior (occipito-parietal) regions, where anticipatory set is suggested to involve predictive processes. By contrast, according to Chapter 4, the omission pattern, which involved prefrontal and parietal regions, may also reflect prediction signals. As both reflect prediction signals, how can they be expressed in separate systems? One useful notion is from the model of cortical message passing under the predictive coding framework. This scheme suggests that prediction signals provide top-down backward connections, whilst the prediction error signals provide

bottom-up forward connections. The other refers to the common understanding that the BOLD fMRI signal is more sensitive to presynaptic, modulatory inputs. We therefore hypothesised that the set is a top-down influence modulating afferent connections to regions reported in Chapter 3. The second question then pertained to the effect of omission-related surprise. We related surprise to prediction error signals, which entail forward message passing. Finally, we hypothesised that surprise could also reflect adaptive modulations which act on the local recurrent connections within the network. We motivated a model space according to the above hypotheses and tested them within the DCM framework. This chapter draws conclusions based on Bayesian model selection and family-level inference, which provided strong evidence that working memory processing follows a cortical message-passing scheme. Supporting evidence also indicates that two antagonising cortico-striatal and cortico-cortical pathways serve to nuance representational flexibility and stability in working memory.

Chapter Six presents a synthesis and general discussion of the empirical chapters from 3 to 5, and remarks on the general findings throughout this thesis. It also states the contributions to systems neuroscience with regards to working memory and higher cognition. Next, I outline future investigations for advancing our understanding of working memory as an integrative part of the predictive brain. Limitations of the current work are critically evaluated such that means of refinement may be brought to bear in the future.

Chapter 2. Materials and Methods

2.1. Participants

Seventeen subjects (eight females; mean ± SD age, 28.0 ± 4.4 years; range, 21–36 years) were recruited via the University College London Psychology Subject Pool. Subjects were screened for right-handedness, unimpaired or correct-to-normal visual acuity, and normal colour vision. All subjects reported no history of psychiatric or neurological illness. English as primary (n = 12) or secondary (n = 5) language was required. Four additional subjects were recruited during the pilot study, in which only the behavioural task was involved and was not included in the analysis reported in this thesis. All subjects were reimbursed monetarily for their time after the study, the reimbursement was part of Wellcome Trust funding. This study was approved by the Institute of Neurology (University College London) Ethics Committee. All subjects provided informed consent before the study.

2.2. Experimental design

2.2.1. Working memory updating task

We proposed a working memory updating task, which was a modification of the delayed match-to-sample paradigm. Two additional components were added to allow testing of relevant hypotheses. Each trial involved five phases: *cueing*, encoding, retention, *action*, and probing. Here, the term 'action' refers to the designated mnemonic processing (updating and maintenance), not any motor or reflexive

command. Subjects were required to match the probe to the content of their working memory and to respond with a binary choice, indicating either a match or mismatch. The subject's working memory content was trial-specific and entailed a serial composition of the two memory arrays, given at the encoding and action phases.

The stimuli comprised predictive cues and memory arrays (Figure 2.1a). The cues reported the likelihood of an update in the ensuing action stage; a high (80% chance of updating) or low (20% chance of updating) probability cue may take place, these cues were displayed as green, upward arrows and red, downward arrows, respectively. We refer to the high cue as the updating cue (UC), and the low cue as the maintenance cue (MC). Note that the need to update was only explicit upon the presentation of the action array, although the predictive cue could establish an appropriate cognitive (anticipatory) set, depending on whether updating or maintenance was *a priori* more likely.

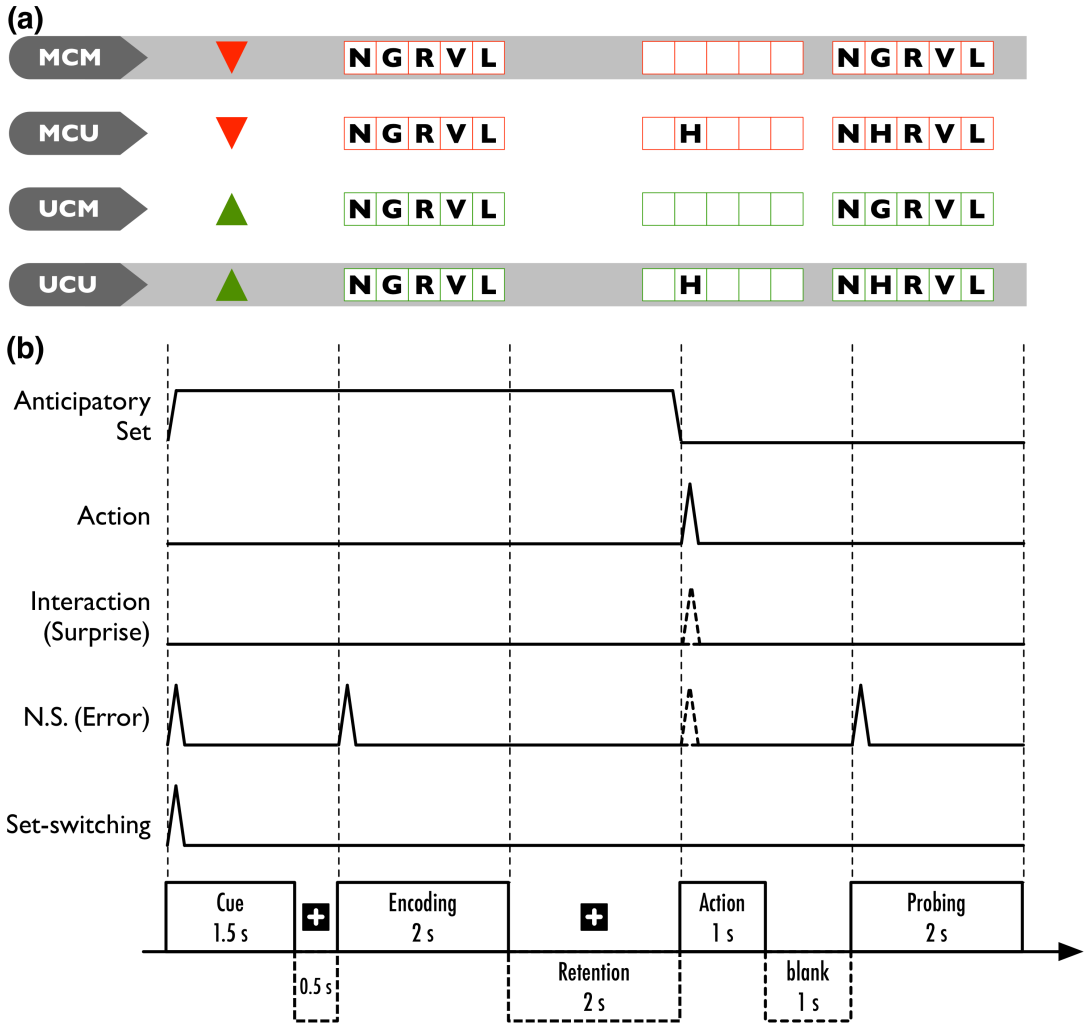

Figure 2.1 Stimuli and task design. ***a***, Each row illustrates an example of a predictive cue and subsequent memory arrays, under one of the four conditions: MCM, MCU, UCM, and UCU. The cues reported the update predictability, in which the probability of updating is 80% given a green cue and 20% given a red cue (equivalent to 80% maintenance probability). The shaded rows represent valid cue-outcome associations. The example shown used a consistent probe array, in which subjects should give positive ('true') responses. ***b***, Events and durations within a single trial shown as stimulus functions. Activity associated with anticipatory set (updating set and maintenance set) was modelled with a boxcar function of 6 s. Action (updating or maintenance) was modelled for both unsurprising and surprising trials, namely, nonspecific effects of updating and maintenance. Surprises (dashed spike in the Interaction row) were modelled separately for MCU and UCM conditions. Nonspecific task effects (NS) were treated as nuisance effects; upon error trial, the action onsets were modelled as nonspecific visual responses (the dashed spike on NS). Although the anticipatory set should be disengaged upon the display of action array, additional set-switching effects were added to model the cue onset.

High cues represented a low likelihood of maintenance and thus weighed more on cognitive flexibility, thus facilitating updating of representation; whereas maintenance was more likely under a low cue, where prior beliefs place high fidelity over the first (encoding) memory array.

Following the predictive cue, subjects were cued sequentially with three memory arrays during the encoding, action, and probing phases (Figure 2.1b. Each memory array contained a set of 5, 1, or 0 randomised English letters, arranged into a 1 x 5 grid presented in the same colour as the preceding updating or maintenance cue. This means some arrays had empty entries, depending on the function of the current phase. For example, the encoding and probing arrays always have five letters, which were non-repeating capital letters sampled randomly from 19 English consonants (excluding W and Y) to ensure phonologically distinct combinations. In addition, arrays were excluded if the letter sequence or its neighbour formed a common acronym or word.

During the action phase, the memory array can be a one-letter array or an empty array, cueing an updating or maintenance event, respectively. Specifically, the letter in the updating array was displayed in a random position, balanced throughout the task: this update letter was generated from the same set of consonants but excluding the five used in the preceding encoding array. The subjects always have to update their working memory upon seeing a letter-containing action array. This was achieved by replacing the encoding letter with the update letter at the corresponding position. A maintenance array would be empty, thus the encoded memory was not updated.

During the probe phase, the subjects made decisions about whether or not the probe array was identical to their working memory – that included an update (if it

had occurred). The subjects were informed that the probe would differ from the subjects' working memory by a single letter, with equal probability in each of the five positions. Subjects responded by pressing a 'true' key when they thought that the probe array matched the array they had in memory, or a 'false' key if they believed otherwise. A fixation cross was presented during inter-trial intervals and during the retention between the offset of the encoding array and the onset of the action array. The stimuli were presented using Cogent 2000 and Matlab (MathWorks).

The predictive cues were informative as they spoke to the statistical regularity embedded in the task. The subjects received explicit instructions regarding the veridicality of cue-action contingencies, and were encouraged to rely on the cue to guide task performance. This design assumed working memory as the realisation of predictive coding (Friston & Stephan, 2007), where cues enabled representations about (familiar) future states. We can therefore define surprise as departures from familiar outcomes, i.e. prediction errors. Accordingly, we defined 'valid' outcomes as trials where subjects had to update after a high probability (updating) cue, or they had to maintain after a low (maintenance) cue. There were two sources of prediction error in our task: *omissions*, an update failed to occur with a preceding high cue; *deviations*, an update of the memory representation under the belief of unlikely updating (low cue). Simply put, omissions and deviations were the interaction of anticipatory set and action.

The task therefore conformed to a 2 x 2 factorial design, with the two factors comprising anticipatory set (high vs. low update predictability) and action (updating vs. maintenance). This provided four conditions with regard to the valid and invalid (surprising) cue-outcome pairings: maintenance cue/maintenance (MCM),

maintenance cue/updating (MCU), updating cue/maintenance (UCM), and updating cue/updating (UCU).

The predictive cue was presented for 1500 ms, followed by a fixation cross for 500 ms. Then, the encoding array appeared for 2000 ms, followed by a fixation cross for 2000 ms, while the subjects maintained the items in the encoding array. The action array then appeared for 1000 ms, followed by a blank screen for 1000 ms. Finally, the probe array was presented for 2000 ms. Subjects were required to response as quickly as possible on the appearance of the probe array. Reaction times were measured from the onset of the probe array. The total duration of a single trial was 12 s, with an inter-trial interval of 2000 ms.

The task consisted of a single session of 100 trials. The maintenance cue and updating cue trials alternated every trial. There were equal numbers of true and false trials, counterbalanced across MCM and UCU conditions, as well as across MCU and UCM conditions. Each session lasted 1200 s. Subjects responded with their index and middle fingers of their right hand using an MRI-compatible keypad. In half of the subjects, the answer ‘true’ was mapped to the index finger and ‘false’ to the middle finger; in the other half, the converse was the case. To minimise nonspecific processing demands, the words ‘True’ and ‘False’ were visible on the lower third of each probe display, in the side of the corresponding response finger.

Immediately before the fMRI experiment, each subject underwent a 1 h instructed training session. Then, a 10-trial version of the task was administered with feedback to confirm the subject had understood the task. Each subject was required to achieve 100% accuracy to enter the second part of the training, which comprised 100 trials without feedback – to prepare the subject for the fMRI experiment.

2.2.2. Working memory capacity

Working memory may share a common neural substrate with the attentional system (Knudsen, 2007), and therefore may be subject to limited resources. The actual limit, depending on the underlying theoretical construct, can be up to four 'chunks' (N. Cowan, 2005) or 7 +/- 2 items (Jensen & Lisman, 1996). The assumption of limited resources predicts that memory updating may be modulated by variations in span limit. This line of reasoning is straightforward – in order to update memory, the brain must represent both the informative and the obsolete items to be able to manipulate them online. Updating while exceeding the capacity limit seems improbable. Moreover, under a more recent, precision-based capacity model (Ma et al., 2014), individual differences may still contribute to the effectiveness of updating, even though the number of items to be remember is less than seven. Indeed, it has been established that individual differences in working memory span may contribute to updating capacity (Ecker et al., 2010). Thus, we conducted a task prior to the training session to measure working memory capacity (WMC) for each subject – in order to control the effect of WMC on updating performance. The task required the subjects to recall a letter sequence in order. The letter sequence was based on the same stimulus set used in the updating task. The subjects viewed the letters on a black background at the rate of one letter per second. Strict forward recall was required, i.e., the subjects had to report the sequence in the order it was presented. The length began with four letters, and then increased by one letter with every two successful trials until the subject committed errors. During responding, the subjects were allowed to type and to make corrections before they submitted their answers. The highest span size performed correctly twice was recorded for each subject.

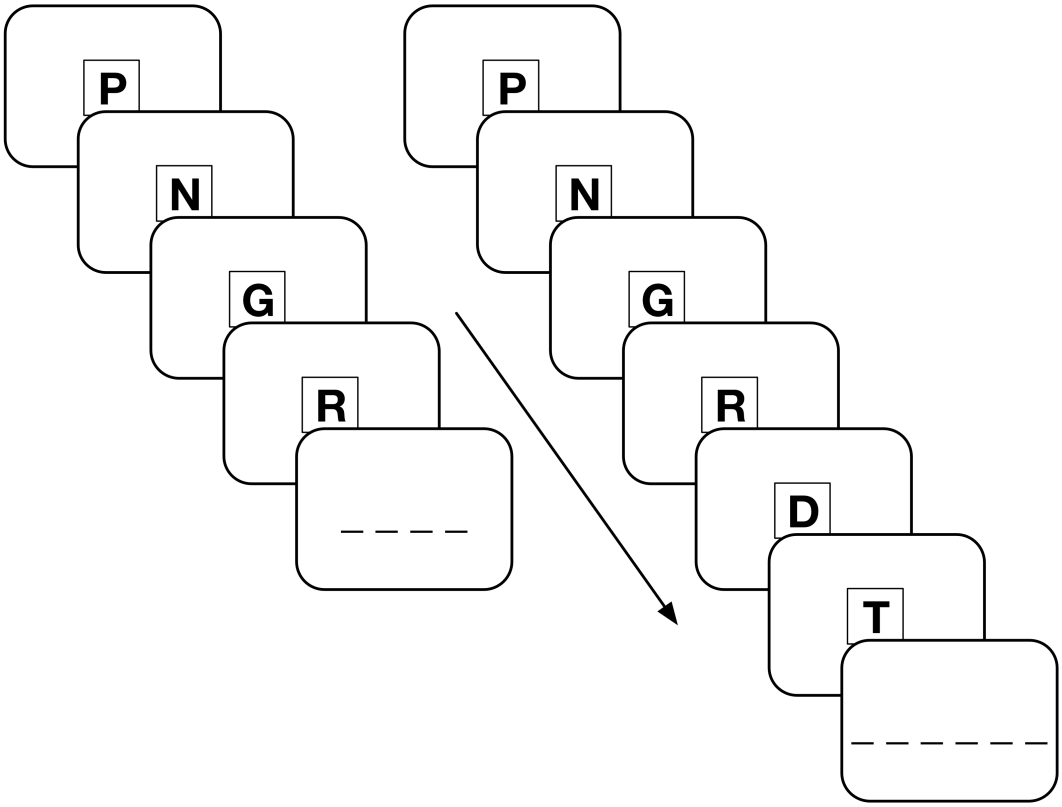

Figure 2.2 A serial recall task for measuring span limit. This diagram shows two levels of span size for testing working memory capacity. Subjects view and memorise letter sequences at one letter per second. After the sequence finished, subjects are prompted to recall the sequence with a keyboard. The span size may increase if subjects made two successful trials in a row. (left) a task showing span size of 4, and (right) a span size of 6.

The measure of individual working memory capacity was validated with an independent index based on Cowan's estimate (Cowan, 2005)

$$k = N(\text{hits} - \text{false alarms}) \qquad (2.1)$$

where *N* is the array size in the updating task. The validation is reported in Chapter 3.

2.3. Data acquisition

Structural and functional images were acquired on a 3 tesla Magnetom Trio MRI system (Siemens Medical Solutions). Functional images were acquired with a 32-channel head coil, using a single-shot echo planar imaging sequence (slice repetition time, 70 ms; echo time, 30 ms; ascending slice acquisition order; 3 x 3 x 3 mm voxel size). During functional acquisition, peripheral physiological variations were monitored by a respiratory belt and a pulse oximeter. Field mapping protocol was applied to sample field inhomogeneity (short echo time, 10 ms; long echo time, 12.46 ms; total EPI readout time, 37 ms). Multi-parameter images, including T1-density, proton density, and magnetisation transfer contrasts were acquired for structural information using 3D FLASH (fast low-angle shot) sequences.

2.4. Data analysis

2.4.1. Spatiotemporal preprocessing

Preprocessing of functional MRI data included: (1) 'unwarping' distorted image due to inhomogeneous B0 magnetic field; (2) approximating slice data to assume identical slice acquisition time with respect to experimentally elicited responses; (3) rigid translations and rotations to anatomically align images across scans; (4) a spatial 'normalisation' with reference to a standard stereotactic atlas, such that image data of all subjects are in voxel-wise correspondence; and (5) smoothing to accommodate small-scale differences in anatomical definition across subjects. All these procedures contribute to the efficiency of statistical tests and inferences made at group level.

2.4.1.1. Realignment

During the course of functional imaging, the head position of each subject may change: breathing, for example, may cause the head to move slightly. This results in inter-scan changes in anatomical alignment, which tends to confound subsequent voxel-wise analyses. Adjusting the functional images into a common frame of reference is therefore necessary. This is first performed within subjects using realignment – an affine registration, which is based on rigid-body transformations parameterised by six parameters (the translation of images in three dimensions, and rotations in three dimensions).

Formally, this procedure starts with a reference image, usually a grand average, and the original set of images as source images. Let $\boldsymbol{q}$ be the model parameters and $b_i(\boldsymbol{q})$ quantify the difference between source and reference image at voxel *i*. Realignment involves minimising the residual sum of square $\sum_i b_i(\boldsymbol{q}-\boldsymbol{t})^2$, given the model parameters are displaced by $\boldsymbol{t}$. Using the first order Taylor expansion, one obtains

$$b_i(\boldsymbol{q}-\boldsymbol{t}) \approx b_i(\boldsymbol{q}) - t_1\frac{\partial b_i(\boldsymbol{q})}{\partial q_1} - t_2\frac{\partial b_i(\boldsymbol{q})}{\partial q_2}\cdots \tag{2.2}$$

which can be rearranged into

$$\begin{bmatrix} \frac{\partial b_1(\boldsymbol{q})}{\partial q_1} & \frac{\partial b_1(\boldsymbol{q})}{\partial q_2} & \cdots \\ \frac{\partial b_2(\boldsymbol{q})}{\partial q_1} & \frac{\partial b_2(\boldsymbol{q})}{\partial q_2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \end{bmatrix} \approx \begin{bmatrix} b_1(\boldsymbol{q}) \\ b_2(\boldsymbol{q}) \\ \vdots \end{bmatrix} \tag{2.3}$$

From here, an iterative scheme is used such that the parameters can be optimised accordingly

$$\boldsymbol{q}^{(n+1)} = \boldsymbol{q}^n - (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{2.4}$$

2.4.1.2. Unwarping

Inhomogeneity in external magnetic field B0 is one of the major sources of MR artefacts. Similar to magnetic susceptibility artefacts, field inhomogeneity tends to distort images. However, it is possible to 'unwarp' the distorted images by means of field mapping, prior to the scanning. Field mapping generates images in which field inhomogeneities are quantified, and can be used to specify a forward model of movement-by-inhomogeneity interactions which is subsequently inverted (Andersson, Hutton, Ashburner, Turner, & Friston, 2001; Hutton et al., 2002).

2.4.1.3. Slice-timing correction

The slice-timing problem refers to different slice acquisition times. Normally, one would expect the acquisition of a single scan (volume), which containing multiple slices, to be in a consistent time frame. This is, however, not possible in practice. For example, given a scanning sequence of repetition time *T*, and a descending slice acquisition order. The time the last slice is acquired is later than the time at which the first slice was acquired by around *T*. In other words, slices within a single volume are acquired at different times. In this case, if the experimentally induced responses are modelled with a single canonical haemodynamic function, and the onset time is set to the beginning of each scan, the parameter estimates in the bottom slices will be biased.

This problem can be adequately alleviated by means of interpolation in time (Henson, Büchel, Josephs, & Friston, 1999). Depending on the interpolation method, the weighted average of two or more time points is calculated. For example, Figure 2.3 illustrates the case where two slices (grey box) are used to approximate a time point (red box) using linear interpolation. The disadvantage of linear interpolation

lies in the potential to smooth the raw data, if the two slices are equally distant from the time point to be interpolated. More complex interpolation schemes, such as those based on cubic, spline, or sinc functions, may be adopted, with the risk of introducing artefacts from distant slices.

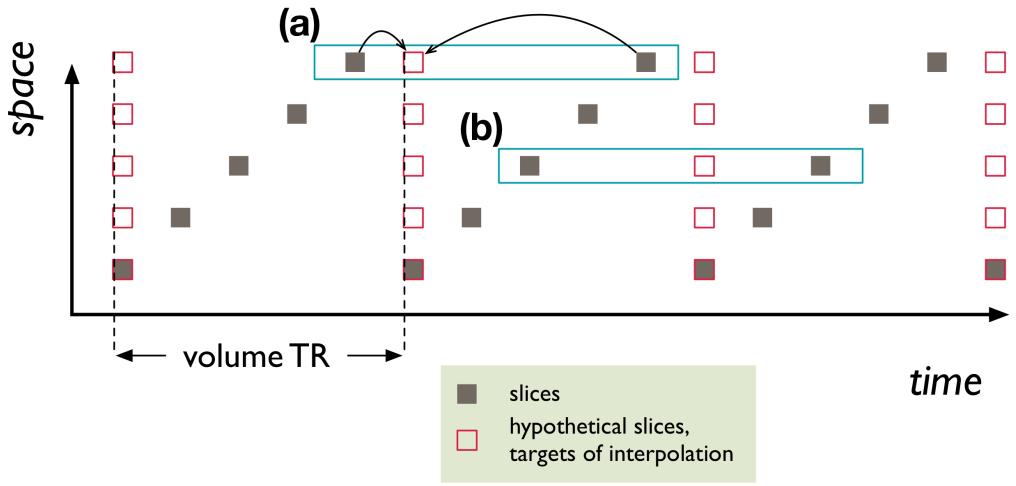

Figure 2.3 Slice-timing correction using a linear interpolation. Slice-timing correction uses interpolation to approximate fMRI slice data to an aligned temporal position. Interpolation here is a weighted sum of neighbouring slices. ***a***, the interpolated time point (red box) will closely resemble the slice data to the left, while the right slice has relatively little contribution. ***b***, slices from both sides have nearly equal contributions to the interpolated slice. Note that in the latter case, the newly generated slice is a smoothed copy of the two. To prevent data from excessive smoothing, cubic, spine, or sinc interpolations may be adopted, which take further slices into account, although this could incur more artefacts.

2.4.1.4. Spatial normalisation

Reporting summary statistics is a common practice in almost every scientific discipline. However, the issue with high dimensional data like structural and functional images is that the geometry of individual brains is never the same. Therefore, we need a standard stereotactic space into which individual brains can be

transformed, making the reference to common anatomical framework at the group level straightforward, and thus enabling efficient statistical analyses.

Spatial normalisation is achieved in two steps (Friston et al., 2004). First, a 12-parameter affine registration is employed, with reference to a standard template in the MNI (Montreal Neurological Institute) space. The 12 parameters stand for transformations along the *x*-, *y*-, and *z*-axis in the form of rotation, translation, sheer, and zoom. Zoom and sheer are needed to register heads/brains of different shapes and sizes, whilst rotation and translation match the orientation to the template image. Prior information concerning the variability of head sizes is incorporated under a Bayesian framework, this prevents over-fitting, as well as increases the robustness and accuracy of the fine-grained warping in the next step. Next, a non-linear warping is introduced to correct differences that cannot be accounted for by the linear transformation. This warping can be seen as a mapping from the native image space into the standard space; the mapping is described in terms of a linear combination of non-linear basis functions. Regularisation is introduced by minimising the sum of squared difference between the warped image and the template in order to avoid over-fitting.

2.4.1.5. Smoothing

In SPM processing pipeline, this is usually the final step, which uses a Gaussian smoothing kernel applied to the 3-dimensional volume of data. The idea is similar to that of moving average, in which each data point is a weighted mean of neighbouring points within a pre-defined window, except the window and the weights in smoothing are now a Gaussian 'sphere' in space. The overall effect of smoothing creates a blurred version of the original images. The extent, i.e., the 'window' width,

of the smoothing is determined by the full-width at half-maximum (FWHM) of the kernel. Apart from accommodating cross-subject small-scale differences in anatomical definition that are unaccounted for by the normalisation process, insufficient or excessive smoothing are suboptimal. This is because statistical inferences within SPM rely on Random Field theory to resolve multiple comparison problems. Random Field theory assumes the response is spatially smooth. To model the dispersion and number of spurious response occurring by chance, a smoothed random noise field is used to evaluate the circumstance stated. In other words, a less stringent correction is applied to highly smoothed data. Briefly, smoothing may help to improve statistical sensitivity by either increasing signal-to-noise ratio, or by inducing normal error distributions in accordance with the assumption of most parametric tests. These come at the cost of being unable to make inferences about smaller cortical structures, and the possibility that focal activation peaks may be merged or completely removed.

2.4.2. Behavioural data

2.4.2.1. Working memory capacity

Our measure of working memory capacity – by means of serial recall – was validated by regression with an independent measure of Cowan's capacity index. The reason for an independent measure of individual span limit is that Cowan's index depends on the hit rate and false alarm rate, it is therefore co-dependent with response accuracy or a score that conflates proportions of correct responses (Bruyer & Brysbaert, 2013).

From Equation 2.1, Cowan's index was calculated by

$$2k = N_u(H_u - F_u) + N_m(H_m - F_m) \tag{2.5}$$

where subscripts denote updating (u) and maintenance (m) trials. N_m is equal to 5, as this is the size of the memory array; whereas N_u is equal to 6 as an additional letter was introduced by the action array. The simple linear regression is now given by

$$\boldsymbol{k} = \alpha + \beta \boldsymbol{C} + \boldsymbol{\epsilon} \tag{2.6}$$

where $\boldsymbol{C}$ denotes individual span size as measured by the serial recall task.

2.4.2.2. Reaction time and response accuracy

Performance of a working memory updating task may be modulated by an individual's working memory capacity. This was suggested by Schmiedek et al. (2009) and was tested by Ecker, Oberauer, and Chee (2010), who showed that both working memory updating and working memory capacity are strongly related and predict higher cognitive abilities to a similar degree (2010). Although, according to Ecker et al. (Ecker et al., 2010), substitution as a component process of updating (the other two being retrieval and transformation) does not seem to be substantially predicted by working memory capacity, we nevertheless considered the influence of individual difference in capacity limit in the following behavioural analysis: substitution is a key manipulation to enable updating in our task, and depends on the retrieval of previously encoded information (Chen & Li, 2007).

The purpose of the analyses reported here is to detect whether there was surprise-induced impairment in task performance. Therefore, individual measures of working memory capacity are treated as covariates in the subsequent analysis of covariance (ANCOVA). Response accuracy was calculated as the proportion of correct responses and the measures of reaction time were summarised within subjects using harmonic mean to control for outliers (Ratcliff, 1993). The number of reaction time measures in invalid trials (updating cue/maintenance or maintenance

cue/updating) was inherently rare, as required by a prediction/surprise paradigm. The use of the harmonic mean is considered suitable to reflect the central tendency in these trials.

The harmonic mean for reaction time is the number of correct sample divided by the sum of inverse reaction times. In other words, it is the reciprocal of the arithmetic mean of the reciprocals, and is given by

$$H = \frac{n}{\sum_i \frac{1}{x_i}} \tag{2.7}$$

It is noted that the harmonic mean has a tendency towards smaller values, thereby alleviating the impact of large outliers.

For each subject, the harmonic means were calculated for each condition. Reaction time and accuracy were analysed in two different ANCOVAs, with working memory capacity entered as a between-subject covariate. The main effect and interaction between conditions were then tested.

2.4.3. Imaging data

2.4.3.1. General Linear Models

2.4.3.1.1. Background

The GLM is a general regression framework from which various types of statistical testing can be realised. The application of GLM in neuroimaging is often referred to as mass-univariate analysis. This is because the method treats voxel-wise time series independently and analyses them as if there are multiple instances of univariate data. The independence assumption is, however, not realistic as spatially

adjacent voxels are likely co-dependent, which, if not taken into account, will render the subsequent statistical analysis inefficient due to an overly severe correction for multiple comparisons. We have briefly covered the idea of controlling false positives previously in the *smoothing* section.

The objective of using GLM is to make inferences about experimental effects of interest. To achieve this, one has to decompose the fMRI time series into task effects and error in terms of model parameter estimates, from which appropriate statistics can be motivated. The idea of decomposition is based on the assumption that the observed fMRI responses are generated as linear combinations of some *explanatory variables* (Figure 2.4). The explanatory variables are also known as independent variables. The fMRI data (*dependent variables*) are known quantities. For example, visual stimuli are displayed in a succession of on/off blocks; in this case regional responses associated with visual inputs may be modelled with an explanatory vector of 0s and 1s in which 1s are present in accordance with the scans of visual onset.

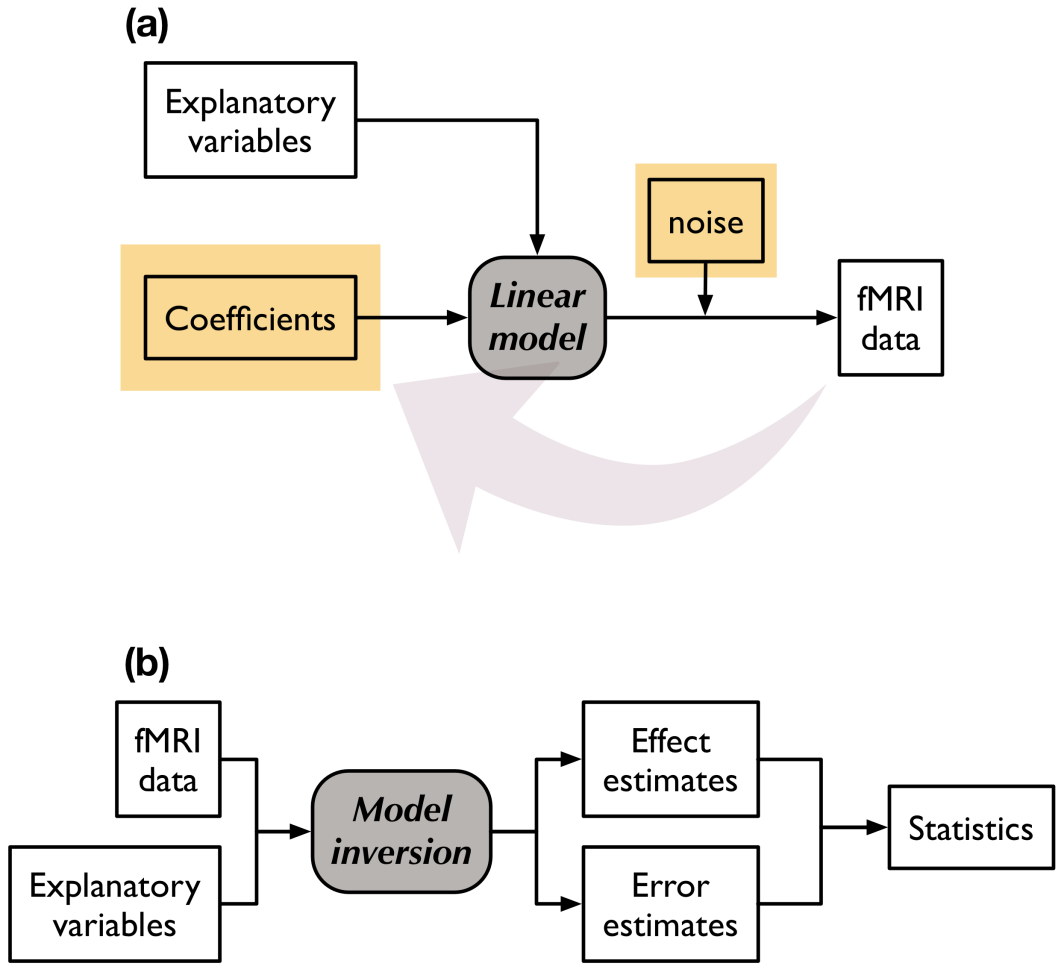

Figure 2.4 A schematic diagram showing data generating process and its inversion. ***a*****,** The diagram shows how fMRI data are generated as linear combinations of designed experimental perturbations as explanatory variables. The unknown quantities are the coefficients that determine the contribution of individual explanatory variables to the final data. The noise is zero mean with unknown variance. ***b*****,** The unknown quantities can be derived through model inversion. Ordinary least square estimates are typical for linear models, and involve minimisation of the sum-of-square error between model predictions and observations. The estimates can be used to form statistics and subsequently make inferences.

From what we have described, we can now write

$$\hat{\boldsymbol{Y}} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 \tag{2.8}$$

where $\boldsymbol{x}_1 = [0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ \cdots]^T$ denotes the visual onsets and offsets and its corresponding coefficient β_1. The second term is just a vector of ones modelling the

mean of the response $\hat{\boldsymbol{Y}}$. This description is, however, not satisfactory in two regards: first, $\hat{\boldsymbol{Y}}$ is the model prediction and differs from the observation by error $\epsilon = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. One therefore needs a way to ensure the model prediction is as close as possible to the observation, such that the error is minimised. Secondly, $\boldsymbol{x}_1$ does not adequately reflect the haemodynamic responses that are observed in BOLD imaging. A common practice for the latter issue is to convolve the stimulus function with the canonical haemodynamic response function to create an explanatory variable that resembles the true response.

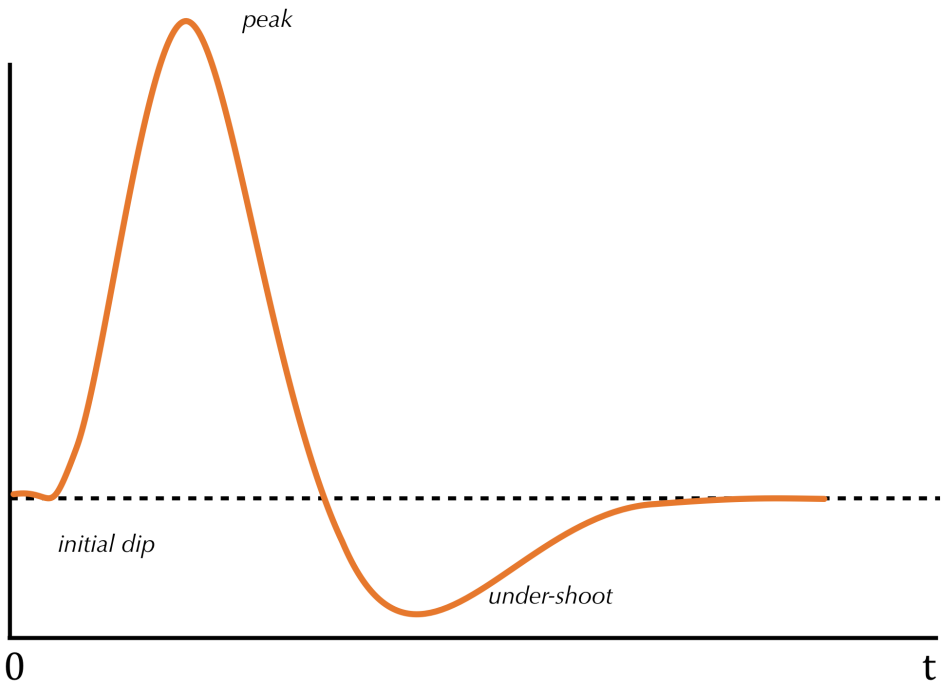

Figure 2.5 The canonical haemodynamic response function. Zero indicates stimulus onset time. A canonical haemodynamic response function is characterised by an initial dip, a peak, followed by an under-shoot.

2.4.3.1.2. Matrix form

Equation (2.8) shows how a time series can be described in terms of a linear combination of explanatory variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, with coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, this may only represent the time series of a single voxel, and with a single experimental variable. We can easily extend this to include multiple experimental variables using matrix notation, given by

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_j \\ \vdots \\ \boldsymbol{Y}_J \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & \cdots & x_{jl} & \cdots & x_{jL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{J1} & \cdots & x_{Jl} & \cdots & x_{JL} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_l \\ \vdots \\ \boldsymbol{\beta}_L \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_j \\ \vdots \\ \boldsymbol{\epsilon}_J \end{pmatrix} \tag{2.9}$$

or

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.10}$$

where $\boldsymbol{Y}$ is a column vector of observations, with each element corresponding to data acquired at time j, $\boldsymbol{\beta}$ the column vector of coefficients (parameters), $\boldsymbol{\epsilon}$ the column vector of error terms. $\boldsymbol{X}$ is the *design matrix*, in which each column corresponds to one observation in time and each column corresponds to one experimental manipulation. Note that the design matrix is a near-complete description of the model, which leaves the remaining, unexplained quantities to the noise terms, the distribution of which is assumed by the model.

Ordinary least squares are used to find the parameters, which generate model predictions that are best fit to our observations. This in effect minimises the residual sum-of-square. Let $\hat{\boldsymbol{\beta}}$ be the optimal parameter estimate, the ordinary least square estimate is given by

$$
\begin{aligned}
\epsilon^T\epsilon &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\
\epsilon^T\epsilon &= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) \\
\epsilon^T\epsilon &= Y^TY - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta} \\
\frac{\partial \epsilon^T\epsilon}{\partial \hat{\beta}} &= -2X^TY + 2X^TX\hat{\beta} \\
0 &= -2X^TY + 2X^TX\hat{\beta} \\
\hat{\beta} &= (X^TX)^{-1}X^TY
\end{aligned}
\tag{2.10}
$$

2.4.3.1.3. Geometrical representation

An intuition is given from a geometrical perspective which, in essence, demonstrates that the least square estimate is in fact an orthogonal projection of the observation $\boldsymbol{Y}$ onto the design space $\boldsymbol{X}$ (Figure 2.6). From the definition earlier, we can see our observation $\boldsymbol{Y}$ as a vector in a *J*-dimensional (Euclidean) space, denoted by $\mathbb{R}^J$. The columns of the design matrix are *J*-basis that spans a subspace in $\mathbb{R}^J$. The perpendicular from $\boldsymbol{Y}$ to the subspace meets the subspace at $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and the distance between them corresponds to the error.

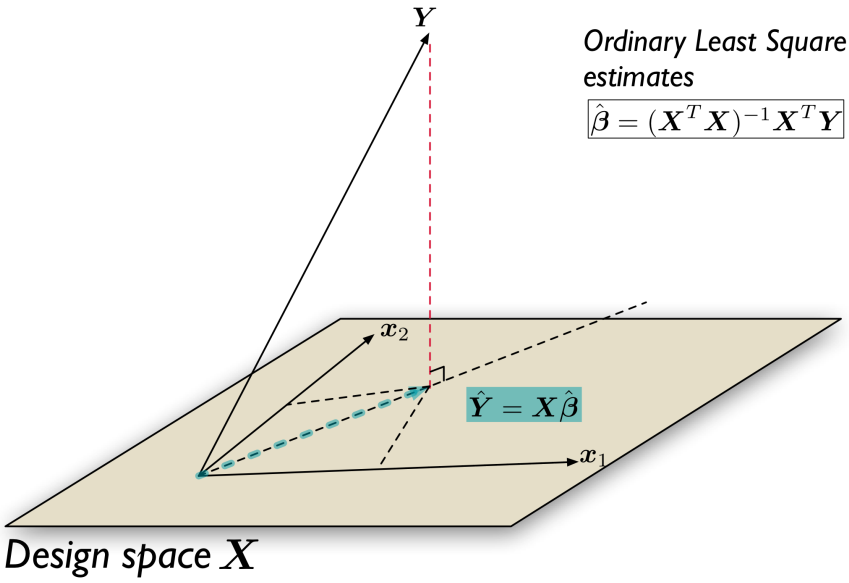

Figure 2.6 Geometrical intuition on linear regression. The ordinary least square estimates (blue shaded arrow) are equivalent to the orthogonal projection of the observation onto the subspace spanned by the columns of the design matrix. The error terms are described by the distance between the observations and the subspace (red dash).

The geometrical notion is also useful for illustrating correlated column vectors (or regressors) in the design matrix. Correlated regressors may be, for example, imposed by the experimental design, in which some conditions inevitably exhibit some co-linearity, such as tasks involving reward prediction. Co-linearity may lead to misinterpretation of the resulting statistical parametric maps and should be avoided. Figure 2.7 illustrates an orthogonalisation process in geometrical terms. Specifically, with correlated regressors, variance explained is shared between regressors. Only when one regressor is orthogonalised with respect to the other regressor, interpretation of an experimental effect can be independently attributed to one regressor.

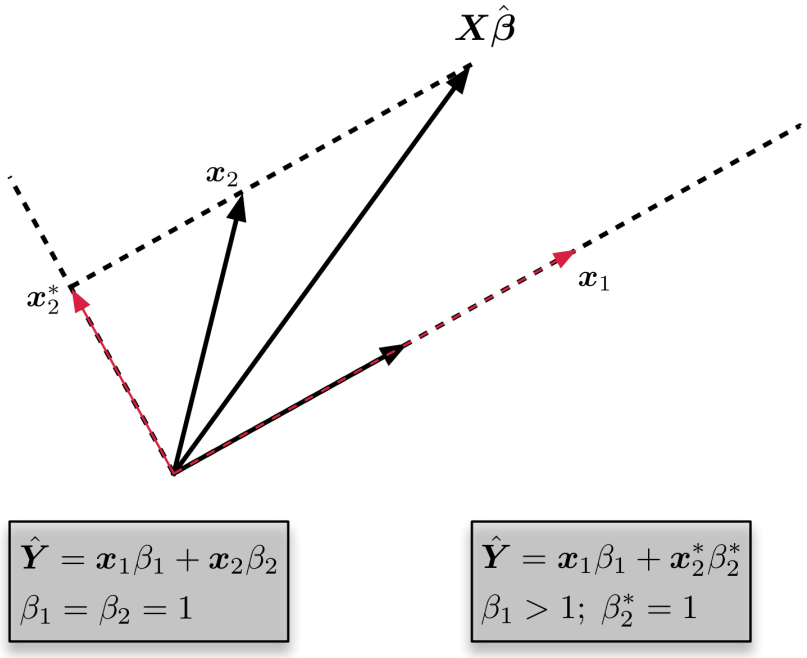

Figure 2.7 Correlated and orthogonal regressors. The schematic illustrates the orthogonalisation of vector $\boldsymbol{x}_2$ with respect to $\boldsymbol{x}_1$. After orthogonalisation, the orthogonal vector $\boldsymbol{x}_2^*$ will have the same effect on $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ but the effect of $\boldsymbol{x}_1$ will apparently increase.

2.4.3.1.4. Remaining issues

There are several other issues commonly encountered in the application of GLM for fMRI data. One of them is of the shape of BOLD response, which we have briefly covered above. One has to translate the experimental effect in terms of input stimulus functions into that of BOLD-relevant response. The solution is to use a convolution model with an impulse response function that generates expected BOLD responses. This impulse response function is the canonical haemodynamic response function. The underlying assumption is that the BOLD signal is the output of a linear time-invariant system, in which (1) responses have the same form irrespective of time and (2) successive responses superimpose linearly (Boynton, Engel, Glover, &

Heeger, 1996). Briefly, the response of a linear time-invariant system is the convolution of the input (experimentally designed perturbations, stick or boxcar stimulus functions) with the (BOLD) system's response to an impulse (HRF).

The second issue is systematic fluctuations or artefacts. Failing to account for such confounds would result in exaggerated noise estimates and have a serious impact on the efficiency of statistical inference. These fluctuations are usually environmental and of low frequency, such as the 'scanner drift' that is caused by variations of the main magnetic field over time. Adjusting the observed signal with a high pass filter is a common solution. There are a variety of high-pass filtering techniques; the one applied in SPM makes use of a discrete cosine transform set (DCT; Figure 2.8). The DCT set is described by a number of cosine bases of 0.5, 1, 1.5 cycles and so on during the time course of a scanner session. These bases can also take a matrix form and be incorporated in the design matrix. Filtering the observed signals corresponds to applying the residual forming matrix of the DCT set to our data. A residual forming matrix is defined by

$$\boldsymbol{r} = \boldsymbol{I} - \boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T \tag{2.11}$$

where $\boldsymbol{I}$ is an identity matrix and $\boldsymbol{S}$ the DCT set. This necessitates the following form of the original GLM problem

$$\boldsymbol{rY} = \boldsymbol{rX}\boldsymbol{\beta} + \boldsymbol{r}\boldsymbol{\epsilon} \tag{2.12}$$

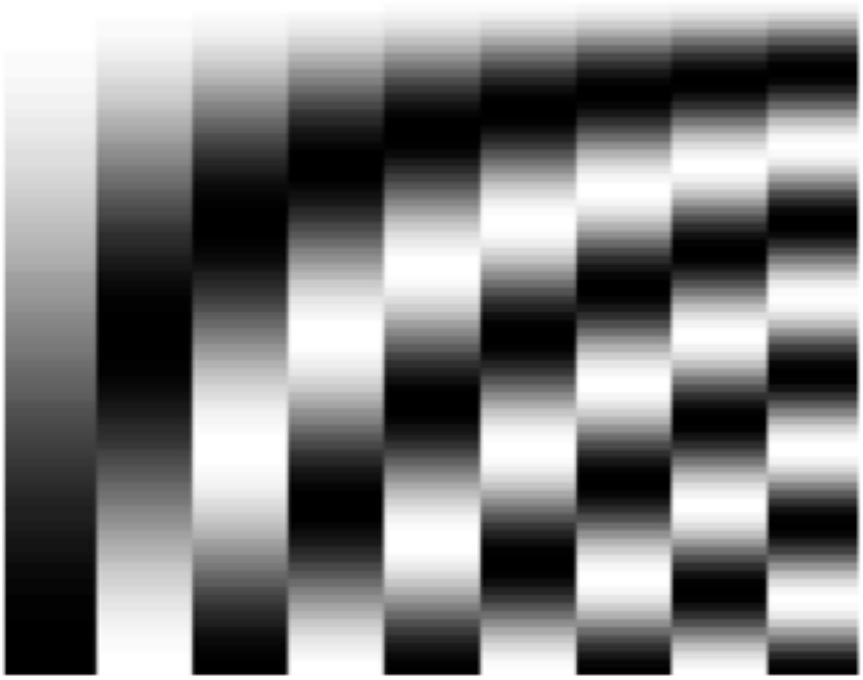

Figure 2.8 A set of bases representing a discrete cosine transform (DCT). Columns from left to right represent cosine functions of 0.5, 1, 1.5 cycles and so on within the time course of a scanning session. Filtering the observed data using the DCT set is similar to the inversion of a GLM, except the design matrix is the null space of the DCT set.

Finally, we turn to the issue of 'non-sphericity'. The term 'sphericity' refers to the assumption that noise is independent and identically distributed (i.i.d.) across observations. Noise that is i.i.d. will have a covariance matrix $\sigma^2 \boldsymbol{I}$ such that its entries are equal to zero except for the main diagonal (Figure 2.9).

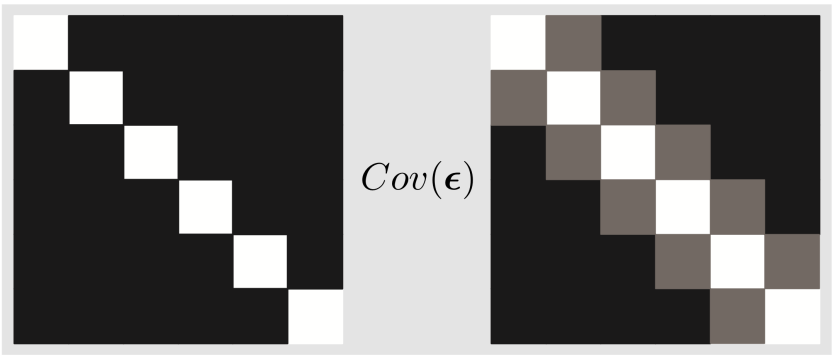

Figure 2.9 Covariance matrices under sphericity and non-sphericity. Noise covariance corresponding to six hypothetical observations. (left) Noise is independent with each other, as depicted by the white pixels along the main diagonal. Positives are brighter. (right) The noise is somewhat co-dependent on the ones adjacent to it, as depicted by the grey pixels off-diagonal.

If the noise terms are co-dependent, the covariance will have non-zero off-diagonal terms. In other words, the noise of some observations is correlated. In order not to violate the sphericity assumption of GLM, one has to 'de-correlate' the noise. A de-correlation technique is called *whitening*. The whitening matrix is given by the inverse square root of the error covariance matrix

$$\boldsymbol{W} = \left(\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\right)^{1/2} \tag{2.13}$$

which is relatively easy to evaluate (see below).

However, we do no know the covariance matrix of our noise, therefore it has to be estimated. One way to estimate the covariance matrix is by assuming a first-order autoregressive process for the noise. Alternatively, an enhanced noise model may be employed; this involves multiple covariance components $\boldsymbol{V}$ which in effect replace $\sigma^2\boldsymbol{I}$ by $\sigma^2\boldsymbol{V}$, given by the enhanced noise model $\epsilon \sim N(0, \sigma^2\boldsymbol{V})$. Suppose we know $\boldsymbol{V}$ and let $\boldsymbol{W}$ be the whitening matrix, we have the following relationship

$$\begin{aligned} &\boldsymbol{W}\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{W}^2\boldsymbol{V}) \\ &\quad \Longrightarrow \ \boldsymbol{W}^2\boldsymbol{V} = \boldsymbol{I} \\ &\quad \Longrightarrow \ \boldsymbol{W} = \boldsymbol{V}^{-1/2} \end{aligned} \tag{2.14}$$

$\boldsymbol{V}$ is a linear combination of several covariance components $\boldsymbol{Q}_i$

$$\boldsymbol{V} = \sum_i \lambda_i \boldsymbol{Q}_i$$

where λ_i is a model hyper-parameter. A common way to derived the proportions of hidden mixtures is through the Expectation-Maximisation (EM) or restricted maximum likelihood (ReML) (Friston et al., 2004).

2.4.3.1.5. Statistical inference

We have considered the issue of variance components in our observation, as well as other issues relating to high-pass filtering and modelling with canonical haemodynamic response functions. Next, we briefly describe how classical inferences with *t* and F statistics are carried out in a GLM.

In fMRI studies, a key question relates to whether or not a voxel is 'activated' by the experimental manipulation. We can rephrase this question by asking whether the mean of its coefficient is different from zero. In statistical terms, the question is expressed in terms of hypotheses. Typically, the approach is to first propose a hypothesis of a null measurement – the null hypothesis H_0. Refuting the null hypothesis implies the outcome of interest – referred to as the alternative hypothesis H_A. To formally test the hypothesis, the distribution of *test statistics* under the null hypothesis is constructed. The distribution may appear differently depending on the type of statistic. For example, a Student's *t* distribution has a bell shape. Figure 2.10 gives an illustration of the null distribution of *t* statistics. The distribution summarises evidence about the null hypothesis; namely, it reports the probability of a specific statistic being observed under the null distribution. The principle of hypothesis testing therefore conforms to the control of an acceptable false positive rate α. The false positive rate can be visualised as the fraction of the total area under

the distribution from one of its tails, as shown in Figure 2.10. This can be expressed as

$$\alpha = p(t > u_\alpha | H_0) \tag{2.15}$$

where u_α corresponds to the *t*-statistic threshold. That is to say, if we specify a false positive rate, say, $\alpha = 0.05$ or $\alpha = 0.01$ for controlling false positive rates at 5% and 1%, respectively, we can determine the lower bound of *t* statistics that yields the rates stated, given an appropriate null distribution. We can then calculate the *t* statistic corresponding to our observations and if the statistic falls into the right side of the threshold, significance is declared, namely, the null hypothesis can be rejected and the alternative hypothesis be accepted.

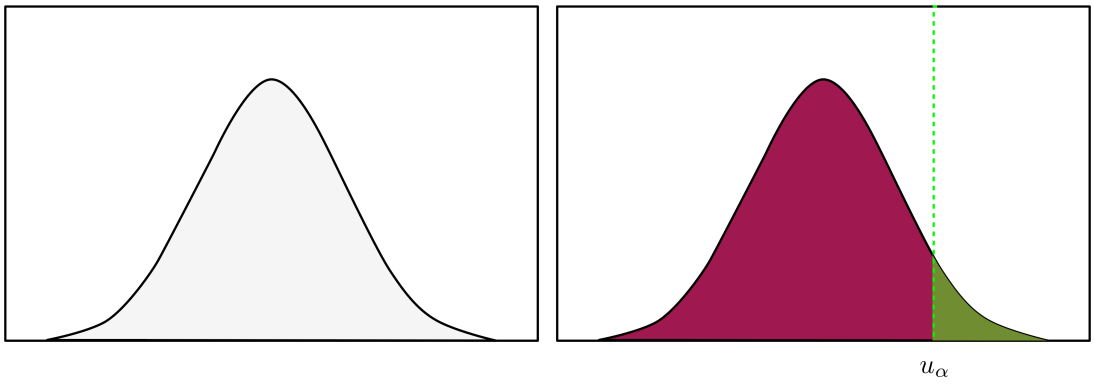

Figure 2.10 Schematic of a null distribution of *t* statistics. (left) the null distribution of *t* statistics has a bell shape; the exact shape of the distribution depends on the degrees of freedom. (right) The green shaded area on the right tail represents an acceptable false positive rate under the specified null distribution. For example, the green area accounts for 5% of the total area under the bell curve. This entails a minimum *t* threshold u_α and if our test statistic falls onto its right side, the alternative hypothesis is accepted. Figure is not drawn to proportion.

In SPM, the first step to test an effect of interest is specifying a 'contrast' $\boldsymbol{c}$. The contrast takes the form of a column vector where the number of elements corresponds to that of the columns of the design matrix. For example, to test whether

a voxel is activated due to the effect encoded in the first column of the design matrix, the contrast vector will have the first element equal to 1, whilst the remaining elements equal to 0. This tests whether voxel-wise amplitudes are greater then zero

$$\beta_1 = \boldsymbol{c}^T\boldsymbol{\beta} > 0? \tag{2.16}$$

and the null hypothesis is given by

$$H_0 = \boldsymbol{c}^T\boldsymbol{\beta} = 0$$

The *t* statistic is given by

$$t = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}}{\sqrt{var(\boldsymbol{c}^T\hat{\boldsymbol{\beta}})}} = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}} \tag{2.17}$$

Following the previous section, where we derived the whitening matrix using the enhanced noise model for variance components, we can re-write Equation 2.17 into

$$t = \frac{(\boldsymbol{W}\boldsymbol{X})^+\boldsymbol{X}\boldsymbol{Y}}{\sqrt{\hat{\sigma}^2\boldsymbol{c}^T(\boldsymbol{W}\boldsymbol{X})^+(\boldsymbol{W}\boldsymbol{X})^{+^T}\boldsymbol{c}}} \tag{2.18}$$

in which

$$\hat{\sigma}^2 = \frac{\sum\left(\boldsymbol{W}\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^2}{tr(\boldsymbol{R})} \tag{2.19}$$

and

$$\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{W}\boldsymbol{X})^+ \tag{2.20}$$

is the residual forming matrix. The notation $(\boldsymbol{W}\boldsymbol{X})^+$ corresponds to $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

From Equation 2.17 it is obvious that the *t* statistic does not depend on the scaling of the contrast vector or that of the design matrix. However, the contrast $\boldsymbol{c}^T\boldsymbol{\beta}$ itself does depend on the scaling of the contrast vector. Crucially, contrasts are often used as dependent variables to construct a second-level inference, for example, as a

group inference. Therefore, one needs to be mindful of scaling biases whilst specifying the contrast vector.

Next, we turn to another frequently used statistic in SPM, the F statistic. The main idea of the F test can be summarised as model comparison. The 'models' being compared here refer to a reduced design matrix and the original (full) design matrix. The test statistic pertains to the ratio of explained variability versus unexplained variability. In other words, the F test can be viewed as testing for the additional variance explained by a model with all parameters with respect to a (nested) one with less parameters. Informally, the F statistic is given by

$$F \propto \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}} \tag{2.21}$$

where 'RSS' is the abbreviation of residual sum-of-square (error).

In SPM, the F statistic is calculated following the form of the previous equation

$$F_{d_1,d_2} = \frac{(\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_0})\boldsymbol{Y} - \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y})/\nu_1}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y})/\nu_2} \tag{2.22}$$

in which

$$\begin{aligned} \nu_1 &= \mathrm{tr}(\boldsymbol{R}_0 - \boldsymbol{R})\boldsymbol{V} \\ \nu_2 &= \mathrm{tr}(\boldsymbol{R}\boldsymbol{V}) \end{aligned} \tag{2.23}$$

and the effective degrees of freedom are

$$\begin{aligned} d_1 &= \frac{\mathrm{tr}(\boldsymbol{R}_0 - \boldsymbol{R})\boldsymbol{V}(\boldsymbol{R}_0 - \boldsymbol{R})\boldsymbol{V})}{\mathrm{tr}(\boldsymbol{R}\boldsymbol{V})^2} \\ d_2 &= \frac{\mathrm{tr}(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}\boldsymbol{V})}{\mathrm{tr}(\boldsymbol{R}\boldsymbol{V})^2} \end{aligned} \tag{2.24}$$

In practice, a multi-column contrast, i.e., a contrast matrix, can be constructed to test multiple linear hypotheses within the same framework. This is extremely useful when testing for an 'effect of interest' that corresponds to variability modelled by

several regressors. We will use this approach to test the haemodynamic state variables for DCM in Chapter 5.

2.4.3.2. Eigendecomposition

In the ensuing chapters, we fit our data to dynamic causal models, which enable inferences in terms of effective connectivity. This involves extracting regional BOLD responses. For each region, SPM uses eigendecomposition to extract the temporal mode of all voxels within a specified region. This method may be superior to, for example, taking the arithmetic mean of the time series across voxels. Consider two extreme but hypothetical cases: one in which voxels within region varies almost identically, resulting a near-perfect correlation, and the other with equal numbers of voxels fluctuating in opposite direction. In the first case, the eigendecomposition method will closely resemble that using arithmetic mean. The other, however, will have a time series of zeros for mean, whilst the eigendecomposition will capture fluctuations over time.

In SPM, the eigendecomposition uses the singular value decomposition (SVD), given by the following form

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{2.25}$$

in which $\boldsymbol{Y}$ is a matrix of our BOLD time series with the size of time points-by-voxels, $(J \times N)$. $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary orthogonal matrices of size $(J \times J)$ and $(N \times N)$, which means they have uncorrelated columns and their respective sum of squares is equal to 1. Each column of $\boldsymbol{U}$ and $\boldsymbol{V}$ can be interpreted respectively as features in space and in time. $\boldsymbol{\Sigma}$ is a $(J \times N)$ matrix. Entries along the main diagonal of $\boldsymbol{\Sigma}$ are singular values, which reflect the amount of the variance expressed by the corresponding eigenvectors. The singular values are usually arranged in descending

order; therefore the first column of $\boldsymbol{U}$ contributes the greatest variability. This is the information extracted by SPM and is referred to as the first (or principal) eigenvariate. The data $\boldsymbol{Y}$ here represents data that has been filtered, whitened and 'adjusted' for null effects (using a F-contrast).

2.4.3.3. Support Vector Machines

In the following section, we briefly introduce the concept of machine learning. I also provide a general summary on machine learning applications in recent neuroimaging studies. This is outlined by asking what machine learning can do for neuroimaging. Specifically, a class of machine learning algorithm known as support vector machines was adopted in the work reported in this thesis. I briefly review its theoretical background, as well as other methodological considerations, i.e., the cross-validation and permutation testing.

2.4.3.3.1. Multivariate pattern analysis

The idea of localisation and modularity of neural activity in association with specific experimental factors is central to many neuroimaging studies, including those reported in Chapter 3. In pursuit of this idea, the use of GLM – or the mass-univariate analysis – is perfectly suitable. Despite fruitful GLM-led studies, the fact that our brain processes information within a constantly interacting, distributed network has motivated the application of Independent Component Analysis (ICA) and principle component analysis (PCA), particularly in resting state studies (e.g., Murty et al., 2014). This is perhaps because there is no need to pose any presumption about the temporal profile of the resting BOLD response. A limitation of ICA is that it may or may not be able to isolate components that speaks to a task-relevant effect

of interest (see Svensén, Kruggel, & Benali, 2002 and for the use of ICA and regression).

Another limitation with multivariate methods like ICA is of the generalisability of a component to another. This speaks to the predictability of one mental state given a relevant mental state. If one can derive such predictions formally, a strong form of reverse inference may ensue (e.g., Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). This is made possible by means of statistical learning theory and machine learning techniques. These techniques are at times referred to as pattern classification or data mining. This is known in neuroimaging as MVPA (Multivariate Pattern Analysis; Haxby, 2012). In essence, MVPA no longer treats fMRI data in a voxel-wise manner. Rather, it takes all voxels in to account. Namely, for a cognitive state, all voxels may contribute to a certain extent. The contribution is quantified by the classifier weight, which corresponds to how informative a voxel is for the classification problem. In addition, the amplitude of a voxel response is secondary. Rather, the overall *pattern* of voxel response is now characteristic of a neural code or a mental state.

Patterns may be obtained in a number of ways. One is to extract the raw fMRI data – if the underlying cognitive state in question in not contaminated by other confounding effects, and if the haemodynamic delay with respect to the timing of experimental effect is adequately accounted for (Schrouff et al., 2013). Alternatively, a GLM may be employed and the corresponding parameter estimate can be used as patterns for classification (Nee & Brown, 2012; Schrouff et al., 2013). The pattern-acquiring process is referred to as *feature selection* in a regular MVPA pipeline. In binary classification problems, two sets of features are gathered from two *known* categories (e.g., visual response to faces and houses). These features form a *training*

set, which is used by the machine learning algorithm of choice to discover the intrinsic regularity that sets the two feature types apart, a decision function is derived as a consequence, with the weight matrix as its parameters. During this process, a subset of data features is left out of the training. The left-out features are instead provided to the classifier after the training. Because we know to which category these features belong, it is therefore easy to tell how the classifier performs. The left-out process is rotated for all feature instances until they are exhausted. This is referred to as *cross-validation* and is used to assess the classifier performance in terms of true/false positives and negatives. Overall, the above falls into a class of machine learning schemes called *supervised learning*, as the classifier is informed about the correct answers.

Note that, however, the opportunity of characterising functional localisation and modularity is lost with the use of MVPA. Although region-of-interest masks may help region-specific inference, this is somewhat contrary to the purpose of MVPA – that assumes distributed neural representation.

Figure 2.11 provides a schematic demonstrating the principal pipeline for working with MVPA. This is given with an example of a hypothetical visual activation task. The task is configured in alternating blocks A and B in which subjects view images from two categories, faces and houses. The brain responses during the corresponding blocks are isolated and separated into training sets and test set. This is followed by estimating the classifier weights from the training set and cross-validating the outcome using the test set. Finally, the performance of the classifier is assessed by calculating the proportion of correct classifications, given by a percent accuracy.

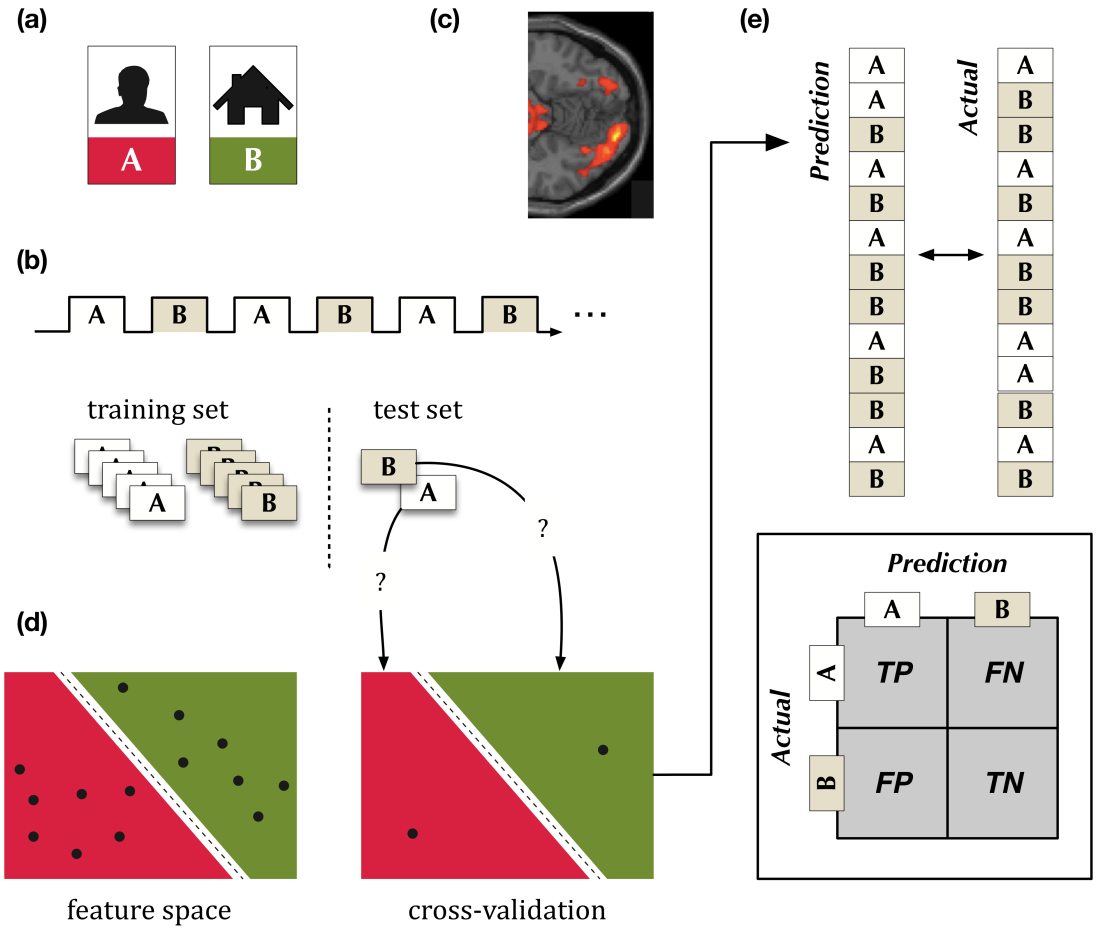

Figure 2.11 A concept of operation for multivariate pattern analysis (MVPA). This schematic demonstrates the concept behind the MVPA. ***a***, Subjects view images from two categories, faces and houses. ***b***, The task is arranged into alternating blocks during which images from each respective category are displayed. ***c***, BOLD responses are extracted as features. Note that the image does not reflect the outcome that a real task would have. ***d***, (left) The training entails finding a 'decision boundary' that separates the two sets of BOLD responses. Features are represented as vectors (points) in the feature space. (right) a 'left-out' test set is used to assess whether the classifier is able to make a correct decision based on the parameters learned from the training set. ***e***, The classifier performance is determined by the proportion of correct classification after all features are rotated to the test set. The performance is given by the true/false positives and negatives.

In the following, we briefly introduce the theoretical background of the classification algorithm, the Support Vector Machine.

2.4.3.3.2. Theory

We start by giving an example of an ideally linear separable problem. Figure 2.12a illustrates, in a 2-dimensional space, two groups of vectors (points), colour-coded in blue and orange, being separated by the dashed line. To describe the separation, let

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b \tag{2.26}$$

be the classification function. This is also known as a decision boundary or hyperplane if the vectors reside in a higher dimensional space, such as the case of fMRI data. Obviously, if $f(\boldsymbol{x}) = 0$, then $\boldsymbol{x}$ is any point along the dash line; $\boldsymbol{w}$ would then be the normal to the line and b the offset from the origin. This means that one can assign an arbitrary blue or orange point to Equation 2.26 and derive that $f(\boldsymbol{x}) \leq 1$ or $f(\boldsymbol{x}) \leq -1$. We can then assign points of respective function values to class labels -1 and 1, respectively. That is,

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b &\geq +1, \text{ when } y_i = +1 \\ \boldsymbol{w}^T\boldsymbol{x}_i + b &\leq -1, \text{ when } y_i = -1 \end{aligned} \tag{2.27}$$

where y denotes the class labels and the subscript *i* is an index to each point (Figure 2.12b).

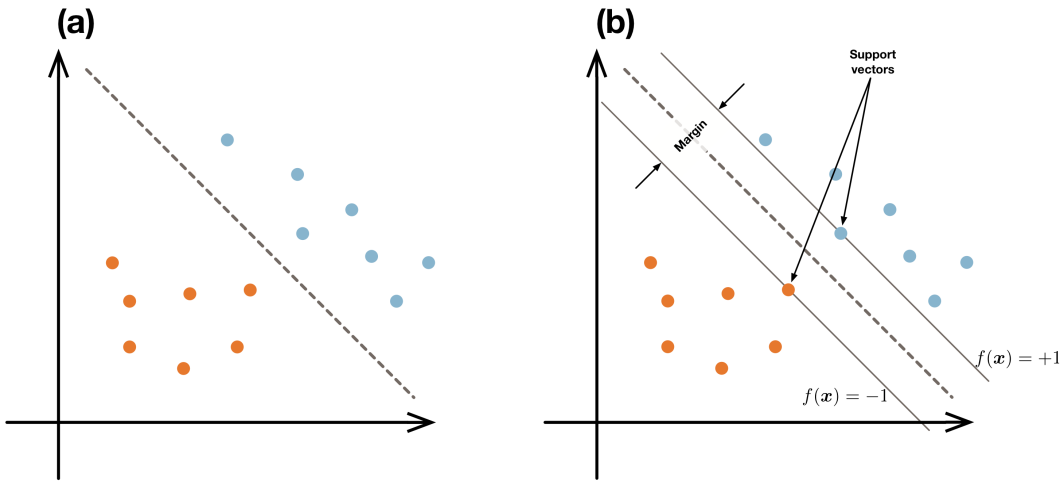

Figure 2.12 Two sets of linear separable vectors in a 2-dimentional space.

The ensuing problem pertains to the determination of unknown variables $\boldsymbol{w}$ and b. With the aim of better discriminating points from different classes, finding $\boldsymbol{w}$ and b will correspond to maximising the margin around the separating hyperplane. The margin is defined by the distance between $f(\boldsymbol{x}) = 1$ and $f(\boldsymbol{x}) = -1$ (Figure 2.12b), which induces the quantity $\frac{1}{2}||\boldsymbol{w}||^2$. We then introduce the Lagrangian and solve the dual variable $\boldsymbol{\alpha}$ to determine the maximised margin and the classification function (Bishop, 2006; Chu, 2009). In short, this is essentially the optimisation with respect to $\boldsymbol{w}$ and b by optimising the dual variable in the context of convex quadratic programming problem.

Functional margin $\hat{\gamma}$

Suppose we have determined the decision boundary $\boldsymbol{w}^T\boldsymbol{x} + b$ in the previous example. We can quantify the effectiveness of classification by measuring the distance of a data point to the decision boundary. This is given by $|\boldsymbol{w}^T\boldsymbol{x} + b|$, and can be replaced with $y(\boldsymbol{w}^T\boldsymbol{x} + b)$ because when the classification is accurate $\boldsymbol{w}^T\boldsymbol{x} + b$ and y will be of the same sign, we can then verify whether the

classification gives a desirable outcome by examining its positivity. This is the idea of the functional margin.

More formally, we define the functional margin

$$\gamma = y(\boldsymbol{w}^T \boldsymbol{x} + b) = yf(\boldsymbol{x}) \tag{2.28}$$

and let

$$\hat{\gamma} = \min \gamma_i, \ i = 1, 2, \ldots, n \tag{2.29}$$

be the minimal functional margin found in the set of available data vectors. So far, we have an interim conclusion that the functional margin is not a regularised measure of distance between the decision boundary and the data points because the size of $\hat{\gamma}$ is obviously scaled by both $\boldsymbol{w}$ and b. Next, we introduce additional constraints that lead to the definition of a geometric margin.

Geometric margin $\tilde{\gamma}$

First, we specify a point $\boldsymbol{x}$, and its corresponding point $\boldsymbol{x}_0$ on the decision boundary $f(\boldsymbol{x})$ along the orthogonal projection, with γ being the function margin. Given that $\boldsymbol{w}$ is a normal vector to the boundary, we can write down $\boldsymbol{x}$ in terms of $\boldsymbol{x}_0$ (Figure 2.13)

$$\boldsymbol{x} = \boldsymbol{x}_0 + \gamma \frac{\boldsymbol{w}}{||\boldsymbol{w}||} \tag{2.30}$$

It follows that

$$\gamma = \frac{\boldsymbol{w}^T \boldsymbol{x} + b}{||\boldsymbol{w}||} = \frac{f(\boldsymbol{x})}{||\boldsymbol{w}||} \tag{2.31}$$

Finally, the geometric margin is given by

$$\tilde{\gamma} = y\gamma = \frac{\hat{\gamma}}{||\boldsymbol{w}||} \tag{2.32}$$

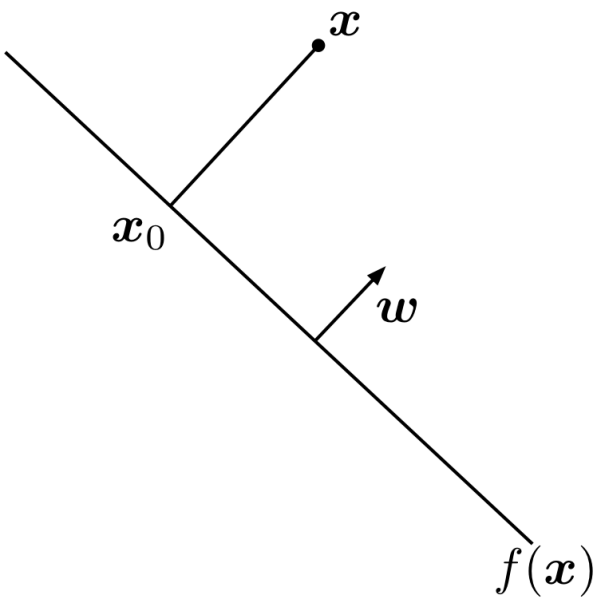

Figure 2.13 A representation of geometric margin.

Support vectors

So far, we have seen that the functional margin differs from the geometric margin by a scaling factor $||\boldsymbol{w}||$. Note that when classifying a group of data vectors, one holds greater confidence in the classifier performance when the margin is maximised. This is most effectively achieved when we deal with a non-trial vector, which lies in close proximity to the decision boundary, and is where the decision is difficult to derive.

We now have the objective

$$\max \tilde{\gamma} \tag{2.33}$$

subject to

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = \hat{\gamma}_i \geq \hat{\gamma},\ i = 1, 2, \ldots, n \tag{2.34}$$

This is equivalent to the *primal* problem

$$\begin{aligned} &\max \frac{1}{||\boldsymbol{w}||} \\ &s.t.\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1,\ i = 1, 2, \ldots, n \end{aligned} \tag{2.35}$$

Support vectors are defined by points satisfying $y(\boldsymbol{w}^T\boldsymbol{x} + b) = 1$.

Dual problem

We can rewrite Equation 2.35 into the following equivalent form

$$\begin{aligned} &\min \frac{1}{2}||\boldsymbol{w}||^2 \\ &s.t.\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1,\ i = 1, 2, \ldots, n \end{aligned} \tag{2.36}$$

We now have a quadratic objective function and a set of linear constraints – this conforms to a convex quadratic programming problem. Moreover, it turns out every *primal* problem in convex programming (Equation 2.35) has an equivalent *dual* problem (cf. to Kuhn-Tucker theorems).

Duality has two advantages: (1) the dual form tends to be easier to solve; (2) it induces a kernel function, which is crucial for non-linear classification problems. We will cover the kernel treatment shortly.

Simply put, we can now transform the original problem of the maximum margin into that of *dual variable* optimisation. In dual form, we can invoke the Lagrange function and the Lagrange multiplier, i.e., the dual variable $\boldsymbol{\alpha}$, and write

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1\right) \tag{2.37}$$

Let

$$\theta(\boldsymbol{w}) = \max_{\alpha_i > 0} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) \tag{2.38}$$

This is fairly straightforward, since the first term in Equation 2.37 should not be negative and we wish the second term, which is also non-negative, to vanish, such that the constraint is satisfied. In this case, $\theta(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||^2$ is what we set out to minimise.

Note that when point $\boldsymbol{x}_i$ is distant from the decision boundary, $\alpha_i \to 0$ since this corresponds to a trivial classification.

The objective function now becomes

$$\min_{\boldsymbol{w},b} \theta(\boldsymbol{w}) = \min_{\boldsymbol{w},b} \max_{\alpha_i > 0} \mathcal{L}(\boldsymbol{w}, b, \alpha) \tag{2.39}$$

which induces the following lower bound

$$\max_{\alpha_i > 0} \min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b, \alpha) \leq \min_{\boldsymbol{w},b} \max_{\alpha_i > 0} \mathcal{L}(\boldsymbol{w}, b, \alpha) \tag{2.40}$$

The intuition here is that the maximum of minima is equal to or smaller than the minimum of maxima.

In the context of convex optimisation, the equality in Equation 2.40 implies that a saddle point exists (Slater's condition; Slater, 2013). Next, we solve the dual problem in two steps. First, minimise $\mathcal{L}$ with respect to $\boldsymbol{w}$ and b to get

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \implies \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \tag{2.41}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \implies \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2.42}$$

Accordingly, Equation 2.37 now reads

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{2.43}$$

Finally, maximise Equation 2.43 to obtain (Platt, 1998)

$$\boldsymbol{\alpha}^T \boldsymbol{y} = 0 \tag{2.44}$$

Soft margin

So far we have only considered the case of perfect linear separation. It is possible to introduce a 'soft margin' to approximate linear separation. This gives an almost identical quadratic programming problem. Specifically, the constraint $\alpha_i \geq 0$ is changed to $0 \leq \alpha_i \leq \lambda$, where λ determines the degree of 'softness', which incurs a penalty on unseparated points. It follows that any point corresponding to $0 \leq \alpha_i \leq \lambda$ is now considered a support vector. Changing the value of λ changes the behaviour of the classifier: with $\lambda \to 0$, the classifier trades accuracy for more robust prediction on new data. Whereas, with $\lambda \to \infty$, the opposite is the case.

Kernel methods

In previous section, we mentioned classification of linear separable data. An issue remains for those are not linearly separable. The adoption of the kernel method circumvents this limitation. In a classification task, the kernel method refers to the use of 'kernels' as inputs, instead of data features. Intuitively, a kernel is a 'similarity matrix' in which the pair-wise similarity of all data points is encoded. Creating the similarity matrix is equivalent to mapping data points in the original space into a higher dimensional space, in which linear separation is possible. Figure 2.14 gives a schematic illustrating the concept of the kernel method.

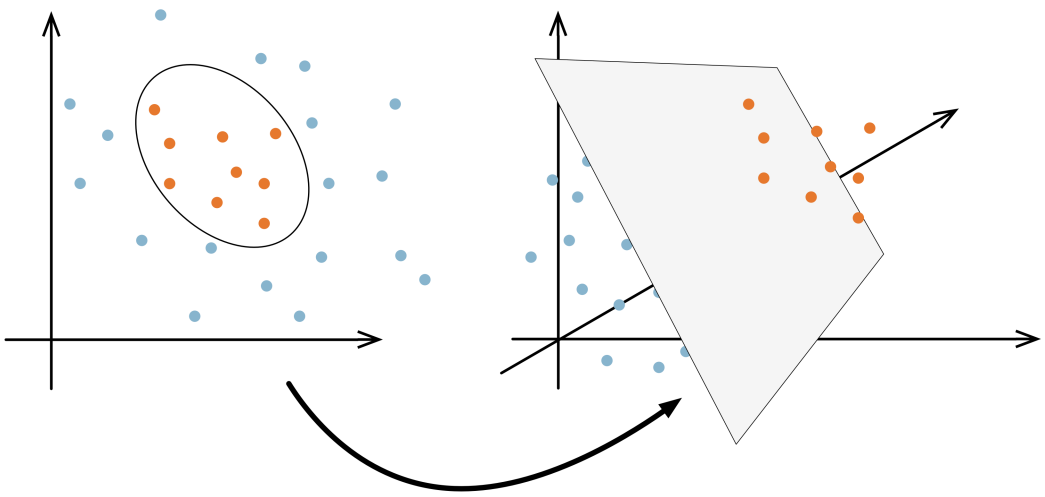

Figure 2.14 Concept of the kernel method. The kernel method can be conceptualised as a mapping of data points into a higher dimensional space. (left) In the original space, separating data points of the two colours requires the classification function to be non-linear which can be difficult to derive. (right) After mapping data into higher dimension, linear separation can be achieved by finding a hyperplane that falls into the framework described in earlier sections.

2.4.3.4. Dynamic Causal Modelling

Effective connectivity quantifies the influence exerted by one node on another in a (neuronal) network. It offers a perspective on how distributed cortical regions interact as an integrative ensemble. Dynamic Causal Modelling (DCM), which analyses effective connectivity, is a framework for the identification of neural networks in the brain that treats the networks as nonlinear input-state-output systems. In setting up a DCM one can estimate: (1) parameters that mediate the driving influence of exogenous or experimental inputs on brain states, (2) parameters that mediate endogenous coupling among neuronal states, and (3) parameters that allow the inputs to modulate that coupling. Issues concerning selection among alternative models naturally arise in DCM analyses. Bayesian model selection (BMS) is a statistical procedure for computing how probable one model is in relation to another. This section presents the motivation and procedures for DCM of evoked brain

responses – as well as the theoretical and operational details on which BMS rests. We describe procedures for parameter, model and family-level inference in the context of data analysis from a group of subjects.

2.4.3.4.1. Background

This section is about Dynamic Causal Modelling (DCM) of the interactions between functionally elicited brain responses, and its applications in neuroimaging. DCM was invented to test hypotheses about neural systems – as opposed to regionally specific correlates – and necessitates a predefined set of plausible structural models, commonly referred to as *model space*. In other words, each DCM embodies a specific hypothesis pertaining to how a neural system interacts and produces observed responses. To allow a hypothetical structure or network model to predict observed responses it is crucial to understand and model how changes in neuronal states are manifest as observed haemodynamic responses.

In what follows we first focus on the conceptual and operational constructs of DCM as a biophysically realistic forward model – as exemplified by an up-to-date implementation of DCM. Then, we turn to Bayesian model selection of DCMs, where models can be considered as fixed effects (e.g., as low-level neurophysiological mechanisms that are conserved over subjects) or random effects (e.g., as high-level cognitive processes that are implemented with different strategies or networks) in the population. Finally, we consider inference about the parameters of a model; for example, how a connection from one region to another is changed by experimental context. We describe how such inferences can be made for the case of single models, and for models derived from averaging over different models or subjects in a group.

2.4.3.4.2. Forward model for fMRI

This section presents the essential operational aspects of Dynamic Causal Modelling (DCM). The theoretical basis of DCM rests on dynamical systems theory and Bayesian statistics. The primary objective of DCM appeals to nonlinear system identification in which a set of differential equations is formulated to capture the (hidden) mechanistic structure of a neuronal system of interest. These equations specify how constituent nodes (or neuronal 'states') of a system exhibit time-varying and causal relations with one another. Specifically, this system is acted upon by exogenous inputs (e.g., visual stimuli) that engender regional neuronal activity that, in turn, generates outputs (e.g., BOLD signals). This necessarily requires DCM to be hierarchical – where a two-layered forward model translates neuronal states into haemodynamic states, and measured BOLD responses. The haemodynamic states are modelled in a regionally independent fashion. Neuronal dynamics emerge from designed experimental perturbations and directed interactions among regions. Specification of effective connectivity within the network of coupled nodes or regions depends on three sets of (neuronal) parameters: (a) parameters that mediate endogenous coupling among the states, (b) parameters that allow exogenous inputs to modulate the coupling, and (c) parameters that mediate the influences of exogenous inputs on the states. These parameters are embedded in a dynamic causal model that is motivated by a particular hypothesis about network structure, and can be estimated by fitting the ensuing forward model to observed data, using standard Bayesian procedures. This model inversion procedure provides posterior estimates of the parameters and an estimate of the model evidence, in terms of probability distributions. Critically, prior densities over parameters constrain parameter estimates to dynamical or physiologically realistic ranges. By default, 'shrinkage'

priors are chosen for endogenous and modulatory parameters, while priors on haemodynamic parameters are derived from previous empirical studies.

In what follows, we first review the neuronal state equations, haemodynamic state equations and the priors over model parameters. We then briefly consider the standard Bayesian scheme used for model inversion. We will briefly review nonlinear DCMs, where one region can modulate the connectivity between another pair. Some of the more recent DCM developments are considered in the closing section.

Notation

Variables in bold face refer to matrices and vectors. States are functions of time, although the dependency on time *t* is not made explicit. The vector $\mathbf{z} = [z_1, z_2, \cdots]^T$ denotes any number of neuronal states of interest. A neuronal state, say z_1, can be taken as the collective dynamics of neuronal activity in the first region. The remaining state variables are biophysical states $\{\mathbf{s}, \mathbf{f}, \mathbf{v}, \mathbf{q}\}$ that model haemodynamics. These haemodynamic states refer to (1) vasodilatory signal, (2) blood inflow, (3) blood volume, and (4) deoxyhaemoglobin content, respectively. The vector $\mathbf{x} = \{\mathbf{z}, \mathbf{s}, \mathbf{f}, \mathbf{v}, \mathbf{q}\}$ denotes all the hidden (neuronal and haemodynamic) states collectively. The vector $\mathbf{u} = [u_1, u_2, \cdots]^T$ denotes any number of exogenous inputs that are specified experimentally. Elements of $\mathbf{u}$ can be, for example, spike or boxcar functions of time that represent the onset/offset of task stimuli or contextual manipulations. Alternatively, exogenous inputs can also be motivated by a neurocomputational or model-based approach (O'Doherty, Hampton, & Kim, 2007). θ denotes the collection of model parameters, including coupling parameters and

haemodynamic parameters. Different model structures are indexed by *m*, i.e., differences may exist in endogenous, modulatory, or exogenous connections.

Neurodynamics

Assuming any number of neuronal states **z** and any number of exogenous inputs, one can posit a model of the general form

$$\frac{dz}{dt} = F(\mathbf{z}, \mathbf{u}, \theta) \tag{2.45}$$

where *F* is some nonlinear function describing the neurophysiological influences exerted by inputs $\mathbf{u}$ and the activity in all brain regions on the evolution of the neuronal states. A bilinear approximation provides a natural and useful re-parameterisation in terms of coupling parameters.

$$\begin{aligned} \frac{dz}{dt} &\approx \mathbf{Az} + \sum_j u_j \mathbf{B}^{(j)} + \mathbf{Cu} \\ &= \left(\mathbf{A} + \sum_j u_j \mathbf{B}^{(j)} \right) \mathbf{z} + \mathbf{Cu} \end{aligned} \tag{2.46}$$

The (effective) connectivity matrix $\mathbf{A}$ represents the first-order coupling among the regions in the absence of inputs. This can be thought of as the endogenous coupling in the absence of experimental perturbations. Note that the state, which is perturbed, depends on the experimental design (e.g. baseline or control state) and therefore the endogenous coupling is specific to each experiment. The matrices $\mathbf{B}^{(j)}$

are the change in endogenous coupling induced by the *j*th input (Figure 2.15a). Finally, the matrix **C** encodes the exogenous (driving) influences of inputs on neuronal activity. The parameters θ^c are the coupling parameter matrices we wish to estimate and define the functional architecture and interactions among brain regions at a neuronal level. Note that the units of coupling are per unit time (Hz) and therefore correspond to rates. Because we are in a dynamical setting, a strong connection means an influence that is expressed quickly or with a small time constant. It is useful to appreciate this when interpreting estimates and thresholds quantitatively.

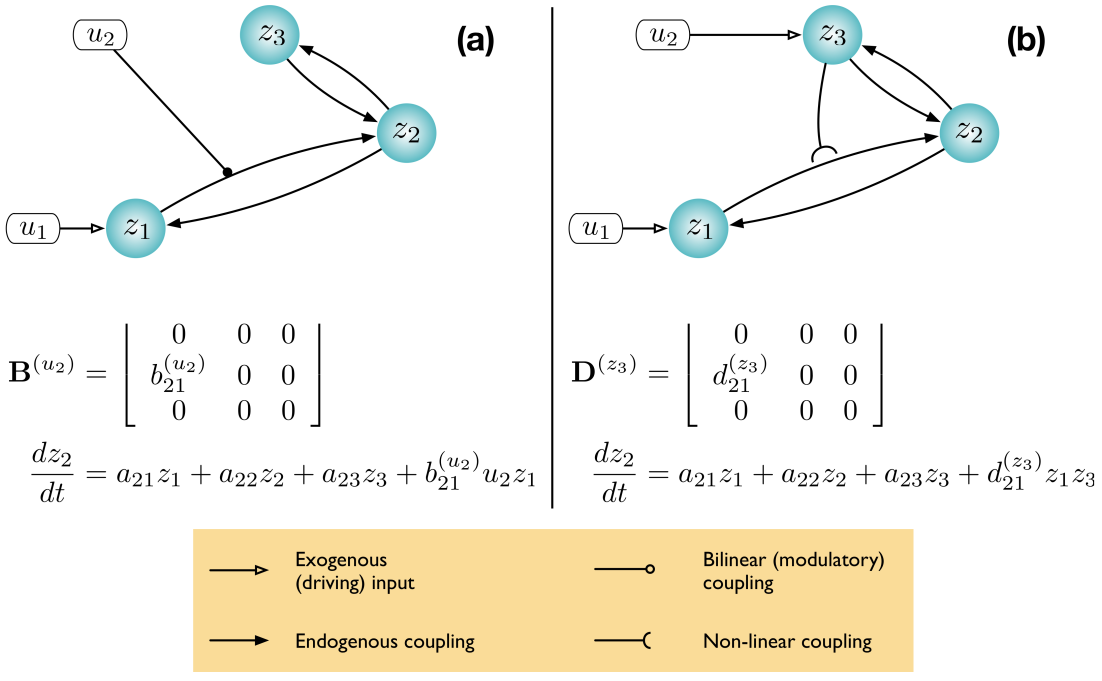

Figure 2.15 Modulatory effects in dynamic causal models. This diagram illustrates two types of 'modulatory' effects – the bilinear (**a**) and the nonlinear (**b**) modulations. The target of modulation is the z_1-to-z_2 coupling. The difference between the neuronal state equations for z_2 is made explicit in the respective panel (see the last term). Specifically, the bilinear model allows and exogenous experimental manipulation (u_2) to induce connectivity change. The nonlinear model, on the other hand, uses the neuronal states (z_3, instead of u_2) as the source of modulation.

Neuronal activity in each region cause changes in volume and deoxyhaemoglobin which engender the observed BOLD response $\mathbf{y}$ as described below. The ensuing haemodynamic component of the model is specific to BOLD-fMRI and would be replaced by appropriate forward models for other imaging modalities, such as EEG or MEG.

The neuronal dynamics in Equation (2.46) operate around a stable fixed point $z = 0$ (strictly speaking, this will only be the case for certain ranges of parameter values – see (Friston, Harrison, & Penny, 2003). This means that, in the absence of exogenous perturbations, the neuronal activity and consequently the fMRI activity will be zero. Briefly, a neuronal state in DCM predicts nothing but a flat line if it is not experimentally perturbed, directly or indirectly [but see (B. Li et al., 2011a)]. This is because DCM for fMRI is based on a dynamic system with a fixed point attractor.

Nonlinear DCM

Nonlinear DCM (Stephan et al., 2008) introduces a parametric matrix $\mathbf{D}$ that allows neuronal activity in one region to change the connectivity between other regions (Figure 2.15b). This is in contrast to bilinear dynamics (Equation 2.46) in which, perhaps unrealistically, effective connectivity can be changed via 'modulatory inputs'. The nonlinear DCM is given by the following equation.

$$\frac{dz}{dt} = \left(\mathbf{A} + \sum_j u_j \mathbf{B}^{(j)} + \sum_j z_i \mathbf{D}^{(i)} \right) \mathbf{z} + \mathbf{Cu} \qquad (2.47)$$

The motivation for this extension is to address 'neuronal gain control' between two neuronal states that are gated by other states (Stephan et al., 2008). The approach also models the neuronal origin of modulatory influences such as 'short-term synaptic plasticity' (Stephan et al., 2008). Applications based on nonlinear DCM can be found in recent works by den Ouden et al. (2010), Dessilles et al. (2011), and Neufang et al. (2011)

Haemodynamics

Neuronal activity is linked to fMRI signals via an extended Balloon Model (Buxton, Uludağ, Dubowitz, & Liu, 2004; Buxton, Wong, & Frank, 1998) and BOLD signal model (Stephan, Weiskopf, Drysdale, Robinson, & Friston, 2007b). The haemodynamic model specifies how changes in neuronal activity give rise to changes in blood oxygenation that is measured with fMRI. For each region i, neuronal activities are translated into BOLD signals via the interactions between the neuronal state z_i and haemodynamic state variables: the vasodilatory signal, the flow induced, changes in volume, and changes in deoxyhaemoglobin. The observed BOLD signals are produced by a nonlinear model that integrates over the states, $\mathbf{y} = g(\mathbf{v}, \mathbf{q})$, where the evolution of $\mathbf{v}$ and $\mathbf{q}$ over time depends on self-regulatory feedback as well as $\mathbf{s}$ and $\mathbf{f}$ (cf. to Figure 3 in (Friston et al., 2003)). The equations for the haemodynamics are described in detail elsewhere (Buxton et al., 1998; Friston et al., 2000; Grubb, Raichle, Eichling, & Ter-Pogossian, 1974; Mandeville et al., 1999).

Priors

Two classes of prior densities are used in DCM; they are placed over coupling and haemodynamic parameters $\theta = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \theta^h\}$. DCM uses 'shrinkage priors'

over coupling parameters. These are zero-mean Gaussian priors with a variance that is chosen to reflect realistic ranges of effective connectivity seen in fMRI studies. These shrinkage priors move the posterior estimates toward zero, especially when the likelihood has a less precise distribution. For example, the posterior expectation will 'shrink' to its prior expectation given a likelihood function with a very large variance. However, a likelihood that has high precision (inverse variance) forces the posterior to deviate from zero. Prior variances can also be chosen to reflect anatomical knowledge; e.g., probabilistic tractography (Stephan, Tittgemeyer, Knösche, Moran, & Friston, 2009b). Haemodynamic priors in DCM reflect empirical knowledge about blood flow and oxygenation dynamics in the brain (Buxton et al., 1998; 2004). The prior densities of the five haemodynamic parameters $\theta^h = \{\kappa, \gamma, \tau, \alpha, \rho\}$ that mediate the interactions among these states are based on empirical measures [see Equation (3) and Table 1 in (Friston et al., 2003)]. These priors have since been updated in light of more recent data (Penny, 2012).

Model fitting

DCMs are fitted to data using the Variational Laplace (VL) algorithm described in (Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007). Simply put, this is an iterative algorithm, which approximates the posterior distribution over parameters with a Gaussian distribution. The parameters of this distribution are updated so as to minimise the distance between the approximate and true posterior, quantified by the Kullback-Leibler divergence – a distance measure between probability densities (MacKay, 2003). The VL algorithm provides estimates of two quantities. The first is the posterior density over model parameters $p(\theta|m, \mathbf{y})$ that can be used to make

inferences about model parameters θ. The second is the probability of the data given the model, otherwise known as the model evidence $p(\mathbf{y}|m)$.

Model evidence

In general, model evidence is not straightforward to compute, since this computation involves integrating out the dependence on model parameters:

$$p(\mathbf{y}|m) = \int p(\mathbf{y}|\theta, m)p(\theta|m)d\theta \tag{2.47}$$

Therefore an approximation to the model evidence is required. DCM uses the free energy approximation to the model evidence provided by the VL algorithm. The model evidence, and the VL approximation to it, naturally embodies the accuracy-complexity trade-off that is the hallmark of a good model (Pitt & Myung, 2002). The VL algorithm uses a 'free energy' approximation to the model evidence which has been shown to be superior to other information theoretic criteria (Penny, 2012). By comparing the evidence of one model relative to another, a decision can be made as to which is the more veridical one (Friston et al., 2008; Penny, Kiebel, & Friston, 2003).

2.4.3.4.3. Model inference

The model inference problem arises naturally in nearly every scientific discipline (D. R. Anderson, 2008). Most importantly, it requires a well thought-out specification of the model space – that is, the set of hypotheses that are to be considered. In the simplest case, one will have a null model and an alternative model and inference can proceed using Bayes factors. Once the evidence has been computed, a model (m_i) can be compared to another (m_j) by means of the Bayes factor (Raftery, 1995)

$$\mathrm{BF}_{ij} = \frac{p(\mathbf{y}|m_i)}{p(\mathbf{y}|m_j)} \tag{2.48}$$

A Bayes factor of 20 (or log Bayes Factor of 3) corresponds to a posterior model probability of 0.95, and is used as the standard decision threshold (Penny, Stephan, Mechelli, & Friston, 2004).

More generally, one might be able to constrain the space of models to a small number. Model inference can then proceed using the posterior distribution over models, which can be obtained from Bayes rule

$$p(m|\mathbf{y}) = \frac{p(\mathbf{y}|m)p(m)}{p(\mathbf{y})} \tag{2.49}$$

The prior distribution over models, $p(m)$, is usually chosen to be a uniform distribution. In larger model spaces it becomes increasingly unlikely that high posterior probability mass will be attributed to any single model. This is because there are likely to be many similar models in large model spaces – and they will share probability mass. This is known as *dilution* and can be dealt with by combining models into families (Penny et al., 2010). Models in the same family share the same characteristics; e.g., nonlinearity, the same driving region or the same modulatory connection.

Group Inference

Next we turn to the topic of model inference for data from a group of subjects. There are two approaches. Fixed effect analysis (FFX) (Stephan, Marshall, Penny, Friston, & Fink, 2007a) assumes that all subjects use the same model, whereas random effects (RFX) analysis assumes different subjects use different models (Stephan, Penny, Daunizeau, Moran, & Friston, 2009a).

Fixed effect analysis

In FFX analysis, a Group Bayes Factor (GBF) is computed by multiplying the Bayes Factors from the group of subjects. As is considered in Stephan et al. (Stephan et al., 2009a), the GBF approach implicitly assumes that every subject uses the same model (Figure 2.16a). This assumption is warranted when studying a basic physiological mechanism that is unlikely to vary across subjects, such as the role of forward and backward connections in visual processing (C. C. Chen, Henson, Stephan, Kilner, & Friston, 2009). Li et al. (2011b), for example, studied the motor network by perturbing it with transcranial magnetic stimulation. With clearly defined timing in eliciting network responses, and the homogeneity of motor circuitry over subjects, GBF was entirely suitable. In other words, inferences relying on GBF will – by default – neglect group heterogeneity, whereas functional tasks engaging higher cognitive processes may show group heterogeneity due to individual differences in cognitive strategies. Moreover, GBF is susceptible to outliers – toward which the inference may be heavily biased.

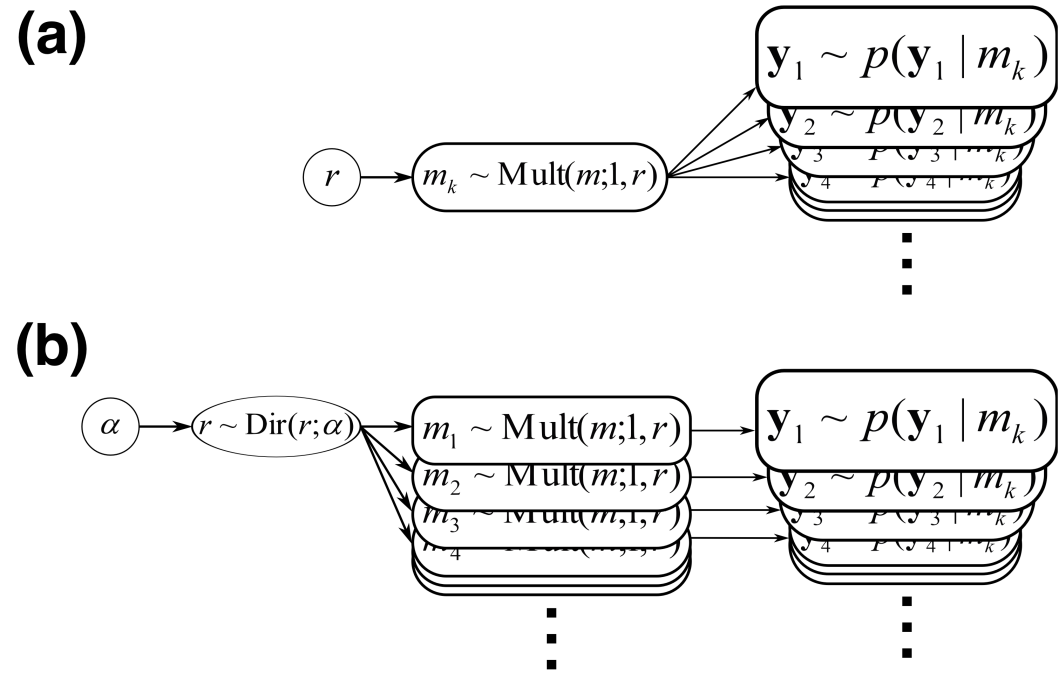

Figure 2.16 Generative models for multi-subject data. The fixed effect model (**a**) assumes that subject-specific data are generated by particular model. The random-effect model (**b**) suggests that the data-generating models (e.g., m_1-m_4) are treated as random variables. As a consequence, the random-effect model allows different causal structures across subjects (α, parameters of the Dirichlet distribution, or model 'occurrence' in the population level; r, parameters of the multinomial distribution, or the model).

Random effect analysis

An alternative procedure for group level model inference allows for the possibility that different subjects use different models (Figure 2.16b). This is more realistic when investigating pathophysiological mechanisms in a spectrum disorder or when dealing with cognitive tasks that can be performed with different strategies. In random effect analyses one makes inferences based on the posterior estimates of the model frequencies. For the *k*th model, r_k denotes the frequency with which it is used in the population. Inferences are therefore based on the posterior density $p(r_k|y)$. This can be computed by combining the table of model evidences with an uninformative prior, $p(r_k)$, using a Bayesian inversion scheme. Such an inversion

can be implemented using a variational approach (Stephan et al., 2009a) or Gibbs sampling (Penny et al., 2010). One should note that the variational approach is only valid for small numbers of models (small in relation to the number of subjects, e.g. 10 or so models for 20 or so subjects) and that Gibbs sampling is now the standard approach. Both algorithms produce approximations to the posterior density on which subsequent RFX model comparisons are based. One can report the result of RFX model comparison using (1) the posterior expected probability of observing the *k*th model or (2) the *exceedance probability* which reflects the belief that one model is more likely than any other in the model space.

Passamonti and colleagues provide an example of this approach: they investigated the neural mechanisms of emotion regulation, and assumed the underlying cognitive processes would vary across the group (Passamonti et al., 2012). Thus, their adoption of random-effects BMS procedure was appropriate. Another example of model-level inference relates to different neural mechanisms giving rise to distinct synaesthetic experiences that can be explained in terms of alterations in the visual processing hierarchy (van Leeuwen, Ouden, & Hagoort, 2011). Generally, RFX is more conservative and is robust to outlying subjects.

Family inference

Family inference (Penny et al., 2010) can proceed using either an FFX or RFX approach. Passamonti et al. (2012) employed a tripartite model space ('meta-family') where numbers of driving exogenous inputs varied across each subspace. The variation of the location of driving inputs further constituted respective families within each meta-family. The authors performed BMS across all the meta-families, regardless of any other difference among the models considered – to establish how many inputs were needed. Models within the winning meta-family were further

compared to determine the location of driving inputs (cf. Figure S1 in Passamonti et al., 2012). To summarise, model-level or family-level inference is appropriate when the hypothesis of interest can be answered in terms of overall model structure (i.e., the existence of multiple sets of parameters) rather than any specific model parameter.

2.4.3.4.4. Parameter inference

Finally, we address inferences made on the basis of connectivity parameters in the context of a group analysis. Assessing the statistical significance of posterior estimates of individual model parameters is usually the last step in a DCM application (Almeida et al., 2009; 2011; Bányai, Diwadkar, & Erdi, 2011; Deserno, Sterzer, Wüstenberg, Heinz, & Schlagenhauf, 2012; X. Li et al., 2011b; Neufang et al., 2011; Passamonti et al., 2012; Schlösser et al., 2010; van Leeuwen et al., 2011).

If random effects on parameters are assumed in the population, a classical approach can be applied (e.g., t-test or ANOVA). Conceptually, this conforms to the classical (frequentist) summary statistics approach – using subject specific MAP (maximum a posteriori) point estimates of the coupling parameters. This application is used widely (Bányai et al., 2011; Deserno et al., 2012; Diwadkar et al., 2012; Neufang et al., 2011; Schlösser et al., 2010) for identifying significant effects between different groups.

The summary statistic RFX approach is readily applied to the MAP parameter estimates for selected parameters from each subject. However, if one has multiple models per subject, then one also needs to average over models (accounting for the possibility that different subjects use different models). This can be implemented using Bayesian Model Averaging (see below) within subject (over models). The

resulting parameter estimates can then enter as summary statistics into a classical RFX analysis (e.g., t-test or ANOVA).

If fixed effects of parameters are assumed in the population then one can compute a 'group' model by averaging over the models from subjects in that group. This is a FFX approach and can be implemented using Bayesian Parameter Averaging, as described in the next section and in Kasess et al. (2010).

Bayesian Parameter Averaging

Bayesian parameter averaging (BPA) has multiple uses. Generally, it is a procedure to combine parameter estimates from the same model of multiple datasets to produce parameter estimates from the entire dataset. The data could come from DCMs fitted to different sessions from the same subject. Or, most often, they could be the same model structure fitted to data from multiple subjects (Kasess et al., 2010). For example, van Leeuwen et al. (2011) summarised model parameters using BPA and found that V4 activation in synaesthetes was dependent on top-down – rather than bottom up inputs – as a function of whether they were a 'projector' or an 'associator'. The common feature of all these applications is that variability over the model fitting is not taken into account. That is, the averaging procedure corresponds to a FFX analysis (because only one model is used). Mathematically, the posterior means from each model to be combined are weighted by their relative posterior precisions; this means estimates with higher precision are given greater weight.

Low-level neurophysiological processing can be considered as fixed effects since they are unlikely to vary across populations, e.g., Desseilles et al. (2011) and van Leeuwen et al. (2011) both interrogated selective colour vision processing mechanisms. If this is the case, Bayesian parameter averaging (BPA) can be used to

summarise individual posterior densities of an identical optimal model across the entire group (Bányai et al., 2011; Desseilles et al., 2011; van Leeuwen et al., 2011).

Bayesian Model Averaging

Another approach to summarise parameters as random effects is through Bayesian Model Averaging (BMA). In this sort of averaging there are multiple models of the same data (as opposed to a single model of multiple datasets). BMA is usually performed within a model family, where no model within the family clearly outperforms all others. It can also be applied to the whole model space. As such, parameter inference no longer depends on a particular model selection. For instance, Deserno et al. (2012) employed BMA in a DCM study of working memory in schizophrenia. They first performed a family-level inference and found that the family of models with modulation of backwards connections from prefrontal to parietal cortex was the clear winner. They then performed BMA for each subject, entered the averaged parameters as summary statistics into a two-sample t-test and found reduced connectivity in the schizophrenic group (cf. Figure 3, 4, and Table 2 in Deserno et al., 2012).

The relationships among the various model and parameter inference procedures perhaps seems complicated on a first reading, but are clearly laid out in, for example, Figure 1 of (Stephan et al., 2010). Once one appreciates the simplicity of pooling evidence for different models and parameters, Bayesian model and parameter averaging can be a powerful approach to testing specific mechanistic hypotheses.

2.4.3.4.5. Conclusions

This section has described the basic principles of DCM – with a focus on how to implement parameter, model and family-level inferences in analyses of data from

groups of subjects. I have not elaborated on some of the more recent developments in DCM. These include the use of two-state DCMs, in which neuronal activity is represented by separate populations of excitatory and inhibitory cells, and stochastic DCMs in which neuronal activity is modelled via a combination of deterministic flow and stochastic innovations – thus better describing the interaction between exogenous and endogenous brain activity. Moreover, there is a library of DCMs for the study of effective connectivity based on EEG, MEG and LFP data (Litvak et al., 2011) that may usefully complement the use of dynamic causal modelling in fMRI.

Chapter 3. The functional anatomy of anticipatory set and memory updating

To behave adaptively, an organism must be able to balance the accurate maintenance of information currently stored in working memory with the ability to update that information when the context changes. This trade-off between fidelity and flexibility is likely to depend upon the anticipated stability of information retained in working memory – and thus the likelihood that updating will be necessary. To address the neurobiological basis of this anticipatory optimisation, we acquired functional magnetic resonance imaging (fMRI) data while subjects performed a modified delayed response task. The modification used cues that predicted memory updating – with high or low probability – followed by a contingent updating or maintenance event. This enabled us to compare behaviour and neuronal activity during conditions in which updating was anticipated with high and low probability, and measure responses to expected and unexpected memory updating. Based on the known importance of the dopaminergic system for cognitive flexibility and working memory updating, we hypothesised that differences in anticipatory set would be manifest in the dopaminergic midbrain and striatum. Consistent with our predictions, we identified sustained activation in the dopaminergic midbrain and the striatum, associated with anticipations of high versus low updating probability. We also found that this anticipatory factor affected neural responses to subsequent updating processes, which exhibited suppressed, rather than elevated, midbrain and striatal activity. Our study thus addresses – for the first time – an important and hitherto understudied aspect of working memory.

3.1. Introduction

Working memory involves actively maintaining and manipulating mental representations in the absence of external stimuli (Baddeley, 1992; 2012). Maintenance and manipulation are often cast in terms of stability and flexibility – as two reciprocal aspects of working memory. Generally speaking, manipulation is studied in the context of memory updating (Veltman, Rombouts, & Dolan, 2003) in which the fronto-striatal circuitry is strongly implicated (Marklund, 2009). Updating requires the encoding of new information and adaptively replacing old information. Crucially, balancing maintenance and manipulation involves trading off flexibility against the robustness of representations, but little is known about how this is achieved.

The striatum, given its role in action selection (Mink, 1996) and the topographically parallel infrastructure (Alexander et al., 1986), seems to enable non-motor cognitive function such as flexible updating. In contrast, the dorsolateral prefrontal cortex, which is an integral part of prefronto-striatal functioning, has the neuronal architecture for maintenance of working memory representations (Goldman-Rakic, 1995). Neurotoxin administration has supported this implicit functional segregation (Crofts et al., 2001). Yet, simply having two functionally segregated systems cannot explain how maintained memories are updated without understanding how their functions are integrated. Dopamine, which exhibits tonic and phasic modes of discharge, is a promising candidate for nuancing the balance between stability and flexibility – given that it exerts antagonistic influences in the two systems by modulating neuronal excitability through dissociable distributions of (D1/D2) receptors (Camps et al., 1989). Tonic dopamine tends to stimulate (high-

affinity) D2 receptors, whereas phasic dopamine generally increases (low-affinity) D1 stimulation level (Dreyer, Herrik, Berg, & Hounsgaard, 2010; Goto, Otani, & Grace, 2007). Prevalent theories have addressed the phasic mode of dopamine in contributing to updating (M. J. Frank et al., 2001; R. C. O'Reilly & Frank, 2006). However, there is little evidence on how tonic dopamine modulates the updating of working memory representations.

Current approaches to memory updating generally focus on the comparison between non-updating (maintenance) and (selective/total) updating (Lenartowicz, Escobedo-Quiroz, & Cohen, 2010; Podell et al., 2012). Existing evidence tends to sit well with theoretical predictions (M. J. Frank et al., 2001; Hazy et al., 2007; R. C. O'Reilly & Frank, 2006), in which the fronto-striatal network controls access to working memory (McNab & Klingberg, 2008) – with phasic dopamine acting a gating signal (D'Ardenne et al., 2012; Murty et al., 2011). Although this provides a compelling mechanistic explanation of updating, it does not address a crucial aspect of adaptive behaviour and brain function: how the brain balances the maintenance of beliefs about the world with the assimilation of new information (Friston & Stephan, 2007; Rao & Ballard, 1999), a balance that is likely to depend upon the anticipated changeability or volatility of environmental cues (Behrens, Woolrich, Walton, & Rushworth, 2007). Manipulating the anticipated likelihood of updating may thus provide a new insight into the functional anatomy of memory updating. Tonic dopamine has been associated with uncertainty on both empirical (Fiorillo, Tobler, & Schultz, 2003) and theoretical (Friston et al., 2012) grounds, suggesting a possible augmentation of the phasic updating model to include a role for tonic dopamine in encoding the precision of – or confidence in – the task-relevance of current representations.

To characterise the functional anatomy of updating in working memory, we used predictive cues to manipulate subjects' anticipatory set or beliefs about the probability that working memory updating would be called upon. Our principal hypothesis was that anticipation about imminent updating would increase cognitive flexibility via modulations of tonic activity in the dopaminergic system and would thereby interact with the subsequent updating *per se*.

3.2. Methods

3.2.1. Pre-processing

Imaging data were analysed using SPM 12 (Statistical Parametric Mapping; Wellcome Trust Centre for Neuroimaging, London, UK). Preprocessing of functional images included correction for geometric distortion using field maps (Hutton et al., 2002; Jezzard & Balaban, 1995), realignment via affine registration to correct for head movement, slice timing correction, coregistration with respect to anatomical images, normalisation to MNI space based on the anatomical normalisation parameters, interpolation to voxel size of 2 x 2 x 2 mm^3, and smoothing with a Gaussian kernel of 4 mm FWHM (full-width at half-maximum).

3.2.2. Mass-univariate analysis

Pre-processed data were entered into the general linear model, which was subsequently inverted to obtain the parameter estimates of interests. The design matrix used in the first (within subject) level analysis included eight task-related regressors: maintenance set (MAI-set), updating set (UPD-set), updating (UPD), maintenance (MAI), omission (surprising maintenance), deviation (surprising

updating), non-specific task effects (NS), and set-switching. MAI-set and UPD-set were models with 6 s boxcar functions, extending from the onset of cue stimuli to the offset of the retention period. These regressors modelled the sustained cue-specific anticipatory set-related activity, during which subjects prepared for the forthcoming action array. The set-switching regressor modelled transient responses at the cue onset, which can be associated with the effect of trial transition. UPD and MAI entered the GLM for all trial types. Omission and deviation modelled the interaction between anticipatory set and action (i.e., UPD-set/MAI and MAI-set/UPD trials), where invalid outcomes violated anticipatory states. Surprises were modelled as transient responses at the onsets of action arrays under the MCU and UCM condition. Using non-specific UPD and MAI regressors, together with regressors encoding omission and deviation is equivalent to comparing valid/invalid trials. Finally, transient responses to encoding, probing, and all cues from error trials were modelled by a NS regressor as a nuisance effect. The eight regressors were convolved with a canonical haemodynamic response function to produce haemodynamic regressors for the GLM. Other effects of no interest, including head motion and low-level physiological variations, were modelled with an additional 20 regressors. Head motion was described using three translation (*x*, *y*, and *z* directions) and three rotations (pitch, roll, and yaw) derived from the realignment procedure. The physiological nuisance effects comprised six cardiac regressors, six respiratory regressors, and two regressors for hear rate change and change in respiratory volume (Hutton et al., 2011).

3.2.3. Region of interest analysis

Empirical and theoretical accounts of the 'gating' hypothesis implicate the dopaminergic midbrain and the striatum in memory updating (D'Ardenne et al., 2012; M. J. Frank et al., 2001; Murty et al., 2011; R. C. O'Reilly & Frank, 2006). We hypothesised that activity in these regions would also be modulated by anticipatory set. We therefore defined regions of interest (ROIs) in the substantia nigra/ventral tegmental area (SN/VTA), the striatum, and the DLPFC and analysed responses within these regions across each level of anticipatory set and action.

Anatomically informed functional ROIs were created for the midbrain and the striatum in two steps: (1) after the SN/VTA region was identified in the mean normalised magnetisation transfer image averaged across subjects (Fitzgerald, Friston, & Dolan, 2012; Helms, Draganski, Frackowiak, Ashburner, & Weiskopf, 2009), we manually traced the SN/VTA to create a (preliminary) anatomical ROI; (2) the anatomical ROI was then masked with the thresholded activation map of set (main effect of UPD-set and MAI-set, uncorrected $p = 0.005$). A similar thresholding procedure was taken for the main effect of action using the preliminary ROI. A small-volume correction was performed on tests for responses within the ROI search volume. The main effect of set (or action) was specified with appropriate contrasts averaging over UPD-set and MAI-set (or for UPD and MAI in the case of action). The resulting contrast images were then entered into a second (between-subject) level analysis using a one-sample *t* test. Importantly, these localising (ROI defining) contrasts are orthogonal to the differential effects of set (UPD-set > MAI-set) and action (UPD > MAI) that were subsequently tested using one-sample *t* tests.

The functional SN/VTA ROI for set consisted of 240 voxels [p = 0.001, cluster false discovery rate (FDR); Figure 3.1a]. This ROI was used to summarise UPD-set

and MAI-set effects in terms of their principal eigenvariates. A one-sample *t* test was performed to test for updating versus maintenance effects of anticipatory set on these summary statistics.

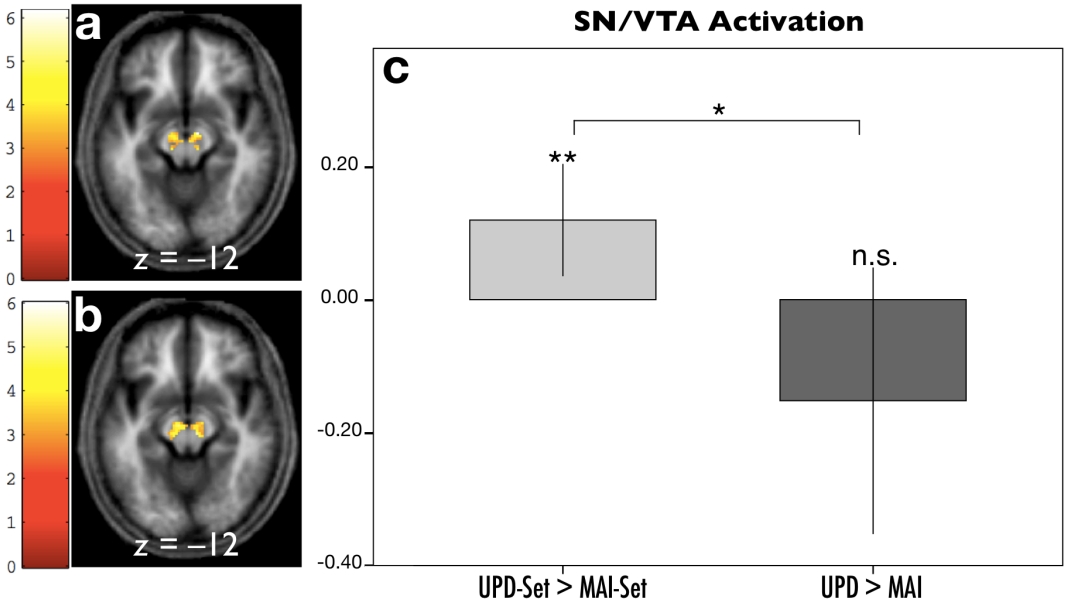

Figure 3.1 SN/VTA BOLD responses of the set and action phases. Regional responses to each level of set and action were extracted using a functional ROI defined with an orthogonal contrast. ***a***, The functional ROI for set was defined by the main effect of set over both of its levels (240 voxels, $p = 0.001$, cluster FDR). ***b***, The functional ROI for action was defined by the main effect of action over both its levels (186 voxels, $p = 0.002$, cluster FDR). Voxels within these functional ROIs were activated, as determined by small-volume corrections using a predefined anatomical ROI based on mean normalised magnetisation transfer images across subjects. ***c***, ROI analysis for the SN/VTA region across experimental phases. The SN/VTA activity was significantly larger when expecting an updating event (left bar, $^{**}p = 0.008$), whereas the SN/VTA was slight decreased on updating *per se* compared with maintenance (right bar, $p = 0.127$). An interaction between set and action in the SN/VTA was also evident ($^{*}p = 0.040$). $^{*}p < 0.05$; $^{**}p < 0.01$; n.s. not significant; UPD, updating; MAI, maintenance.

The functional SN/VTA ROI for action consisted of 186 voxels ($p = 0.002$, cluster FDR; Figure 3.1b). Principal eigenvariates were then extracted to summarise

UPD and MAI effects. The effect of updating versus maintenance was then tested with a one-sample *t* test.

Striatal set activation was only observed in the left putamen. For the sake of consistency, ROI analysis of the action phase was reported in the same region. We referred to the Automatic Anatomical Labelling atlas (Tzourio-Mazoyer et al., 2002) for the anatomical ROI of the left putamen. This anatomical ROI was masked with the thresholded activation map of the main effect of set and action separately (both used uncorrected $p = 0.001$). This yielded two functional ROIs: the putamen-set ROI consisted of 205 voxels ($p < 0.001$, cluster FDR; Figure 3.2a); the putamen-action ROI consisted of 460 voxels ($p < 0.001$, cluster FDR; Figure 3.2b). The same contrast, UPD-set > MAI-set and UPD > MAI, were tested with one-sample *t* tests after extracting the principal eigenvariates for corresponding conditions.

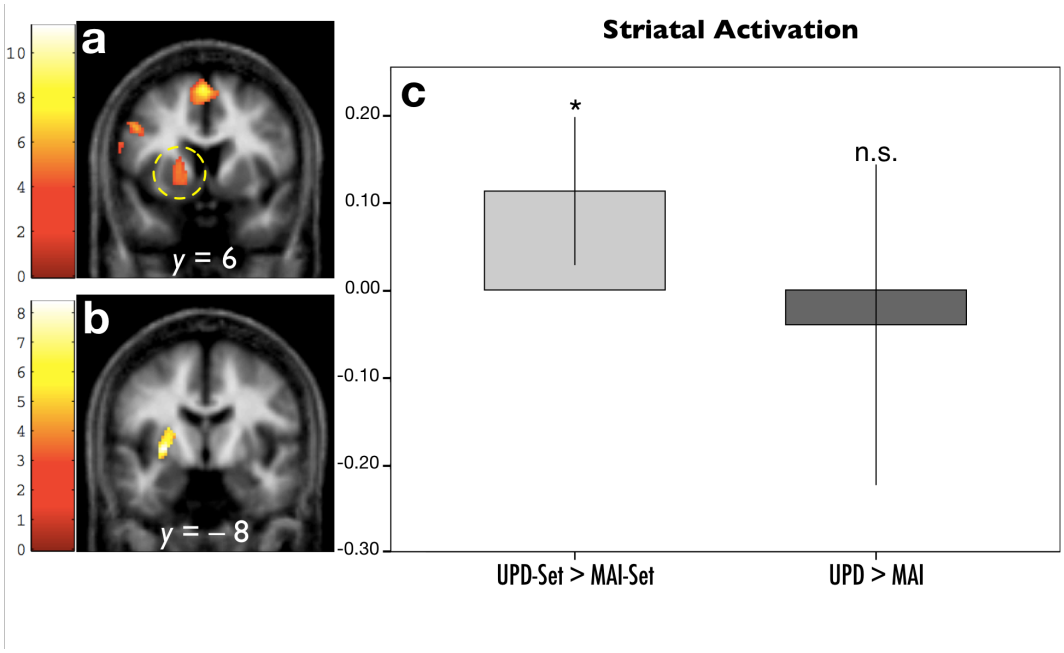

Figure 3.2 Striatal BOLD responses for the set and action phases. Regional responses to each level of set and action were extracted using a functional ROI defined with an orthogonal contrast. ***a***, The functional ROI for set was defined by the main effect of set over both its levels

(205 voxels, $p < 0.001$, cluster FDR). ***b***, The functional ROI for action was defined by the main effect of action over both its levels (460 voxels, $p < 0.001$, cluster FDR). Voxels within these functional ROIs were activated, as determined by small-volume corrections using the left putamen mask from the Automatic Anatomical Labelling (AAL) atlas. ***c***, ROI analysis for the left putamen across experimental phases. Anticipatory activity in the left putamen was significantly larger with high update probability, compared with low update probability (left bar, $*p = 0.012$). There was no difference in the striatal activity between updating and maintenance (right bar, not significant $p = 0.652$). $*p < 0.05$; $**p < 0.01$; n.s. not significant; UPD, updating; MAI, maintenance.

Activation in the DLPFC was identified in the right hemisphere during action. Given no *a priori* anatomical constraint, the ROI specification was based on an isolated cluster in the right middle frontal gyrus [peak (44, 30, 24), $p < 0.001$, cluster FDR, 203 voxels; Figure 3.3a] on the main effect of action. Features of UPD and MAI parameter estimates were extracted accordingly, followed by testing the contrast UPD > MAI using a one-sample *t* test.

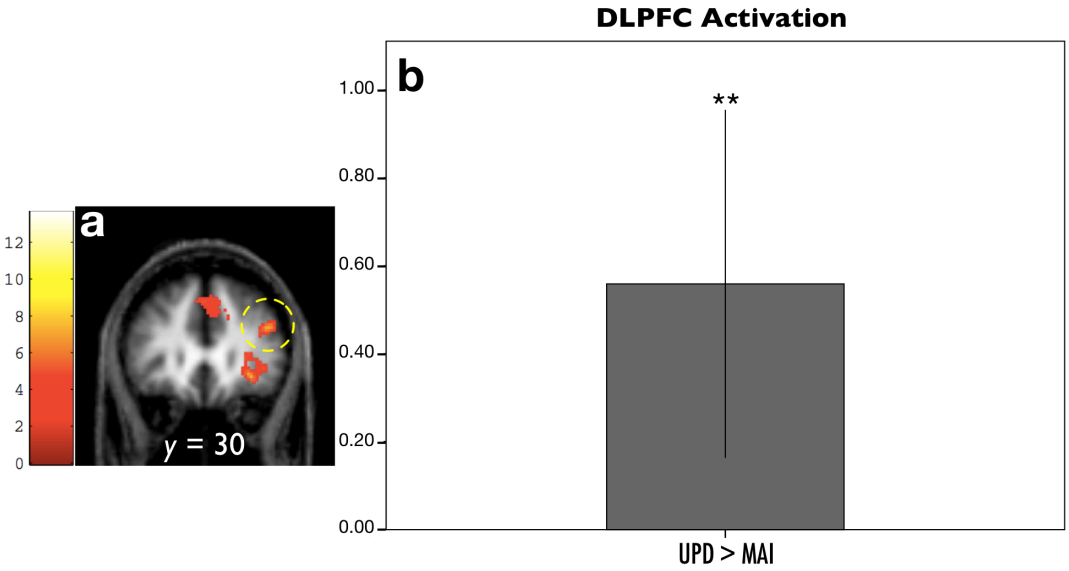

Figure 3.3 DLPFC BOLD responses during the action phase. ***a***, The functional ROI for extracting UPD-specific and MAI-specific DLPFC responses was defined by the main effect of

action over both levels, where an isolated cluster was able to be identified in the right hemisphere (arrow; 203 voxels, $p < 0.001$, cluster FDR). The ROI localisation was orthogonal to the contrast being tested. ***b***, ROI analysis revealed a significant UPD > MAI contrast in the right DLPFC, showing a larger response to updating events (**$p = 0.009$). *$p < 0.05$; **$p < 0.01$; UPD, updating; MAI, maintenance.

3.3. Results

3.3.1. Behavioural results

A repeated-measure ANCOVA, including individual working memory capacity (WMC) measures as a covariate, demonstrated a significant crossover interaction between cue (high or low probability) and action on the RTs (Figure 3.4; $F_{(1,15)} = 9.43$, $p = 0.008$; mean RTs ± SD in seconds: MCM, 1.15 ± 0.17; MCU, 1.22 ± 0.20; UCL, 1.18 ± 0.18; UCU, 1.17 ± 0.20). There was no main effect for cue ($F_{(1,15)} = 1.080$, $p = 0.315$) or action ($F_{(1,15)} = 0.005$, $p = 0.945$). A cue x action x WMC interaction was detected ($F_{(1,15)} = 7.18$, $p = 0.017$). On average, the subjects performed the task to an accuracy level of 80.71% (SD, 15.72%) during the scanning session. Statistical tests revealed no significant main effect or interaction for response accuracy. The average measure for the subjects' WMC was 6.2 letters. Performance on the task, as measured by Cowan's capacity index (Cowan, 2005), significantly predicted the WMC measure in a linear regression ($R^2 = 0.418$, ß = 0.646, $p = 0.005$).

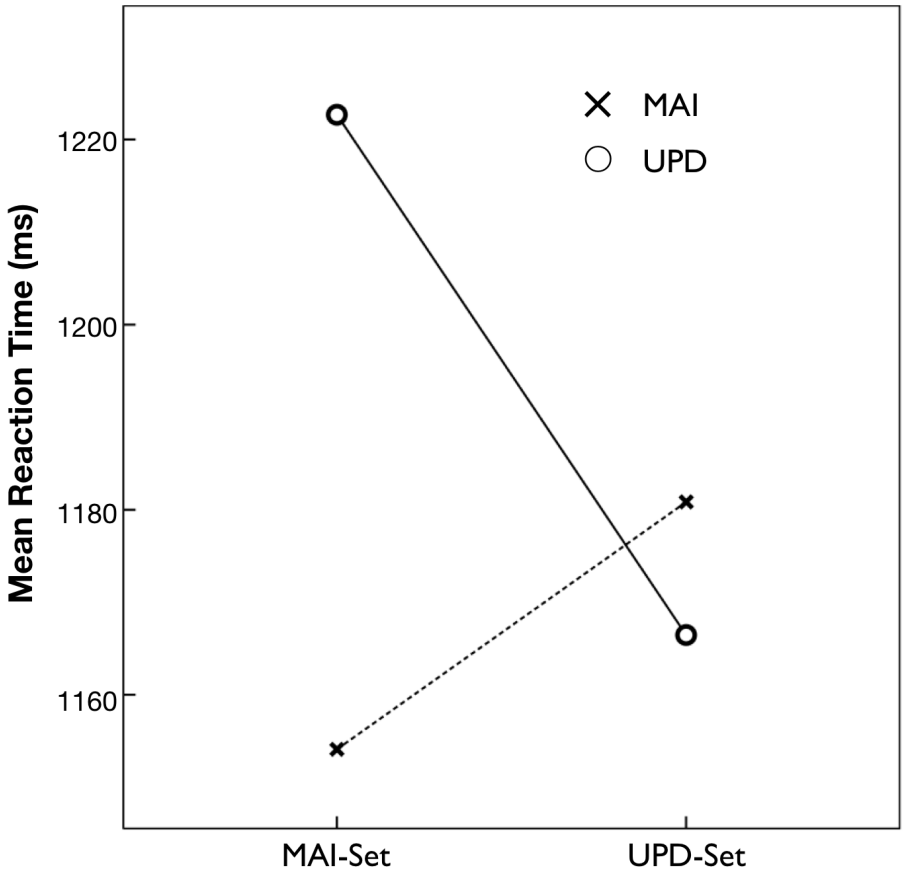

Figure 3.4 Analysis of covariance for reaction time data. RT results indicated a significant interaction between surprising (invalid) and unsurprising (valid) conditions (ANCOVA controlled for individual differences in WM capacity; $F_{(1,15)} = 9.43$; $p = 0.008$), suggesting that the subjects were able to discriminate the cues behaviourally. The MAI-set predicts the MAI event, whereas the UPD-set predicts the UPD event, both at 80% probability, explicitly instructed to the subjects.

3.3.2. Neuroimaging results

3.3.2.1. SN/VTA responses during set and action

Set activity for UPD-set was significantly larger than that for MAI-set (Figure 3.1c, left; $t_{(16)} = 3.003$, $p = 0.008$) in the SN/VTA. Then, we tested UPD > MAI using a one-sample *t* test and showed a slight trend decrease in SN/VTA activity for UPD (Figure 3.1c, right), albeit insignificant ($t_{(16)} = 1.609$, $p = 0.127$). A significant

interaction was observed between the effect of set and action (Figure 3.1c; $t_{(16)}$ = 2.237, p = 0.040).

The main effect of set revealed a common response over the UPD-set and MAI-set in the SN/VTA, the left putamen, the left premotor cortex, the SMA, the left posterior parietal cortex, and the bilateral visual cortices (Table 3.1, upper section; see also Figure 3.5). No activation in the DLPFC was detected. Clusters showing differential activation (UPD-set > MAI-set contrast) under whole-brain correction were summarised in the lower section of Table 3.1, where activations were almost restricted to the posterior brain, including the right calcarine cortex, the left middle occipital cortex, the left inferior parietal; the left premotor cortex was detected, as well.

Table 3.1 Localisation of set-related activation

				Local maxima			
Anatomical regions	Hemisphere	BA	Size	*x*	*y*	*z*	*t* value
Main effect of set[a]							
Calcarine cortex	Right	18	415	18	−96	0	11.21
SMA	Bilateral	6	546	−4	0	66	9.59
Inferior occipital cortex	Left	18	553	−16	−94	−8	9.52
Premotor cortex	Left	6	423	−48	−2	40	8.65
Cerebellum	Right	—	170	26	−66	−24	7.63
Putamen	Left	—	205	−18	2	10	7.21
Thalamus	Left	—	422	−6	−26	−6	6.61
Inferior parietal cortex	Left	40	116	−34	−46	40	6.58
Midbrain: SN/VTA	Bilateral	—	64	10	−12	−12	6.16
Premotor cortex	Left	6	62	−28	−8	52	5.88
Middle occipital cortex	Left	18	100	−30	−90	12	5.84
Superior parietal cortex	Left	7	132	−22	−66	42	5.00
Contrast UPD-set > MAI-set[b]							
Calcarine cortex	Right	17	2046	8	−74	16	5.54
Middle occipital cortex	Left	18	495	−34	−92	−2	5.50
Inferior parietal cortex	Left	40	254	−38	−44	38	5.08
Premotor cortex	Right	6	342	28	−8	52	5.06

[a]Cluster-wise significance at 0.05 FDR, using a cluster-defining threshold of $p = 0.001$; critical cluster size was 62.

[b]Cluster-wise significance at 0.05 FDR, using a cluster-defining threshold of $p = 0.01$; critical cluster size was 254.

Widespread activation under the main effect of action was observed primarily in the occipital cortices, extending into the superior parietal cortices and the bilateral frontal cortices (Table 3.2, upper section). Subcortical activation included bilateral striatum, the SB/VTA, and the thalamus. The contrast UPD > MAI revealed a distinct recruitment in the fronto-parietal network, including the left superior parietal lobule, the right middle frontal gyrus, bilateral superior parietal lobules, and the left premotor cortex (Table 3.2, lower section).

No activation was detected either in our ROIs or whole-brain correction for either the omission or deviation contrasts.

Table 3.2 Localisation of action-related activation

				Local maxima			
Anatomical regions	Hemisphere	BA	Size	*x*	*y*	*z*	*t* value
Main effect of action[a]							
Thalamus	Left	—	11022	−16	−22	10	13.62
Lingual gyrus	Right	18	13752	16	−86	−8	12.75
Parietal operculum cortex	Right	2	228	54	−22	22	9.80
Middle frontal gyrus	Right	45	203	44	30	24	7.09
Cingulate cortex	Bilateral	24	298	−4	2	40	6.85
Precuneus cortex	Right	7	240	8	−76	52	6.73
Inferior orbital frontal cortex	Right	47	211	42	48	−8	5.95
Precuneus cortex	Left	7	168	−10	−74	40	5.64
Cingulate cortex	Right	32	205	14	24	36	5.63
Cingulate cortex	Bilateral	23	162	2	−26	26	5.39
Inferior parietal cortex	Right	40	364	46	−50	44	5.16
Contrast UPD > MAI[b]							
Superior parietal lobule	Left	7	1041	−28	−60	44	10.58
Middle frontal gyrus	Right	45	141	38	30	30	8.67
Superior parietal lobule	Right	7	1104	34	−60	46	8.32
Superior parietal lobule	Right	7	140	14	−74	58	6.38
Premotor cortex	Left	6	84	−42	2	32	5.39
Premotor cortex	Left	6	73	−28	−8	50	5.18

[a]Cluster-wise significance at 0.05 FDR, using a cluster-defining threshold of $p = 0.001$; critical cluster size was 158.
[b]Cluster-wise significance at 0.05 FDR, using a cluster-defining threshold of $p = 0.001$; critical cluster size was 73.

3.3.2.2. Striatal responses during set and action

Striatal activation was only observed in the left putamen for the main effect of set, while responses in the bilateral basal ganglia were observed for the main effect of action. For consistency, we report set and action effect for the left putamen. Comparing activations in the putamen revealed that the UPD-set elicited a significantly larger response than the MAI-set (Figure 3.2c, left bar; $t_{(16)} = 2.832$, $p = 0.012$). Comparing the principal eigenvariates extracted from UPD and MAI using the functional mask showed no significant difference (Figure 3.2c, right bar; $t_{(16)} = -0.460$, $p = 0.652$).

3.3.2.3. DLPFC responses during action

We specified a cluster in the right DLPFC as a functional mask. Activity in the DLPFC during updating was significantly larger than during non-updating (Figure 3.3b; one-sample *t* test; $t_{(16)} = 2.993$, $p = 0.009$).

3.3.2.4. Neurobehavioural correlations

In neural terms, exploiting cognitive set usually speaks to optimal gain and efficiency in the presence of limited resources, thereby favouring behavioural outcomes in the absence of surprising outcomes (Fuster, 2008; Gazzaley & Nobre, 2012). We therefore tested for correlations between set activity and behavioural responses. Specifically, non-parametric correlations were performed to discover whether a greater neurophysiological set activity improved response accuracy. The

response accuracy in the UCU condition was positively correlated with the UPD-set activity in the SN/VTA (Spearman's rho = 0.546, $p = 0.023$). Similarly, the MAI-set activity in the SN/VTA was positively correlated with the response accuracy in the MCM condition (rho = 0.483, $p = 0.049$). No correlations were observed for invalid conditions: UCM accuracy and UPD-set activity (rho = 0.415, $p = 0.098$); MCU accuracy and MAI-set activity (rho = 0.188, $p = 0.469$). These significant correlations between measures of neuronal responses and behaviour lend a further validity to the physiological effects reported above.

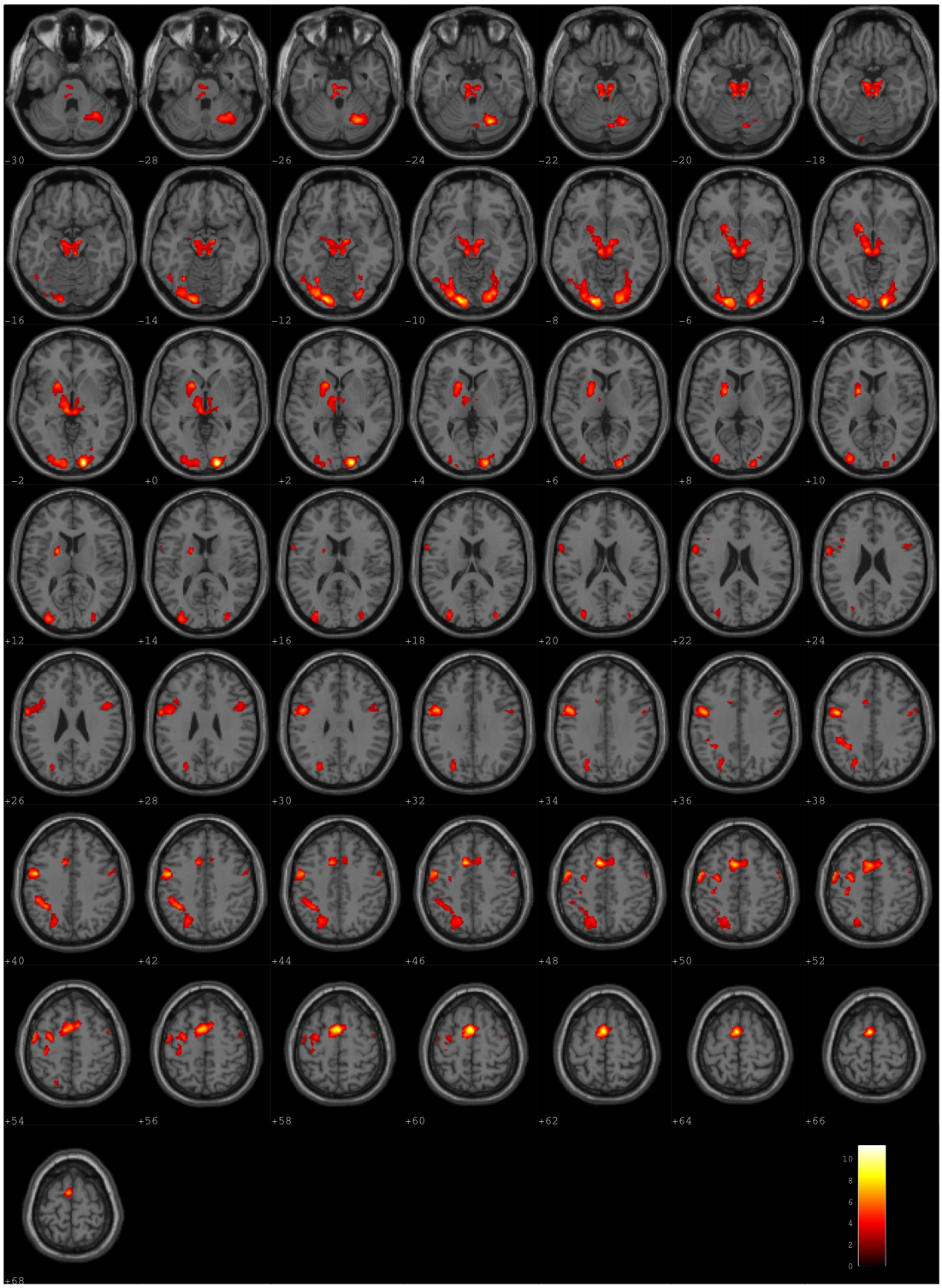

Figure 3.5 Significant clusters showing the main effect of anticipatory set. Voxel thresholding criteria, p = 0.01; cluster size threshold, 227; clusters were corrected for multiple comparison using false discovery rate.

3.4. Discussion

We tested the hypothesis that anticipating a working memory update is accompanied by activation in the dopaminergic midbrain. Consistent with our hypothesis, we found that updating-related anticipation induced sustained activity in the midbrain and striatum, suggesting a key role for tonic dopamine in the maintenance of anticipatory set (rather than maintaining memory *per se*). In addition, the amplitude of set-related activity in the midbrain and striatum was positively correlated with response accuracy in valid conditions, i.e., UCU and MCM. Memory updating *per se* did not elicit significant activity in the midbrain and striatum, as compared with non-updating activity: these results are contrary to previous studies (e.g., Baier et al., 2010; Bledowski, Rahm, & Rowe, 2009; Nee & Brown, 2013) but are discussed in light of set-related responses and the neurochemical underpinning in later sections.

3.4.1. Cue utility and anticipatory set in the midbrain

The connection between the SN/VTA BOLD response and dopamine neuron firing speaks to several plausible cellular mechanisms (Düzel et al., 2009). Although our set-related SN/VTA activations are likely to be dopaminergic – D'Ardenne and colleagues (2008) have established the correspondence between midbrain BOLD and neuronal firing in both rewarding and non-rewarding (D'Ardenne et al., 2012) paradigms – it is difficult to pinpoint their tonic nature. Fiorillo *et al*. (2003) have demonstrated that the level of tonic dopamine firing varies with the uncertainty about future events. One may accordingly speculate that the predictive cues entailed uncertainties about updating, thereby inducing tonic dopamine changes. An alternative hypothesis states a consistent perspective, that tonic DA provides

necessary level of DA to support anticipatory states in behaviour and cognition (Hong, 2013).

Task-related dopamine functions may also be explained from the information-seeking perspective. Bromberg-Martin and Hikosaka (2009) demonstrated that the midbrain dopamine signals the expectation of information, targets generating informative contents gave rise to more dopamine discharge. In line with this view, the updating cues convey more information on average (hence a higher entropy) than the maintenance cue. The availability of information in the environment may then motivate an individual to actively engage in collecting it. This can be associated with tonic enabling of parallel neural pathways (Hong, 2013) that promote working memory encoding via signal-to-noise trade-off, possibly mediated by tonic levels of dopamine. Indeed, Niv et al. (2007) have suggested that tonic dopamine reports the long-term availability of reward and may account for motivation and vigorous responding.

The SN/VTA BOLD response is, among other pathways, driven by glutamatergic afferents from prefrontal cortex (Düzel et al., 2009). It is proposed that these glutamatergic projections modulate tonic dopamine discharges (Grace, 1991). In addition, by using mixed task regressors with different temporal profiles (Donaldson, 2004), we were able to distinguish state-related processing from transient responses. It therefore seems plausible to associate the sustained activation we observed to a tonic mode of dopamine release. It is nevertheless possible that the SN/VTA activation we observed might not reflect changes in (tonic) dopamine discharge rates. In order to implicate dopamine definitively in the set-related responses we observed in the midbrain, one would require a pharmacological intervention (e.g., L-DOPA). I hope to test this assumption in future work.

3.4.2. A mechanistic remark on tonic dopamine and memory updating

How might tonic dopamine contribute to updating? A plausible model of the role of dopaminergic activity in this context can be considered in terms of the energy landscape of attractor dynamics for working memory. It has been argued that the prefrontal cortex maintains working memory representations (J. D. Cohen et al., 1997; Courtney, Ungerleider, Keil, & Haxby, 1997; Goldman-Rakic, 1995; Sreenivasan et al., 2014) in multiple attractor states. In order to achieve a more flexible switching of representations, a lower energy barrier, i.e. a relatively flat energy landscape, is required such that the system can easily move from one metastable attractor state to another (Durstewitz et al., 2000). The tonic level of dopamine discharges may play a role in modulating this transition by activating D2 receptors (Dreyer et al., 2010; Rice & Cragg, 2008; Schultz, 2007) but not the D1 receptors that have lower dopamine affinity. In effect, a prefrontal 'D2 state' would reduce the stability of attractor network dynamics and facilitate updating or transitions (Durstewitz & Seamans, 2008), thus rendering the system more responsive to inputs or revision. Tonic dopamine release might have greater influence in the striatum due to higher D2 prevalence as compared with the prefrontal cortex, where D1 receptors are more abundant (Camps et al., 1989; Goldman-Rakic et al., 1992). In this setting, flexible updating may come at the cost of lowered precision or signal-to-noise ratio. In other words, the neuronal instantiation of anticipatory set for updating may be accompanied by changes in the precision or confidence afforded to cues, with an inherent susceptibility to distracting cues.

We observed elevated BOLD responses in the left putamen when updating was expected, which might reflect D2 stimulation of striatal spiny neurons or prefrontal

afferents in layer V (Cools & D'Esposito, 2011). A possible consequence would be inhibiting the default 'NoGo' indirect pathway that, in turn, disinhibits thalamo-cortical connections (O'Reilly & Frank, 2006).

3.4.3. Neurobehavioural accounts of anticipatory set

The between-subject correlations suggested that the midbrain responses to anticipatory set predicts response accuracy in valid, but not invalid, trials, which is consistent with the idea that the brain optimises performance according to anticipated outcomes (Garrido, Dolan, & Sahani, 2011; Posner, 1980). The correlations may, nevertheless, seem somewhat counter-intuitive. Given that the midbrain was more active in updating than in maintenance, one might expect that dopamine release is essential for updating but not for maintenance. However, this is not necessarily the case. It is likely that dopamine release is optimised for specific contexts (Hong, 2013), and that for each anticipatory set there is an optimal range of dopamine levels. Cognitive performance may then have an 'inverted-U' dependency on dopamine levels (Cools & D'Esposito, 2011; Cools & Robbins, 2004), leading to positive correlations between performance and tonic dopamine in both maintenance and updating set (Figure 3.6).

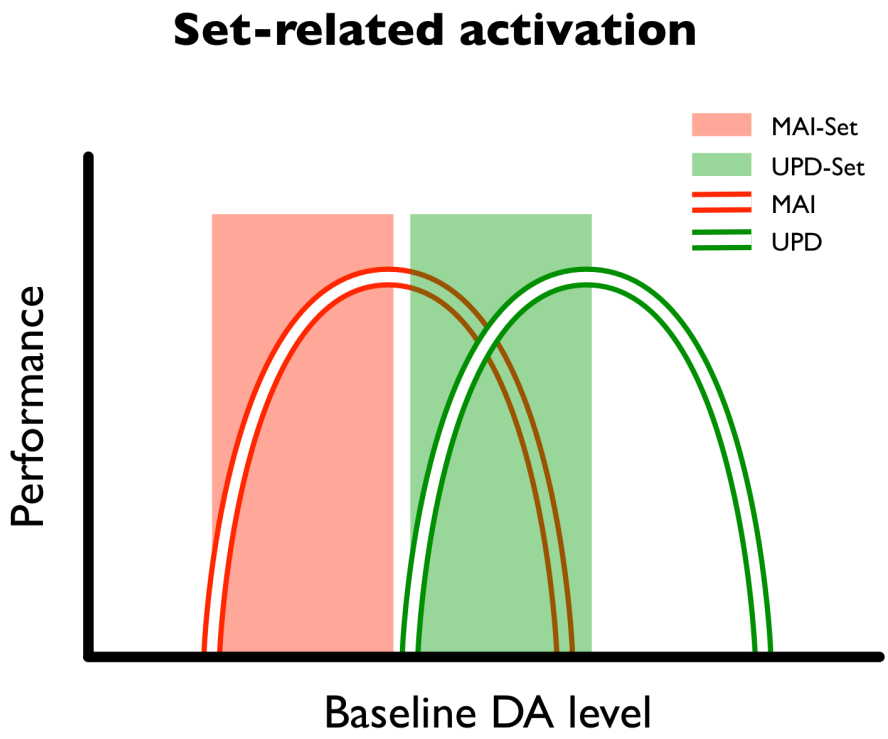

Figure 3.6 Dose-performance functions under different anticipatory sets. A schematic illustrating 'dose-performance' functions under different anticipatory set: the two curves indicate that the relationship between baseline dopamine level and behavioural performance may have an inverted-U shape. They also show that there is an optimal range in which an increase of dopamine level would improve performance. Crucially, the position of the curve may vary depending on the cognitive operation (e.g., maintenance or updating, as red and green curve, respectively). The anticipatory optimisation (i.e., set) may help defining the range of baseline dopamine (represented as shaded areas), thereby optimising performance for anticipated outcomes. This nonlinear relationship may partly account for our observation of positive correlations (between set activity and accuracy) and differential set-related activation in the midbrain. Here, we depict a possible mechanism that appeals to the notion that different behaviours (e.g., updating and maintenance) implicate different brain systems, in which optimal dopamine range is system-dependent. If set-related activity during maintenance and updating fell under the rising parts of the curves (as illustrated by the red and green box), one would expect a positive correlation, as reported in the main text. At the between-subject level, set-related modulation of dopamine could change performance in a way that depends upon subject-specific baseline dopamine levels (c.f. the "law of initial value;" for review, see Cools & D'Esposito, 2011).

The crossover interaction in reaction time data indicated that subjects discriminated between the predictive cues. Predictive cues in perceptual decisions

are known to enable better detection and discrimination of percepts, an enhancement that is attributable to attention (Feldman & Friston, 2010). Notably, a recent study demonstrated that anticipation induced shifts in baseline activity in association cortex and subsequently mediated the transfer of perceptual representations into working memory (Bollinger et al., 2010; Schmidt, Vogel, & Woodman, 2002). Such working memory representations maintained in the prefrontal cortex were also shown to be robust against distractions (Miller, 1996). These findings lend the explanation of the behavioural relevance two complementary aspects. Firstly, anticipatory set may be analogous to predictions in perceptual inference (Feldman & Friston, 2010; Friston, Friston, Kiebel, & Kiebel, 2009) that facilitate context-sensitive percepts, or possibly percepts embedded in a *repertoire* prediction in which categorical, procedural, or cognitive constructs are expected. Second, it may have optimised the (prefrontal) neuronal substrate of working memory — by modulating overall network stability, such that items can be refreshed or exchanged more flexibly. Theoretical models have outlined plausible mechanisms by which the brain can learn flexible updating in this setting (Frank et al., 2001; O'Reilly & Frank, 2006; O'Reilly, Cohen, Braver, & O'Reilly, 1999).

3.4.4. Updating activity in the meso-cortico-striatal circuitry.

Dopamine has long been implicated as an integral component of working memory function, both in terms of the stability (maintaining) and flexibility (adaptive updating) of active representations (Miller & Cohen, 2001). In particular, neurocomputational models have proposed biologically realistic mechanisms by which dopamine can contribute to working memory (M. J. Frank et al., 2001; Gruber, Kleinschmidt, Binkofski, Steinmetz, & Cramon, 2000; Hazy et al., 2007; R. C. O'Reilly & Frank, 2006). In these accounts, phasic bursts of dopamine are

associated with selective updating of working memory representations through the fronto-striatal circuitry that is equipped with a 'gating' mechanism (Baier et al., 2010; e.g., M. J. Frank et al., 2001). Presumably, the brain can learn when to gate information (R. C. O'Reilly & Frank, 2006), such that the midbrain dopaminergic neurons will dispatch a 'gating' signal (in the form of phasic bursts) to enable fast encoding (updating) given context-relevant percepts. Recent empirical findings in human functional imaging have demonstrated midbrain activation when updating working memory of visual stimuli or contexts (D'Ardenne et al., 2012; Murty et al., 2011). Murty *et al*. (2011) concluded that updating selective elements in verbal working memory activated the SN/VTA region, relative to simply maintaining or completely overwriting the working memory content. D'Ardenne *et al*. (2012) reported a phasic increase in BOLD response in the SN/VTA during updating as compared with non-updating. In general, midbrain responses are potentially attributable to updating. However, our results suggest the midbrain was less active during updating, relative to maintenance (Figure 3.1 right). One possible explanation for this may lie in our experimental design: in our study we manipulated anticipation about updating, which affected (putative) dopamine activity *before* updating. This may be relevant if the same (dopaminergic) systems are implicated in updating *per se*. In other word, the anticipatory set interacts with the effect of updating.

Interactions between anticipatory set and update phases may be explained with reference to the hypothesis of tonic-phasic homeostasis (Bilder et al., 2004; Grace, 1991). This hypothesis states that the level of extracellular dopamine – as determined by tonic release – provides the mechanism to up- or down-regulate the magnitude of phasic discharge. In principle, this is based on the D2 auto-receptor stimulation located on dopamine terminals. Only tonic dopamine release is proposed to be

capable of stimulating these D2 auto-receptors – without being affected by the re-uptake process that acts primarily on phasic dopamine. As a consequence, background dopamine levels would suppress the neuronal responsiveness and hence spike-dependent (phasic) discharge. In other words, increased tonic levels would result in a depressed phasic responsiveness. According to this hypothesis, the tonic release of dopamine is likely to be elicited by the presynaptic glutamatergic afferents from the prefrontal cortex. The relationship between prefrontal cortical activity and the midbrain may be an important determinant of sustained BOLD responses in the midbrain (Düzel et al., 2009).

We noted that our striatal responses (Figure 3.2c right) were inconsistent with previous studies of working memory updating (Bäckman et al., 2011; Bledowski et al., 2009; Kuhl, Bainbridge, & Chun, 2012; Murty et al., 2011; Podell et al., 2012). These studies suggested striatal recruitment with working memory updates, whereas our results indicated that the striatum was activated during updating and maintenance. More importantly, striatal activity did not differ significantly between updating versus non-updating conditions. This finding may be sensible when viewed from a predictive coding perspective. Several studies have suggested that the striatum has a role in processing salient or unexpected events; namely, the response in the striatum can be related to prediction error (O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; O'Doherty et al., 2004; Ouden et al., 2010). Failing to observe significant updating-specific recruitment in the striatum may reflect the fact that subjects anticipated updating. As such, an updating event, once predicted, was less surprising. This argument could be further extended to cover that reporting updating-specific striatal activation. That is, without manipulating anticipation, working memory has the propensity to occupy a low entropy (stable) state (that

supports robust maintenance, or 'D1 state'; see R. C. O'Reilly, 2006). Sudden but infrequent updating may incur greater prediction error as expressed through striatal activation. Whereas, for frequent and expected updating, staying in low entropy states could be suboptimal; instead, migrating to a state of higher entropy may reduce the average surprise over time (cf. Friston, 2009). In other words, expected surprise is not really surprising and may not be accompanied by prediction error or surprise responses that would be seen when the surprise was unexpected.

Ample evidence has demonstrated the involvement of dorsolateral prefrontal cortex (DLPFC) in various working memory tasks. Studies fractionating working memory subprocesses have suggested that the DLPFC is responsible for encoding, maintenance, and manipulation of working memory representations (for review, D'Esposito, Postle, & Rypma, 2000). It is therefore difficult to disentangle the actual role of updating-related DLPFC activation in the current study. Particularly, the functional implication of right-lateralised activation is unclear. However, TMS-induced disruption in the working memory network may shed light on the time course of information flow, where the right DLPFC appeared to be critical at an early phase of updating (for review, Linden, 2007). More recently, D'Ardenne *et al*. (2012) showed that time-locked, TMS-induced disruption was only effective on the right DLPFC after the onset of updating information.

3.5. Conclusions

In summary, our data suggest that anticipating to update working memory representations activates the dopaminergic midbrain and striatum, which speaks to a key role for tonic dopaminergic activity in modulating the flexibility of representations based on the volatility of the environment. These anticipations

interacted with subsequent updating processes in the same regions to suppress transient responses in the midbrain and the striatum, which otherwise respond strongly updating events. While these latter findings are *prima facie* inconsistent with previous findings (D'Ardenne et al., 2012; Murty et al., 2011), they can be easily accounted for from the perspective of predictive coding (Friston & Stephan, 2007; Rao & Ballard, 1999) — in that expected updates are not inherently surprising. In general, our data speak to a role for dopamine in modulating the precision, or gain, on sensory information during working memory processing (M. J. Frank & Badre, 2012; Friston et al., 2012). Our findings thus represent a step towards understanding both how working memory flexibility is modulated in response to the demands of environment, and the likely role of tonic dopamine in working memory function.

Chapter 4. Multivariate correlates of anticipatory set

Observing a mechanism that works by predictions gives insights into its hierarchical structure. In particular, empirical evidence based on electro-magnetophysiology and functional MRI suggests that the deviation and omission of anticipated sensory signals can be distinguished. However, analogous evidence with regard to higher cognition – if one were to assume both computations rest on a common predictive principle – has yet to be established. To extend our understanding in this regard, we employed a working memory updating task, which entailed an update to previously encoded memory that was anticipated with high or low probability. Using multivariate pattern analysis (MVPA), we demonstrate that the brain response to omission and deviation can be dissociated. Surprising events diverge from statistical regularities and are inherently rare; the use of MVPA revealed the distributed and covarying nature of responses to these events. We conclude that monitoring and maintaining working memory is a predictive process that involves distributed systems that underpin adaptive control and stable set-maintenance control, i.e., the fronto-parietal and cingulo-opercular networks, respectively.

4.1. Introduction

A hallmark of intelligent organisms is the ability to tolerate error-making (Carpenter & Doran, 1986) and assimilate error by learning from them (Schultz, Dayan, & Montague, 1997). This often entails a predictive model constantly estimating the states of the world. The estimates may adjust if the predicted state is incongruent with the actual state, through a process resembling hypothesis testing, or model comparison. Observing a predictive system making error gives crucial insights into how it works (Frith, 2007). This observation is fundamental to many modern cognitive neuroscience studies that treat the brain as an inference device, notably in the area of perception (Garrido, Kilner, Stephan, & Friston, 2009), decision-making (Ouden, Friston, Daw, McIntosh, & Stephan, 2009), and reward processing (O'Doherty et al., 2007); with special cases referring to theoretical frameworks of the Bayesian brain (Mathys, Daunizeau, Friston, & Stephan, 2011) or Kalman filter (Doya, 2007).

The human neocortex exhibits a hierarchical organisation, with repetition of highly similar lamination and laminar connectivity up the hierarchy (Felleman & Van Essen, 1991). It has been speculated on both theoretical and empirical grounds (Friston, Friston, Kilner, & Harrison, 2006; Markov & Kennedy, 2013; Mumford, 1992; Rao & Ballard, 1999; Bastos et al., 2012) – that the brain is an active inference device which constantly updates its hierarchical model of the world by processing only the discrepancy between bottom-up sensory inputs and top-down predictions, i.e., the prediction error. Many studies are concerned with the corresponding generative model underlying perceptual inference. However, the possibility as to whether higher cognition works under the same predictive principle, and whether it

engenders error signals – not to a stimulus-bound surprise but to contextual violations (e.g., improbable events) is not so clear.

This has led us to the proposition that working memory, which bridges sensory-related representations and goal-related attentional bias and task-level control (Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008; Gazzaley & Nobre, 2012; Sreenivasan et al., 2014), may well adhere to the same hierarchical predictive principles (Friston, 2008). We reasoned that stimulus-bound and context-dependent surprise signals might reveal a hierarchical organisation in working memory.

Detecting neural codes – representing surprise-related states – using classical mass-univariate analysis may be a challenge. Surprise, in probabilistic terms, violates statistical regularities and is inherently rare in experimental settings. Its detection therefore suffers from low statistical power. In addition, unlike sensory signals, which are highly localised, neural representations of higher constructs may be sparse and vary considerably across individuals. To make use of weak but informative voxel data and to accommodate individual variability, exploring the covariance structure amongst surprise-related voxel data can be effective (Norman, Polyn, Detre, & Haxby, 2006). This is referred to as multivariate pattern analysis (MVPA). MVPA is based on statistical learning theory and classification algorithms which, in practice, takes joint information across all data features, as opposed to treating the features independently (Schrouff et al., 2013). We therefore applied MVPA to surprise-related responses.

To characterise surprise-related responses of a conceptual nature, we employed a working memory updating task based on the delayed match-to-sample paradigm. The principal manipulation was to enable context-sensitive updating or maintenance of sequential working memory. Trial-specific expectations about the propensity of an

imminent update were afforded by a preceding predictive cue. Task-related surprise pertains to the omission of an anticipated update, or the deviation from anticipated maintenance with an unexpected update. In particular, the omission response may speak to the neural substrates implementing prediction signals *per se* (SanMiguel, Saupe, & Schröger, 2013; Wacongne et al., 2011). This is in contrast with the deviation response, in which unexpected inputs may contribute to neuronal responses. Our hypothesis was that omission and deviation speak to abstract and concrete levels of processing and may therefore be differentiated in multivariate patterns that are evoked by these events.

4.2. Methods

4.2.1. Data Pre-processing

Functional data were analysed using SPM12 (Statistical Parametric Mapping; Wellcome Trust Centre for Neuroimaging, London, UK). Preprocessing of functional images included correction for geometric distortion using field maps (Hutton et al., 2002; Jezzard & Balaban, 1995), realignment via affine registration to correct for head movement, slice timing correction, co-registration with respect to anatomical images, normalisation to MNI space based on the anatomical normalisation parameters, interpolation to voxel size of 2 x 2 x 2 mm. No spatial smoothing was performed.

4.2.2. Mass-univariate analysis

When attempting multivariate pattern analysis with fMRI data, the temporal delay of the haemodynamic responses with respect to the stimulus onsets must be

taken into account. Modelling brain responses with the canonical haemodynamic response function (HRF) in the general linear model (GLM) offers a robust way to obtain responses at corrected time points – if individualised parameters for HRF are unavailable (Schrouff et al., 2013). The responses obtained were in the form of beta maps (GLM parameter estimates), encoding evoked responses specified in the design matrix. In short, GLM constitutes a pre-processing stage prior to the pattern analysis (Norman et al., 2006).

The effects of interest were the omission and deviation of updates, and their dissociable spatial distributions were hypothesised. Regional responses to these effects were modelled with an impulse response function in the design matrix. To capture the responses in a trial-by-trial manner, as many regressors as *correct* instances of omission and deviation were placed in the design matrix. The total number of regressors was less or equal to 20, depending on individual performance.

Another four regressors were used to model remaining task effects, including the main effect of anticipatory set (prolonged epoch from high and low cues), and that of action (non-specific transients of updating and maintenance). Other variables included non-specific visual onsets (encoding, probing), and switching (transition between trials). Events in an error trial were modelled as non-specific visual onsets.

All the regressors were convolved with a canonical HRF to produce haemodynamic regressors for GLM. The full model also included head motion and low-level physiological variations that were of no interest. Head motion was summarised using three translations (x, y, and z directions) and three rotations (pitch, roll, and yaw), derived from the realignment procedure. The physiological measures comprised six cardiac regressors, six respiratory regressors, and two regressors for heart rate change and change in respiratory volume (Hutton et al., 2011). Beta maps

were obtained by inverting subject-specific GLMs, which in turn provided the data features (GLM parameter estimates) for subsequent pattern classification.

Contrasts for the main effect of action were computed for each subject, a one-sample t-test was used to test these contrasts at group level, with a relatively liberal threshold (uncorrected p = 0.01; cluster size, 10). This provided a 'functional localiser' of the neural substrate of non-specific mnemonic processing in question. The ensuing voxels were defined as a mask for the subsequent classification procedure (kernel construction).

4.2.3. Pattern classification

A classical task in machine learning (or MVPA) is to derive, in the feature space, a decision hyperplane, whereby exemplary feature vectors are adequately separated to reveal *known* categorical discrimination. This is specifically referred to as *supervised* learning – as opposed to unsupervised learning – since the category to which each 'example' belongs is identified *a priori*. Features are usually instances of observations, e.g., whole brain voxel data of BOLD responses, or a subset (excluding non-brain tissue). One can surmise that voxel responses to different experimental conditions exhibit differentiable spatial patterns and infer that they must belong to distinct categories.

Multiple classification algorithms are capable of defining a decision boundary given training data. The one applied in this study was Support Vector Machines (SVM). Derivation of the hyperplane in SVM yields classifier estimates (or weights), which represents a vector orthogonal to the hyperplane. SVM optimises the (binary) decision by maximising the margin between two groups of points, where a set of 'support vectors' is of particular interest. The support vectors are feature vectors

representing non-trivial classification problems, in that they are in the immediate vicinity of the true decision boundary, therefore they are the most difficult ones to tell apart. An optimised weight vector is essentially the linear combination of the support vectors (using Lagrange multiplier as coefficients; for details please refer to Chu, 2009).

The classification procedure was carried out using the PRoNTo toolbox (Schrouff et al., 2013). Within-subject classification was conducted. Feature vectors were beta maps of individual omission and deviation trials derived from the GLM analysis, voxels from outside the brain were excluded using the whole brain mask provided in SPM. A second-level mask constituting the main effect of action was subsequently applied, this allowed one to frame the classification problem in accordance with the functional anatomy of working memory (see Schrouff et al., 2013). A kernel matrix was computed based on the masked feature set. Kernel methods have computational advantages when dealing with high dimensional data, and therefore are ideal for neuroimaging dataset. In our case, using linear kernels, the kernel matrix was a pair-wise inner product of feature vectors. This enabled the maximum-margin optimisation to take place in a transformed feature space (Hofmann, 2008).

4.2.4. Cross-validations

SVM performance was assessed by means of a 'leave-one-scan-per-group-out' cross-validation scheme. This tested the generalisability of the classifier by systematically removing one scan from each condition during the training phase. Expected accuracy of the classifier can then be estimated by testing the hold-out set. This prevented over-fitting, as is often the case when the dimensionality of the

feature space is larger than the size of feature set (Mahmoudi, Takerkart, Regragui, Boussaoud, & Brovelli, 2012). The performance of the classifier was expressed in terms of balanced accuracy

$$\text{Acc} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{4.1}$$

where F/TP and F/TN are false/true positives and negatives, respectively. Also note that True Positives corresponded to the correctly classified deviation conditions, whilst True Negatives to that of omission conditions.

A common issue in assessing classifier performance is an unbalanced dataset, the consequence of which is invalid detection and fraudulent estimate of classifier accuracy. In the case of imbalance, the class having more instances is referred to as a majority class, whereas the lesser set is a minority class. To prevent unbalanced feature points in neuroimaging studies, balanced experimental designs are frequently used, such that the number of trials in respective conditions of interest is equal. This was the case in our study. Nevertheless, individual variations in task performance required some trials to be discarded, thus inevitably creating unbalanced data. Various approaches have been proposed to tackle this problem (Japkowicz & Stephen, 2002). For example, under-sampling the majority class was performed in Nee and Brown (2012). Alternatively, over-sampling the minority class is also viable. The method adopted here was SMOTE (Synthetic Minority Over-sampling Technique; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). We applied SMOTE by randomly selecting a feature vector in the minority class. We then determine another feature vector amongst the *k*-nearest neighbour of the one under consideration, followed by taking the difference of the two vectors. This represented a line segment along which the synthesised feature vector would then be selected at a random point.

This procedure was repeated until both classes were of equal size. Depending on individual subject's performance, the number of feature vectors synthesised was between 0 and 3.

4.2.5. Permutation testing

Permutation tests were performed on the classification outcome. This procedure supplemented the accuracy estimate from the cross-validation; due to the small size of the test set, and co-dependent cross-validation trials, the variance estimate and the extent to which the observed test accuracy occurred by chance cannot be accessed. Permutation tests furnished an empirical cumulative distribution of the statistics (here, the accuracy estimates) by repeatedly shuffling the class labels corresponded to respective feature vectors. This treatment was proposed in Golland and Fishl (2003), and has been implemented in the PRoNTo toolbox.

For each subject, 100 permutations were performed. The significance level was set at 0.05.

4.2.6. Visualising weight maps

The objective of MVPA in the scope of current work was to obtain individual weight maps. In practice, the use of linear SVMs produces linear boundaries in the original feature space. This results in a straightforward interpretation of the weight maps. In this case, the size of the weight value corresponds, in a voxel-wise manner, to how much a voxel contributed to the decision (Mahmoudi et al., 2012).

It was therefore informative to summarise weight maps across subjects to visualise voxel-wise contribution to classification. To achieve this, the original weight maps were transformed into standard scores and averaged over subjects.

Then, the mean weight map was overlaid on a montage of anatomical image. Blue and red colour codes were used to indicate the voxel's contribution to the omission and deviation classes, respectively. No thresholding was applied to any weight map during the procedure. In addition, to determine the class-dependent importance with respect to anatomical regions, the number of informative voxels was recorded. Between-class comparison of these numbers was carried out at the group level, in a region-by-region manner. Specifically, a mask was created using the WFU PickAtlas toolbox (Maldjian, Laurienti, & Burdette, 2004; Maldjian, Laurienti, Kraft, & Burdette, 2003) based on the 116-region AAL atlas (Tzourio-Mazoyer et al., 2002). The resultant mask contained voxels labelled according to regions. Positive and negative weights, corresponding to deviation and omission, respectively, of each subject were then identified and grouped based on the labelled mask. From the AAL atlas, we combined regions to form the following seven systems: (1) fronto-parietal network (FPN); (2) cingulo-operculum network (CON); (3) basal ganglia (BG); (4) thalamus (Th); (5) temporal cortex (Tpx); (6) cerebellum (Cbx); and (7) visual cortex (Vix). The contrast "omission count > deviation count" was performed to obtained a table of subject-by-region count difference. The differences were evaluated for normality using the Shapiro-Wilk test, followed by a one-sample *t* test.

The region combinations were as follows (a) the fronto-parietal network included: precentral, superior frontal, middle frontal, inferior triangularis, orbital frontal, SMA, superior parietal, inferior parietal, supramarginal, angular, and precuneus regions; (b) the cingulo-operculum included: superior orbital frontal, middle orbital frontal, frontal opercular, rolandic opercular, superior medial frontal, medial orbital frontal, insula, cingulum, and postcentral regions; (c) the basal ganglia included: caudate, putamen, and pallidum regions; (d) the temporal cortex included:

fusiform, Heschl's, superior temporal, superior temporal pole, middle temporal, middle temporal pole, and inferior temporal regions; (e) the visual cortex included: calcarine, cuneus, lingual, superior occipital, middle occipital, and inferior occipital regions. All regions were bilateral.

4.2.7. Correlation analysis

The state-dependent responses associated with the informative voxels were tested for correlations with corresponding reaction time measures. This was performed within subjects on a trial-by-trial basis. Firstly, BOLD response of voxels bearing positive weights, which informed the deviation class, were extracted. The multi-voxel BOLD data were then compressed in terms of the principal eigenvariate and paired with reaction time of the corresponding trial. Spearman's rho was calculated for each subject. The correlation statistics were then tested at the group level.

For the omission class, the informative voxels carried negative weights. The same procedure was repeated.

Brain responses underlying counter-informative voxels were also tested to detect non-specific correlations. The stated statistics were obtained by – in deviation trials, for example – correlating reaction times with responses extracted from omission-informing voxels. Likewise, omission responses extracted from deviation-informing voxels were correlated with reaction times of omission trials.

The correlations between surprise-related response accuracy and pattern-informed BOLD responses were tested at between-subject level. For each subject, data features (i.e., first eigenvariate) of pattern-informed activity were extracted and averaged across trials. The accuracy was calculated as the proportion of correct

responses for the omission and deviation conditions, respectively. Spearman's non-parametric correlations were performed to detect whether surprise activity predicts corresponding performance.

4.3. Results

4.3.1. Classifier performance

Overall classification accuracy was 89.58%, as calculated by averaging balanced accuracy estimates across subjects (range, 56.25% - 100%). The mean accuracy by class was 90.44% and 88.73% for the deviation class and omission class, respectively (range, 50.00% - 100% and 62.50% - 100%). The permutation test indicated 16 out of 17 subjects were significant ($p<0.05$).

4.3.2. Visualising weight maps

The mean weight map revealed an apparent discrimination between the pattern of deviation responses (Figure 4.1, red blobs) and that of omission responses (Figure 4.1, blue blobs).

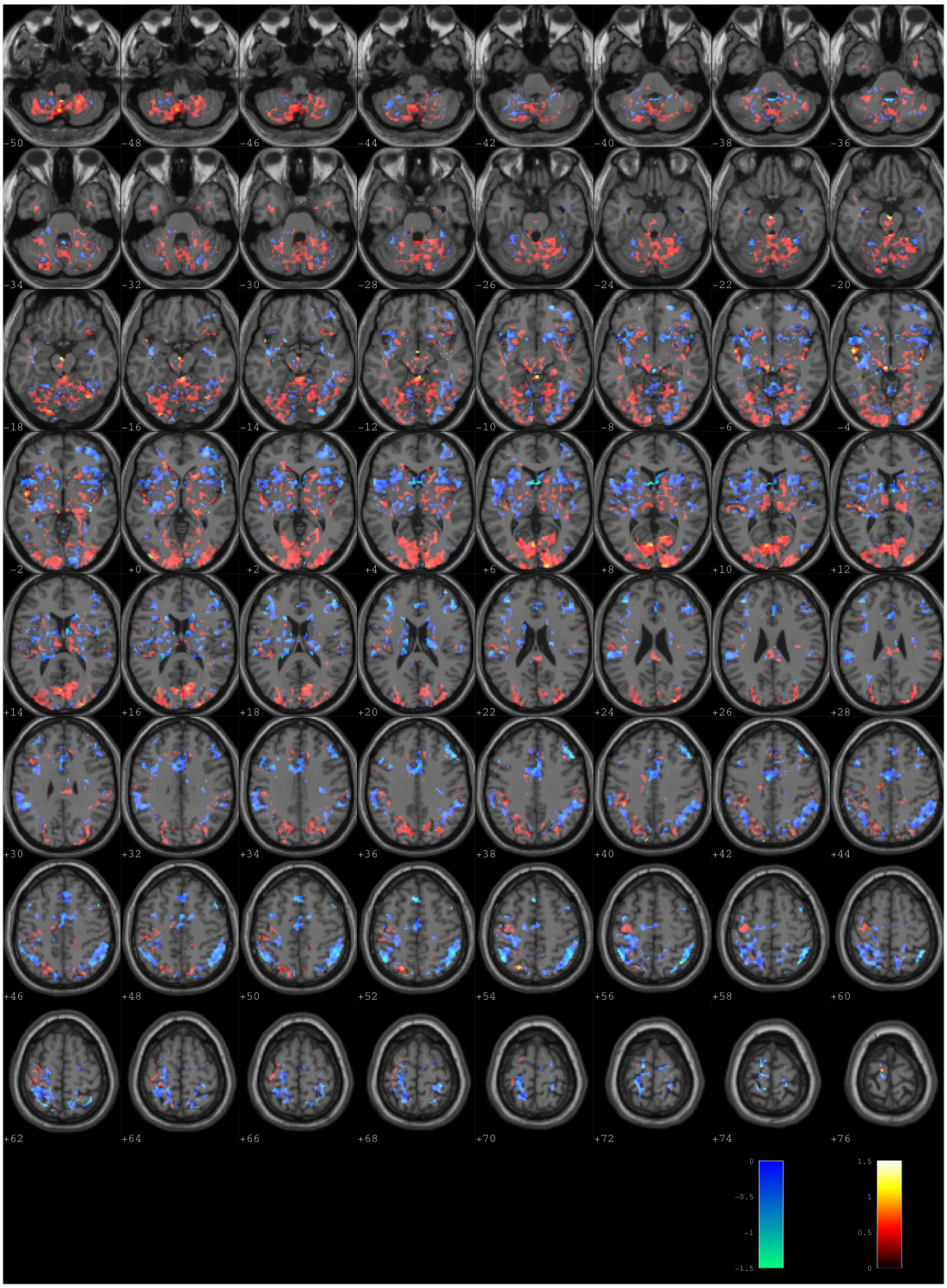

Figure 4.1 Mean weight map. Classifier weights of each subjects were transformed into standardised scores and averaged in a voxel-wise matter at the group level. The resultant matrix was overlaid on an structural MR image, and colour-coded according the respective class of brain response: red, deviation; blue, omission. Voxel brightness corresponds to how informative they are in relation to the omission and deviation processes. A trend separation between two patterns can be observed, in which the deviation pattern occupied the posterior brain, including the visual cortex and the cerebellum. The inclusion of the midbrain was also noted. In contrast,

the omission pattern encompassed voxels that conform to the fronto-parietal and cingulo-opercular networks.

The distribution of the deviation voxels was mainly in the caudal regions, including most of the visual cortex, precuneus, and the cerebellum in both hemispheres. The pattern extended over the thalamus, and the SN/VTA in the midbrain.

The omission pattern, on the other hand, closely resembled the fronto-parietal network, including the bilateral inferior parietal cortices and the bilateral middle frontal cortices. Other regions enclosed by the pattern constituted the cingular-opercular network. The involvement of the striatum was observed in the ventral part, including the nucleus accumbens, and the bilateral caudate. Bilateral lingual gyri and the right occipital cortex also contained informative voxels about omission processing.

Informative voxels in the basal ganglia shared mixed contributions to the omission and deviation classes. Among the two classes, omission mainly implicated the caudate tail and the caudate body (Figure 4.1, z = 18 to z = 24), with limited contributions from the caudate head and putamen. Whereas, substantial contributions to the deviation class was observed in the caudate head (Figure 4.1, z = -8 to z = 4), as well as the putamen.

The most informative voxels (defined by the 90th percentile for respective class weights; Figure 4.2) were in the visual cortex, precuneus, cerebellum, thalamus, pallidum, and SN/VTA for the deviation pattern, and in the insular, ventral striatum, caudate, SMA, DLPFC, and inferior parietal cortex for the omission pattern.

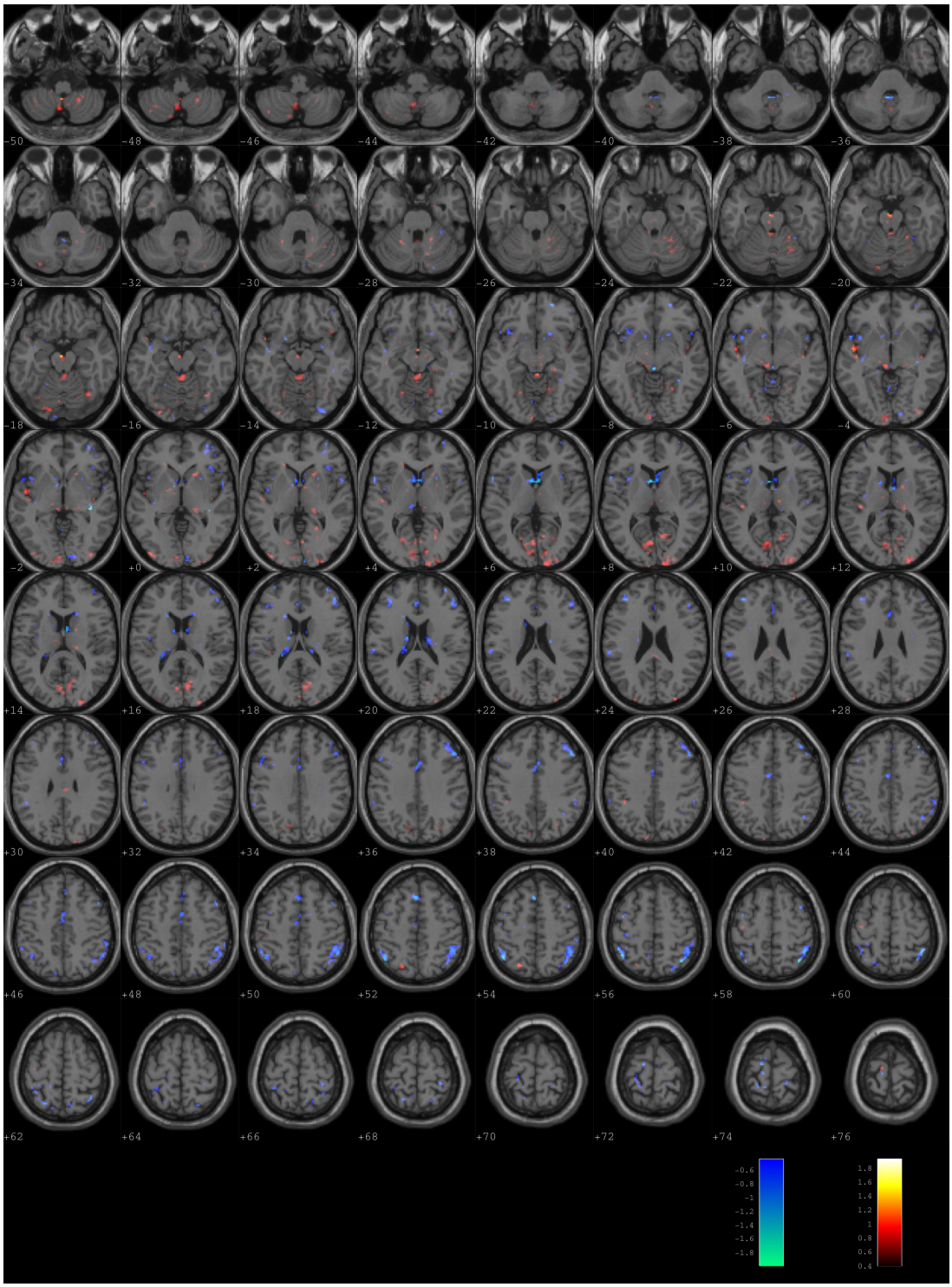

Figure 4.2 Mean weight map showing voxels above the 90th percentile voxels. Voxels of average classifier weight above the 90th percentile from both classes were overlaid on structural MR images. This presents the most informative voxels for the omission and deviation patterns. For the deviation pattern (red), these voxels were observed in the posterior occipital cortex, the left precuneus, the thalamus and the midbrain. For the omission pattern, the ventral striatum, the caudate nucleus, the anterior cingulate cortex, and the fronto-parietal regions contributed the most informative voxels.

The voxel count difference (omission count > deviation count) across the seven major regions was normally distributed (Table 4.1). The one-sample *t* test showed that omission was greater than deviation in voxel count in two regions, FPN ($p < 0.001$) and CON ($p = 0.003$). By contrast, more voxels in Cbx ($p = 0.037$) and Vix ($p < 0.001$) were informative to the deviation class. Voxel dominance in BG, Th, and Vix was indistinguishable ($p = 0.926$, 0.054, and 0.178, respectively; Table 4.1). Figure 4.3 shows region-specific voxel count based on the group mean weight map.

Table 4.1 Statistical tests on the voxel count difference (omission count > deviation count).

	Normality test : Shapiro-Wilk		One-sample *t* test		
	Statistic	Significance	*t*	df	Significance
FPN	.955	.540	4.401	16	<0.001
CON	.957	.580	3.502	16	0.003
BG	.984	.984	0.095	16	0.926
Th	.944	.365	-2.078	16	0.054
Tpx	.916	.128	-1.408	16	0.178
Cbx	.895	.055	-2.276	16	0.037
Vix	.938	.298	-4.773	16	<0.001

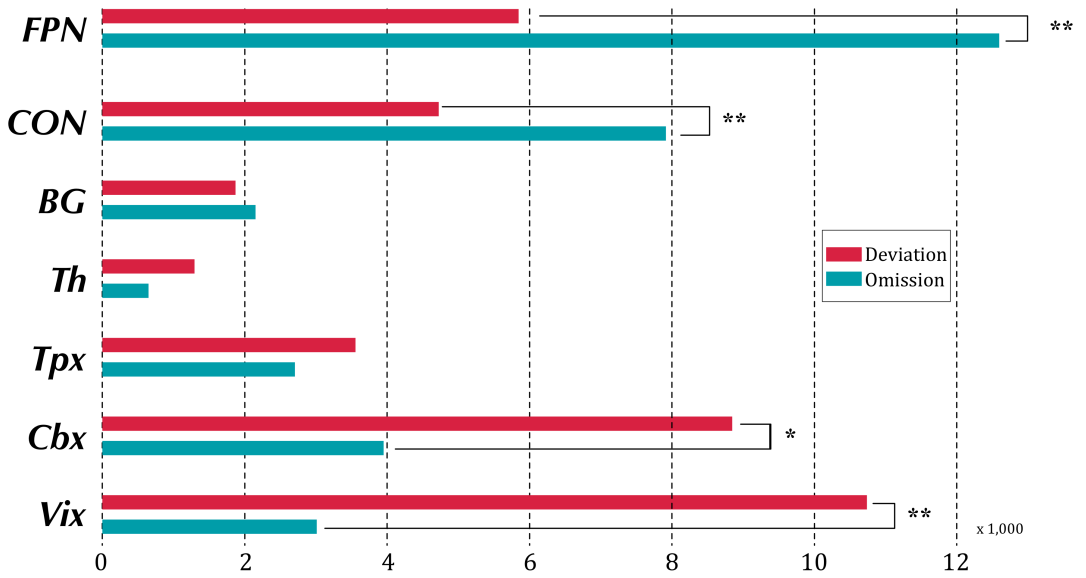

Figure 4.3 Region-specific voxel count based on the mean weight map. The number of informative voxels from the omission and deviation patterns was counted in accordance with the anatomical regions of interest. The voxel count presented here was based on the count of mean weight map; whilst the statistical significance was based on voxel counts of single subject weight maps. The contrast on which the one-sample *t* test was based is omission count > deviation count. Region definition was based on a mask created using WFU PickAtlas and the 116-are AAL atlas. Informative voxels were present in all regions but the relative dominance varied as reflected by the classification. FPN, fronto-parietal network; CON, cingulo-opercular network; BG, basal ganglia; Th, thalamus; Tpx, temporal cortex; Cbx, cerebellum; Vix, visual cortex. All regions are bilateral. *$p < 0.05$; **$p < 0.01$.

4.3.3. Pattern-informed neurobehavioural correlation

Correlation analysis was performed using within-subject weight maps. Significant negative correlations suggested that, during deviation trials, BOLD responses in informative voxels predicted response speed ($p = 0.0319$, $t_{(16)} = -2.351$). Counter-informative voxels, i.e. deviation responses extracted with omission-informative voxels, did not predict performance in deviation trials ($p = 0.1567$, $t_{(16)} = -1.486$).

No correlation with reaction time was detected for omission responses extracted from omission voxels ($p = 0.8706$, $t_{(16)} = 0.166$), or for counter-informative omission responses ($p = 0.5220$, $t = 0.655$).

At between-subject level, there was no correlation between omission responses and omission trial (UCM) accuracy (Spearman's rho = 0.2924, $p = 0.2548$), or between deviation responses and deviation trial (MCU) accuracy (rho = -0.1897, $p = 0.4658$).

4.4. Discussion

Using machine learning technique and blood oxygenation level-dependent (BOLD) imaging, we tested the hypothesis that the following two classes of surprise-related cortical responses are dissociable in terms of multivariate patterns: (1) *omission*, events in which category-specific environmental cues were anticipated (i.e., an update) but failed to occur; (2) *deviation*, events in which unexpected updates occurred. Non-specific effects of updating and maintenance were controlled for both types of responses. Consistent with our hypothesis, the classification exceeded better-than-chance performance. Inspecting the mean weight map revealed two well-defined brain systems: (1) the meso-cerebello-occipital network, responsible for deviation processing; (2) the joint fronto-striato-parietal and cingulo-opercular networks, responsible for omission processing. In light of the circumstantial evidence that surprise impairs perceptual performance, we supplemented our results with the finding that the within-subject amplitude of pattern-informed deviation responses predicts faster reaction time, whilst the same responses did not predict corresponding accuracy, suggesting little evidence for

speed-accuracy trade-off. The negative correlation was not detected in relation to omission activity.

4.4.1. Fractionating the sources of prediction

To what extent do multivariate patterns speak to underlying neural mechanisms? This question cannot be answered without referring to the experimental design; namely, we direct the ensuing discussion by treating the multivariate patterns as surprise (prediction error) responses shaped by the preceding anticipatory set. In other words, we view surprises as the discrepancy between the environmental cues and outcomes (Bubic, Cramon, & Schubotz, 2010; Bubic, Cramon, Jacobsen, Schröger, & Schubotz, 2009; Eshel, Tian, & Uchida, 2013; SanMiguel et al., 2013). Firstly, what does the anticipatory set entail, and what does it predict? For the high (updating) cue, the induced prediction is two-fold: (1) an expectation of representational updating and flexibility (Yu, Fitzgerald, & Friston, 2013); this imposes a low precision on the *ad hoc* integration of subsequent letters (Kahneman et al., 1992; Kahneman & Treisman, 1984), which may be monitored by a sustained task-set control signal (Dosenbach et al., 2008; 2006; Sestieri, Corbetta, Spadone, Romani, & Shulman, 2014). Note that anticipatory set does not generate concrete sensory predictions, which can only be represented upon the encoding cue, but will nuance the item-wise 'stickiness' and perhaps the rehearsal process involved. (2) The anticipatory set also entails a form of prediction about visual sensations, which engender the updating process. This, however, should be distinguished from classical perceptual inference (e.g., visual occlusion) in that the latter pertains to exteroception in an instant in time as opposed to a prospective sense, and that the perception being

called forth bears a non-specific nature. Therefore, it can be plausibly argued that the sensation being predicted is *categorical*, and the perceptual inference necessary to carry out an update is enabled.

By contrast, the low (maintenance) cue involves predictions in two parts: (1) it portends representational stability of working memory, thus placing a high precision on the sequence memory. (2) It is unlikely that any visual sensation would ensue, given an update is unlikely.

Briefly, the updating cue predicts that the working memory representation is to be hierarchically updated. That is, it entails a prospective sub-goal and necessitates that the encoding stimulus generates an inadequate/insufficient goal representation. The updating stimulus then completes the goal representation (sub-goal completion). Whereas, the goal representation is almost certainly informed by the encoding stimulus under the maintenance cue which predicts the probing stimuli. This speaks to a construct relevant to the idea of 'task set', which we will consider shortly.

Naturally, one may argue that working memory updating *per se* without explicit 'prediction' is in itself surprising. Thus, prediction error responses provided by pattern classification may have been confounded by those originated from the updating operation, making the responses not entirely attributable to anticipatory set. To prevent this, we modelled the effect of updating and maintenance using non-specific regressors to account for surprise effects irrelevant to that of the cues.

4.4.2. The prediction error

Having unpacked the predictive processes involved in the task, it is now more straightforward to outline the responses when predictions are violated. Specifically, violation of the prediction pertaining to the updating set is the *omission* of the

updating stimulus. Because the brain has no means to represent exactly what is used to achieve updating, prediction errors to stated sensations are considered minimal, possibly including those accounting for more abstract, categorical representations. In addition, the effect of any visual cue is removed due to the inclusion of non-specific updating/maintenance regressors. Another response that can be associated with omission is the processing of sustained set control, as well as that of rapid adaptive control (Dosenbach et al., 2006; 2008; Sestieri et al., 2014). In other words, omissions to the predicted outcomes may induce error responses pertaining to pure (top-down) prediction signals. In our case, they are reflected mainly in the fronto-parietal and cingulo-opercular systems. Indeed, den Ouden et al. (2009) demonstrated prediction error responses when the *absence* of visual stimuli is surprising. Our observation extends this finding to include error responses not directly related to sensations. Along the same line of evidence, Wacongne et al. (2011) used the auditory mismatch paradigm to show that omission to an anticipated auditory pattern revealed prediction signals from hierarchical predictions. They also concluded that higher-order predictions encompass multiple frontal and associative cortices, which is consistent with our findings. However, SanMiguel et al. (2013) pointed out, with a self-paced trigger-to-sound task, that both the timing and the identify of the sensation must be represented by the sensory system to formulate appropriate predictions, followed by the induction of error responses. Our finding is not restricted to the sensory system but nevertheless offers an alternative perspective.

The error response associated with the violation of maintenance set is relatively simple to interpret. Due to unexpected visual presentation, which is inherently salient (Ouden et al., 2010; Zink, Pagnoni, Martin, Dhamala, & Berns, 2003), sensory responses beyond the predicted, non-specific effect of updating are pronounced. This

explains widespread visual involvement found in the deviation pattern. The sensory saliency, perhaps along with unexpected allocation of cognitive capacity for updating, may account for the pattern comprising the meso-thalamo-striatal network (Baier et al., 2010; D'Ardenne et al., 2012).

Based on the findings so far, as well as the notion that the omission responses reflect prediction error signals, we argue that predicting and implementing memory updating in the context of cognitive meta-stability have distinct neural substrates. Our data speaks to a putative hierarchical organisation of these substrates.

4.4.3. A free-energy perspective

The Bayesian principle considers neuronal computations to represent the cause of environmental states and the uncertainty of these states. These states are not stationary but are rather represented in terms of their motion. The motion refers to a trajectory through state-space that contains the variables responsible for generating sensory data. In other words, neuronal representations of states encode prospective states along the trajectory. Working memory clearly falls into this category. For instance, in a simple delayed-response paradigm, the subject is cued to match a target. Once the cue is extinguished, the encoded memory conforms to attractor dynamics, towards a basin encoding the target. In a more complex setting, such as our updating task, the trajectory may follow variable dynamic regimes based on the uncertainty afforded by cues. The maintenance cue, on the one hand, may provide an energy landscape resembling the simple case stated. On the other hand, the updating cue may induced a wider dynamic regime in which set-switching can be accomplished given multiple attractor states. A useful concept here is the notion of a winnerless competition – or a stable heteroclinic channel (Bick & Rabinovich, 2009;

Bick, Rabinovich, & Rabinovich, 2010; Rabinovich, Huerta, & Laurent, 2008), and can be associated with an adaptive cognition-stable set control complex known as the FPN-CON network (fronto-parietal-cingulo-opercularis network; see below and Figure 4.4). In light of the relationship between omission and the predictive coding hypothesis, one may argue that the omission response is a representation of prospective predictive process.

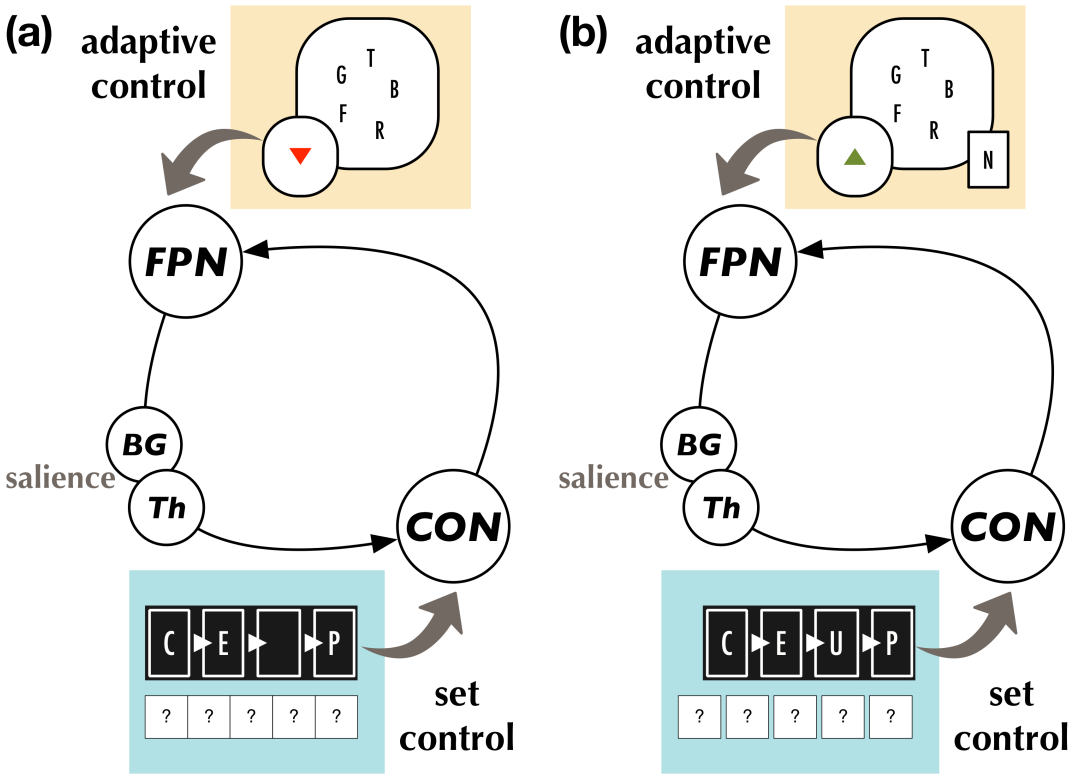

Figure 4.4 A proposed model of information processing flow in the FPN-CON network. This model is based on the hypothesis of Dosenbach et al. [Dosenbach, N. U. F., et al. (2008). A dual-network architecture of top-down control. Trends in Cognitive Sciences, 12(3), 99–105]. The CON provides task parameter, and the framework in which information is processed. This may entail a flow control governing trial transition, and the events within trial. Also, it may also provide a level of information binding. The FPN is more specific for stimulus associations. In the face of salient events, the basal ganglia report the saliency; the ensuing update may take place in a rapid/stimulus-dependent manner (updating a sequence), or a change in slow dynamics (e.g., predictive cue-induced trial transition). FPN, fronto-parietal network; CON, cingulo-

opercular network; BG, basal ganglia; Th, thalamus; (blue shade) experimental phases: C, cue; E, encoding; U, updating; P, probing.

4.4.4. Regional-specific functional implications

Both the fronto-parietal and the cingulo-operculum networks are implicated in working memory (Gordon, Stollstorff, & Vaidya, 2012; Repovš & Barch, 2012), suggesting intrinsic connectivity and coherent specialisation in task-dependent information processing (Gordon et al., 2012). In particular, the fronto-parietal network is proposed to support attentional set, based on graph theory (intrinsic connectivity networks; Markett et al., 2013) and delayed-response paradigms (Corbetta & Shulman, 2002). The neural mechanism of attentional set overlaps functionally with that of working memory, which enables the top-down selection of behaviourally relevant stimuli, and is adaptive to unexpected, salient events (Owen et al., 1993). Such set may involve divided attention (Baddeley, 2012; 2007) to prepare for multiple stimulus selection. Santangelo and Macaluso (2013) employed a divided attention task, in a load-dependent delayed-response task. The authors demonstrated that the bilateral intra-parietal cortices were more activated under divided attention, and under incremental working memory load, which indicated an effect of unnecessary storage in the parietal lobe (unnecessary in the sense that memory items are behaviourally irrelevant to the performance of the attentional task; see McNab & Klingberg, 2008; Vogel et al., 2005). We argue that the non-specific storage in the parietal cortex forms the basis of cognitive flexibility that allows multiple representations, which can be subject to multiple selections via the prefrontal-basal ganglia network (M. J. Frank & O'Reilly, 2006; R. C. O'Reilly & Frank, 2006). One may accordingly reason that the anticipatory set about updating, and thus the error

response on omission, entails preparation for divided attention under non-specific storage capacity implemented via the fronto-parietal network. By contrast, under the maintenance set, in which representational stability is expected, access to non-specific storage is not required. Instead, the maintenance of high-fidelity working memory information may call upon domain-specific sensory cortices (Sreenivasan et al., 2014).

Dosenbach et al. (2006) hypothesised a system for general task control, which comprises the fronto-parietal and cingulo-opercular networks (Dosenbach et al., 2006; 2008). The system is proposed to exhibit (1) sustained task set-maintenance signals; (2) trial-specific transients in response to cues; and (3) error-related feedback to optimise task set. It may constitute a core resource, limited in capacity, shared across concurrent tasks (Dosenbach et al., 2006). Therefore, it is plausible that the same system underpins 'sub-goaling' (Fincham, Carter, van Veen, Stenger, & Anderson, 2002; Miller & Cohen, 2001; Oosterwijk et al., 2012). Specifically, set-maintenance signals are associated with the cingulo-opercular network, including anterior insula, frontal opercularis, dorsal ACC, and medial superior frontal cortex. Whereas, the fast adaptive control of sensory signals is proposed to implicate the fronto-parietal network. Error-related responses are found in both systems (Dosenbach et al., 2006). This hypothesis sits well with our interpretation that the omission response speaks to two levels of prediction – stimulus- and task-dependent. The response therefore suggests a reconfiguration of task set.

4.4.5. Neurobehavioural correlations

Increased amplitude of deviation BOLD responses that predicts subsequent response speed may indicate a faster retrieval at the probing phase. This can be

related to whether memory items are within the current focus of attention – an accompanying effect of surprise (saliency), which is arousing and causes state-switching (Zink et al., 2003). Namely, we speculate that the amplitude-speed relationship may be an intrinsic property of behaviour-relevant prediction error in the context of delayed tests. More formally, the deviation amplitude may report memory accessibility (McElree, 1998). This is in light of the interference theory. Specifically, the proactive interference – whereby recall of the newly acquired information interacts with existing information, especially in a delayed test (Tehan & Humphreys, 1995; 1996) – is introduced by the encoded item *in situ* as per (unanticipated) updating items.

Another possibility is that the update causes the entire sequence in working memory to be reconfigured (Kessler & Meiran, 2006), resulting in a refreshed representational state. This is plausible when a subvocal rehearsal process is involved (which is a common strategy for verbal stimuli), thereby improving memory availability (McElree, 1998; 2001).

Overall, these reflect the neural re-instantiation of goal representation that is to be recognised in a delayed probe. The same line of reasoning can also be applied to the omission condition to interpret the lack of detection of amplitude-dependent speed gain. Specifically, anticipation of an updating stimulus precedes the updating event, which is argued to induce a meta-stable attractor state that trades (representational) stability for flexibility (see Chapter 3; Yu et al., 2013). The surprise then follows that no expected stimulus is presented; therefore there is no induced sensory salience. The (noisy) representations are still in the focus of attention in working memory, which means they are available, but they have no means to modulate memory accessibility. As a consequence, in processing the

omission, the brain does not have appropriate cues to nuance the accessibility or availability of memory retrospectively; it is therefore relatively independent of the subsequent retrieval speed.

4.4.6. Conclusions

In summary, we demonstrated that the multivariate patterns of omission and deviation responses are regionally segregated and dissociable. This suggests distinct neural mechanisms of surprise (prediction error) with respect to prior beliefs. The prior belief (or anticipatory set) may entail neural instantiations at both concrete and abstract levels, which speak to the idea of identity (perception and selection) and structural conformation (task-set and stimulus binding). This idea sits comfortably with the model proposed by Dosenbach et al. (2006, 2008), in which the FPN and the CON control adaptive, stimulus-dependent cognition and stable task set-maintenance, respectively. The omission pattern, which reflects prediction signals, identified both systems, showing two levels of predictive control pertaining to working memory. The deviation pattern, on the other hand, speaks to a surprise in the high-fidelity memory representations maintained in the sensory cortex (Sreenivasan et al., 2014). This work offers an interpretation from the free-energy perspective that working memory may involve slow, itinerant dynamics within the FPN-CON network under the generalised predictive coding framework, which stands for robustness-adaptiveness trade-off in working memory maintenance and manipulation. We reported that, although surprise impairs overall reaction time, within-subject surprise response predicted faster reaction time without trading off accuracy.

Chapter 5. Causal models of anticipatory processes in working memory

This chapter shows that working memory, compared to perceptual inference and motor responses, is shaped by predictive cues affording contingencies that entail representational flexibility and stability. The underlying question being investigated appeals to hierarchical inference in the brain: whether cortical connectivity encodes beliefs about fluctuations in working memory representation and whether improbable outcomes under those beliefs are conveyed back to the cortical hierarchy. We conducted a working memory task in which updating or maintaining memory items was contingent upon a preceding predictive cue. The cue induces an anticipatory set that is maintained until the realisation of the required working memory operation. The cue was probabilistic therefore surprising outcomes may ensue. We used dynamic causal modelling (DCM) to model Blood-oxygen-level dependent (BOLD) responses where the anticipatory set and surprise entered as forward, backward, or local recurrent modulations. Bayesian model selection (BMS) and family-level inference revealed that the anticipatory set modulates backward connections, whereas surprise modulates forward and local recurrent connections. Furthermore, statistical inference based on parameter estimates of the optimal model showed that the anticipatory set exerts differential modulatory effects across two working memory-related circuits. Our results suggest that working memory processing may follow the principle of hierarchical inference and that information flow is contingent upon top-down belief.

5.1. Introduction

Our brain benefits from representing the causal structure of the environmental states that generate its sensations, thereby allowing it to react appropriately to a sensory input (e.g., perceptual inference; Rahnev et al., 2011; Summerfield et al., 2006), and maintain these representations as necessary. This notion confers working memory with properties of predictive codes (Rao & Ballard, 1999) or hierarchical inference (Friston, 2008) in an anticipatory sense. To put the notion to the test requires a demonstration of cortical message passing in the working memory network. However, few studies have addressed this question (but see Bollinger et al., 2010; Rahnev et al., 2011). Moreover, evidence with regard to predictive codes that enhance perceptual working memory performance has been reported to involve domain-specific sensory cortices (Bollinger et al., 2010; Rahnev et al., 2011; Summerfield et al., 2006). But working memory representation is not merely about sensory codes; rather, it represents future goal variables with nuanced stability and flexibility (Miller & Cohen, 2001; Sreenivasan et al., 2014). It therefore raises the question as to whether cortical hierarchies may represent such biased beliefs (or anticipatory sets) about the necessary representational stability or flexibility.

Previously, we have shown that the neural correlates of anticipatory set are in the striatum and parieto-occipital regions. These regions showed an elevated activity when an imminent update to working memory content was more predictable. We therefore regarded activity in these regions as reflecting prediction signals that provide top-down modulations. However, in a different study, another line of evidence emerged that the fronto-parietal, as well as the cingulo-opercular networks may also reflect prediction signals derived from the anticipatory set. This was

revealed, using machine learning techniques (Schrouff et al., 2013), by the multivariate pattern of neural responses during which an anticipated update was omitted. It has been argued that omission to an expected sensory event elicit prediction error responses that reflect the prediction signals (SanMiguel et al., 2013; Wacongne et al., 2011). This notion appears to be a viable consequence of message-passing under hierarchical inferences in the brain (Bastos et al., 2012; Friston, 2008). Briefly, we observed the prediction-related responses that have distinct cortical substrates. This discrepancy may be ostensible yet and speak to the information flow during hierarchical inference that can be disambiguated by means of effective connectivity.

Dynamic causal modelling (DCM) is a hypothesis-driven technique for neuronal system identification (Friston et al., 2003). By using a two-layered forward model, it allows inferences to be made at a neuronal level in terms of inter-regional coupling due to experimentally designed perturbations, i.e., effective connectivity. Multiple plausible hypotheses are then motivated by assuming the (fMRI) data observed were generated by a certain connectional configuration, in which stimulus-bound and stimulus-/trial-free factors may drive or modulate the neuronal responses. Parameters of these models are estimated through a Bayesian inversion scheme during which individual approximates of model log-evidence are derived (Friston et al., 2007). Bayesian model selection (Stephan et al., 2009a) relies on these log-evidence estimates to determine which model, or model family (Penny et al., 2010), is optimal.

Our principal hypothesis, which stems from the predictive coding framework, is that the anticipatory set is a top-down modulation exerted on backward connections that encode prediction signals, whereas surprise – i.e., the violation of prediction –

modulates forward connections that drive higher regions with prediction error signals. These modulatory effects entered a DCM comprised of regions that mediate flexible updating (Baier et al., 2010; McNab & Klingberg, 2008) and robust maintenance (Vogel et al., 2005) of working memory respectively.

5.2. Methods

5.2.1. Pre-processing of functional data

Data pre-processing and DCM specification were carried out using SPM12 (Statistical Parametric Mapping; Wellcome Trust Centre for Neuroimaging, London, UK). Pre-processing of functional images included correction of geometric distortions due to B0 magnetic field inhomogeneity using pre-acquired field maps (Hutton et al., 2002; Jezzard & Balaban, 1995), inter-scan realignment via affine rigid-body registration to model head motion, slice-timing correction, coregistration with respect to anatomical images, normalisation to MRI space based on the anatomical normalisation parameters, interpolation to voxel size of 2 x 2 x 2 mm^3, and, finally, smoothing with a Gaussian kernel of 4 mm FWHM (full-width at half-maximum).

5.2.2. General linear model

As part of the preprocessing prior to the DCM analysis (Figure 5.1), a GLM was set up for subsequent region-of-interest (ROI) selection and the adjustment of data features using a reduced model. In addition, the task-specific temporal profiles within the design matrix served as the input specification to the DCMs. The design matrix consisted of eight task-related regressors. The UPD-set and MAI-set

regressors represented sustained activity induced by the predictive cues and were modelled as a boxcar function of 6 s. The set regressors extended from the onset of cue stimuli to the offset of the action stimuli. The UPD and MAI regressors were impulse response functions at the onsets of the action phase, indicating non-specific neural transients associated with updating and maintenance processes, respectively. By 'non-specific', it means the two regressors were irrespective of the cue-action interactions. The interactions were modelled by the omission (Om) and deviation (Dv) regressors, which were both impulse response functions at the onset of maintenance action under UPD-set and updating action under MAI-set, respectively. An impulse stimulus function was used to model the effect of set-switching (Sw) at the onsets of predictive cues. This was to account for possible nuisances due to trial transition, although the UPD-set and MAI-set may have been disengaged upon the offset of the action phase, owing to the nature of the underlying anticipatory process. Finally, all visual transients of little interest with regard to the DCM analysis were modelled in a single non-specific regressor (NS). This included the predictive cues, the encoding cues, and the probing cues. The visual onsets of action cues may be included in the NS regressor if the subject had made an error in a specific trial. Accordingly, task-related effects of interest (set, action, and interactions) were not modelled for error trials. All task-related regressors were convolved with the canonical haemodynamic response function to create haemodynamic regressors.

Motion and physiological regressors were also included to factor out non-specific nuisances in the BOLD responses. The motion regressors were derived from the realignment procedure, parameterised by 3 translations (along *x*-, *y*-, and *z*-axis) and 3 rotations (pitch, roll, and yaw). The physiological regressors reflected peripheral readings of heart rate and respiration, using pulse-oximeter and respiratory

belt readings, which comprised of six cardiac, six respiratory regressors, as well as two regressors for heart rate change and change in respiratory volume (Hutton et al., 2011).

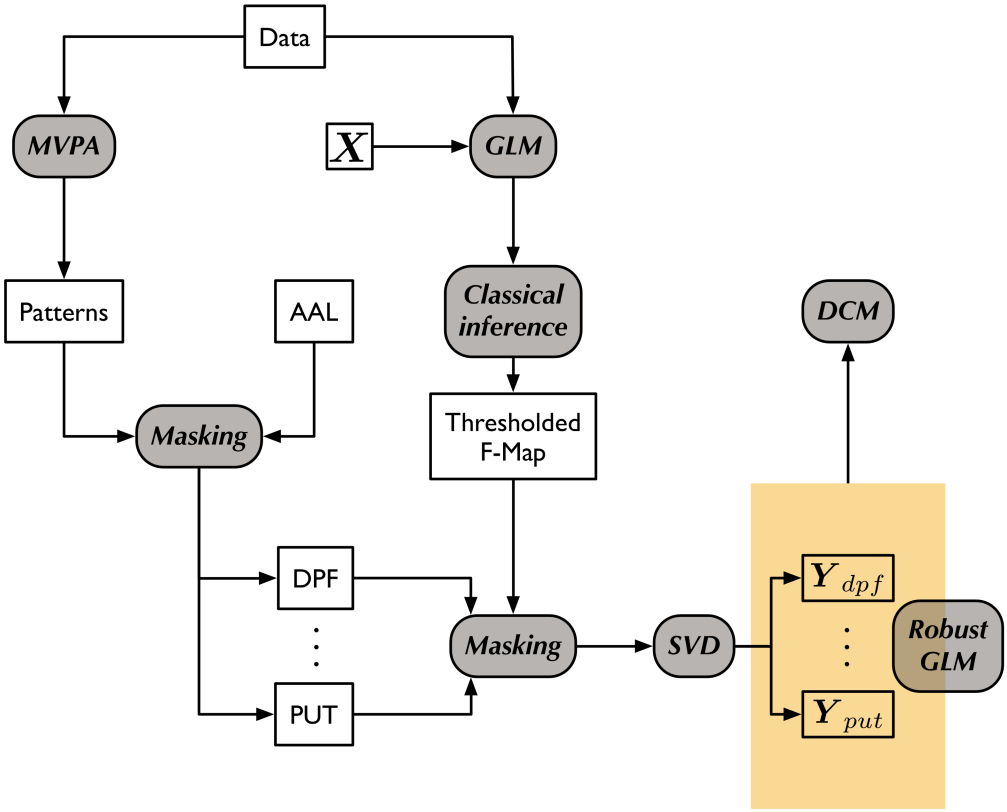

Figure 5.1 Preprocessing pipeline prior to the DCM analysis. This diagram outlines the procedures taken to extract the fMRI time series for subsequent DCM analysis. Firstly, the data were modelled using a general linear model with the design matrix containing regressors that encode all experimental manipulations and nuisance variables. An F-contrast was then specified to test the 'effect of interest' (i.e., with multidimensional contrasts to test multiple linear hypotheses for the experimental effects) within each subject. The resultant statistics were thresholded and corrected for multiple comparisons (family-wise error rate, $p < 0.05$) to generate individual F-maps. Voxels showing significant effects of interest in the F-map were regarded as candidate regions of interest (ROI) for DCM. These voxels were further filtered using information provided from previous MVPA results and region-specific AAL masks. The multivariate patterns associated with the omission (Om) event represent stimulus-free voxel responses and were considered to reflect pure prediction signals (see Chapter 4). Therefore, these voxels were considered appropriate to base our DCM analysis, in which connectivity changes with regard to predictions and the violation of predictions are of interest. Subsets of voxels of

individual Om patterns were selected based on the following four AAL regions respectively: the right middle frontal cortex (DPF), the left putamen (PUT), the left inferior parietal cortex (IPS), and the left inferior occipital cortex (Vix). As a consequence, four Om-informed ROI masks were created. This was followed by masking the F-maps accordingly using the ROI mask. The temporal mode of the fMRI time series within each mask was computed using the principal eigenvariate.

5.2.3. Regions of interest (ROI)

In DCM for fMRI, ROIs were specified in a hypothesis-driven, regionally specific manner, followed by extracting the temporal mode of multi-voxel BOLD time series using eigendecomposition. To restrict ROI selection to voxels exhibiting task-dependent responses that are relevant to the DCM hypothesis, we applied a two-level masking procedure (Figure 5.1). First, for each subject, a reduced model was specified by creating an F-contrast to include only eight task-related effects in the design matrix (UPD-set, MAI-set, UPD, MAI, NS, Om, Dv, and Sw). This was followed by a classical inference which tested voxel-wise F statistics with a thresholding *p*-value of 0.05 (corrected for family-wise error rate). This created a thresholded F-map, from which ROIs were selected using secondary masks. The secondary masks served a region-defining purpose, which, according to our hypothesis, included the DPF, PUT, IPS, and Vix masks. These abbreviations stand for the right dorsolateral prefrontal cortex, the left putamen, the left inferior parietal cortex, and the left visual cortex, respectively. The construction of the secondary masks depended on two primary masks, the multivariate patterns from classifier weights (see Chapter 4) and the AAL atlas (Tzourio-Mazoyer et al., 2002). Specifically, the weight maps contained Om-informative voxels that can be used to motivate the ROI specification. Because the Om patterns represent stimulus-free voxel responses and underlie pure prediction signals they are considered appropriate

for DCMs that address connectivity changes due to predictions (anticipatory sets) and the violations of predictions (surprises). Details as to how the classifier weight maps were derived using MVPA was described in Chapter 4. In short, each ROI was the union of the Om-informed mask and designated AAL masks, excluding sub-F-threshold voxels.

The specific AAL regions used for the DPF, PUT, IPS, and Vix masks were the right middle frontal cortex, the left putamen, the left inferior parietal cortex, and the left inferior occipital cortex, respectively. Previously, we used multi-region AAL masks for the MVPA study. This, however, may not be feasible to motivate haemodynamic state variables for the DCM analysis. Because the DCM models the haemodynamic parameters on a region-by-region basis (Stephan et al., 2007b) and the inter-regional haemodynamic responses may vary irrespective of the underlying neuronal activity, conflating spatially distant ROIs into a single temporal mode may confound the underlying haemodynamic model, thereby possibly resulting in an improper model estimation and subsequence inferences.

fMRI time series were isolated by combining the two levels of masking, followed by adjusting the remaining voxel data with respect to the null space, i.e., using the reduced design matrix. Next, the temporal mode of each ROI time series in terms of their first eigenvariate was calculated by means of eigendecomposition (SVD).

5.2.4. Robust general linear model

An optional noise modelling scheme was applied when spurious spikes (amplitude > 5 standard deviations) were present in the ROI time series. This was

rare but was nonetheless observed in one of the subjects. The approach followed the Robust General Linear Model (RGLM; Penny, Kilner, & Blankenburg, 2007) in which the noise is modelled with a mixture of Gaussians. This allows different data points to have different levels of noise and provides robust estimation of regression coefficients via a weighted least square approach (Penny et al., 2007). In practice, the spikes were modelled as high-variance outliers within a 2-component mixture of Gaussians noise model, as opposed to the standard one-component. The eigenvariate was modelled using the RGLM with the reduced design matrix described previously. The outlier component was then subtracted from the original time series.

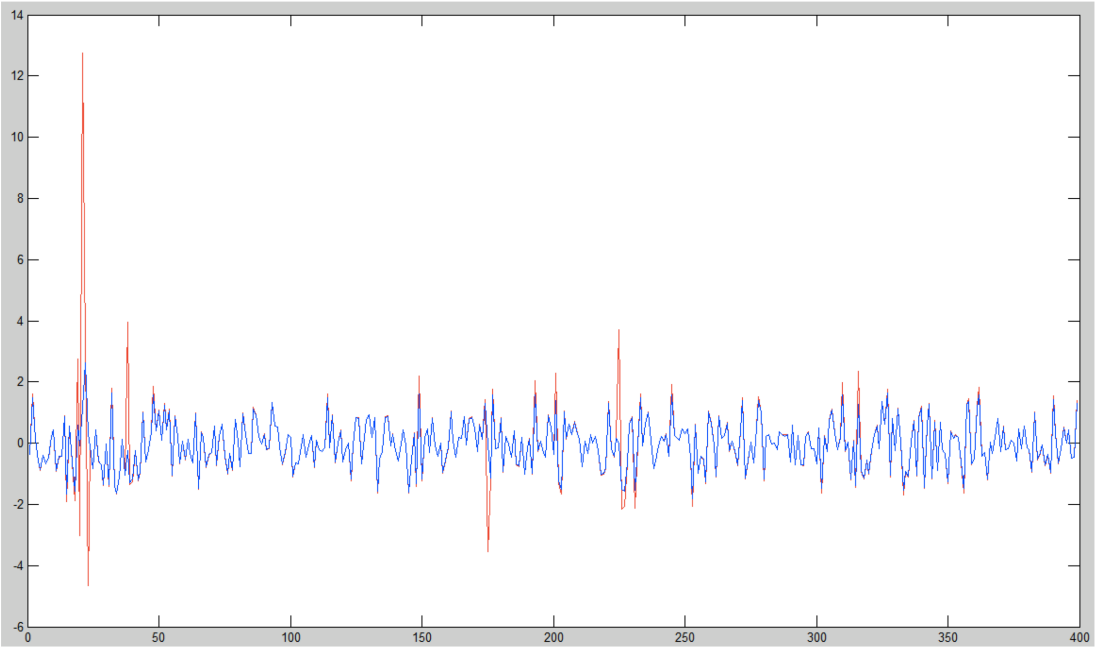

Figure 5.2 Spike removal with the Robust General Linear Model (RGLM). This figure illustrates the presence of spurious spikes in the BOLD response of Vix in one subject. RGLM uses an enhanced noise model that treats noise as being generated by a mixture of Guassians. Here, we used a two-component mixture. A high variance noise component (red) was detected by the model, which was subsequently subtracted from the signal (blue line is the result).

5.2.5. Dynamic causal modelling

DCM for fMRI (DCM12) treat brain responses as deterministic consequences of regional (inter-state) coupling prescribed by a set of dynamic equations under the influence of experimental perturbations (Friston et al., 2003; later versions posit stochastic dynamics but these are beyond the scope of this thesis). It has model parameters that (1) mediate external influences on the states, e.g., exteroceptions; (2) exert influences amongst states in the absence of external inputs; (3) allow intrinsic coupling to be modulated by external inputs. Under the DCM framework, similar to that of other structural models, multiple hypotheses may be motivated in terms of model alternatives to construct a 'model space'. These models can then be inverted and compared as competing hypothesis under a Bayesian framework (Penny et al., 2004; Stephan et al., 2009a), which enables different levels of inference (Penny et al., 2010).

5.2.6. Model space

Our model space was defined on the basis of a fixed structural configuration that defines intrinsic coupling, as well as a set of driving inputs (Friston et al., 2003). We assumed a hierarchical organisation between the four neuronal states: DPF, PUT, IPS, and Vix. The hierarchy was defined such that DPF is at the highest level, PUT and IPS intermediate, and Vix the lowest. This means all regions are inter-connected, except for DPF and Vix. The driving inputs were UPD, MAI, NS, and Sw regressors entering Vix. The temporal profiles of these inputs were as specified in the earlier design matrix (Figure 5.3a). Although we are only interested in the effect of anticipatory sets and the violations of set, other experimental effects of no interest

were also included. The reason for including all experimentally designed effects is to reduce the residual error during model fitting.

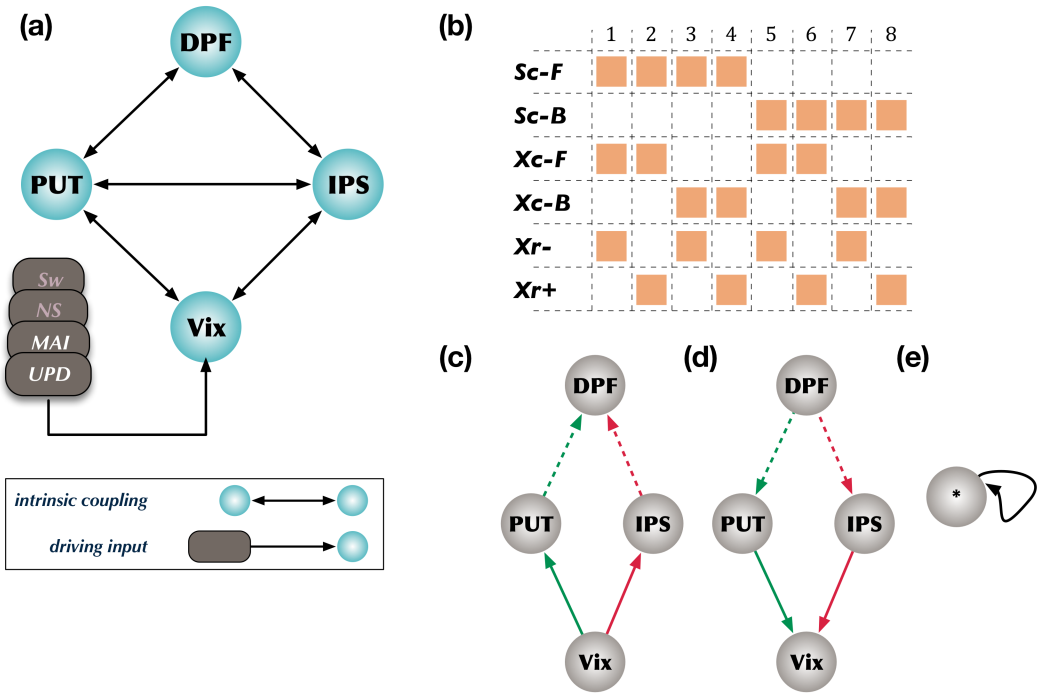

Figure 5.3 DCM specification and model space. ***a***, the basic model architecture illustrating a hierarchical organisation with four constituent regions (high to low): the right dorsolateral prefrontal cortex (DPF), the left putamen (PUT)/inferior parietal cortex (IPS), and the left inferior occipital cortex (Vix). All regions are reciprocally connected, except DPF and Vix. The strength of these connections represents intrinsic coupling in the absence of experimental perturbations. Stimulus-bound experimental effects, UPD, MAI, NS, and Sw, entered Vix as driving inputs. ***b***, model space defined by three factors: set modulations on inter-regional coupling, surprise modulations on inter-regional coupling, and intra-regional surprise modulations. Each factor has two levels; all combinations yielded a model space of eight models. Coloured squares indicate the types of modulatory effects under each model. The modulatory effects can be further divided into that of anticipatory sets (Sc) and surprises (Xc/Xr): Sc-F, inter-regional coupling-forward; Sc-B, inter-regional coupling-backward; Xc-F, inter-regional coupling-forward; Xc-B, inter-regional coupling-backward; Xr^-, intra-regional recurrent modulations absent; Xr^+, intra-regional recurrent modulations present. Note that the 'Sc' entails updating and maintenance sets and 'Xc/Xr' entails omissions and deviations. There is no variation within Sc or Xc/Xr, all levels of set and surprise entered the model concurrently. ***c-e***, illustrating inter-regional forward coupling, backward coupling, and intra-regional recurrent connection, respectively. ***c-d***, the inter-regional connections can be subdivided into factors so

that they are related to different hierarchies (e.g., solid arrows versus dashed arrows) or to different routes (e.g., the green route that uses PUT as a waypoint versus the red route that uses IPS as a waypoint). Such division allows classical inferences on the parameter estimates using a factorial design (e.g., ANOVA).

The principle hypotheses associated with the model space pertained to the connections on which the set (UPD-set and MAI-set) and surprise (Om and Dv) exerted modulatory influences. They were systematically constructed along the following dimensions. Firstly, the UPD-set and MAI-set were treated as top-down modulations. This means they may modulate the strength of DPF→PUT/IPS and PUT/UPS→Vix connections. However, an anatomically plausible alternative exists: the top-down modulations exerted pre-synaptic influences to the dendritic tree of the lower region, thereby tuning the output (i.e., forward) of the lower region (Penny et al., 2004). In other words, it is equally possible that the set modulated the Vix→PUT/IPS and PUT/IPS→DPF connections. Secondly, the surprise (Om and Dv) is hypothesised to modulate forward connections if they encode prediction error signals (Friston & Friston, 2005) and speak to the notion of inter-regional model adjustment (Garrido et al., 2008). That is, Om and Dv are hypothesised to modulate the Vix→PUT/IPS and PUT/IPS→DPF connections. To design competing model alternatives, we allowed Dv and Om to modulate the backward connections: DPF→PUT/IPS and PUT/IPS→Vix. Finally, predictions may contribute to intra-regional adjustment to the statistical regularity of the environment, hence a surprise can be regarded as adaptation modulations that exerts modulatory effects on recurrent connections (Garrido et al., 2008). In this case, Dv and Om entered the models as modulatory inputs in the self-connections: Vix→Vix, PUT→PUT, IPS→IPS, and DPF→DPF. This was tested against a set of null models, i.e., those

without recurrent connections. To summarise, our model space was generated by asking the following questions (the figure inside the brackets indicate the number of levels):

1. Is the anticipatory set a top-down modulation on the forward or backward connections? [2]
2. Does the surprise represent prediction error signals and inter-regional model adjustment via forward connections? [2]
3. Does the surprise have a role in modulating intra-regional adaptation? [2]

Overall, the combination of the three dimensions, each with two levels of variations, gave rise to the model space of eight models (Figure 5.3b). Each model was inverted to obtain parameter estimates using a variational free energy minimisation scheme under the Laplace assumption (Friston et al., 2007). This means the coupling strengths are expressed in terms of their conditional expectations and covariances (Friston et al., 2003).

5.2.7. Bayesian model comparison

The questions we raised above to motivate our model space can be tested using Bayesian model comparison with appropriate model space partitioning. Model space partitioning creates 'families' or comparison sets in which models within one family share a common structural aspect that the other families do not have (e.g., one has recurrent modulations and the other has none). Inferences can then be made with regard to the commonality while ignoring idiosyncratic model structures within each family. This is called family-level inference (Penny et al., 2010), the underlying concept closely resembles factorial experimental designs in psychology where data from all cells are summarised to assess the size of main effects. Under the

assumption of random effects, we tested (1) Sc-F vs Sc-B (inter-regional set modulation, forward versus backward; model 1 - 4 versus model 5 - 8); (2) Xc-F vs Xc-B (surprise modulations on inter-regional coupling, forward versus backward; model 1, 2, 5, 6 versus model 3, 4, 7, 8); (3) Xr^- vs Xr^+ (surprise modulations on intra-regional recurrent connection, absence versus presence; model 1, 3, 5, 7 versus model 2, 4, 6, 8). The results were summarised in terms of family exceedance probabilities. In addition, model-wise random-effect Bayesian model comparison (Stephan et al., 2009a) was applied to determine whether an optimal model exists at the group level. The results were reported in terms of model exceedance probabilities.

5.2.8. Classical inferences with DCM parameter estimates

For the optimal model, three separate statistical analyses were performed at the group level on the parameter estimates: (1) set-related inter-regional couplings, (2) surprise-related inter-regional couplings, and (3) surprise-related recurrent connections. For the respective set- and surprise-related inter-regional couplings, their modulatory effects were associated with eight parameter estimates: that between Vix and PUT/IPS and that between DPF and PUT/IPS, multiplied by the two levels of update predictability (i.e., UPD-/MAI-set or Dv/Om). This allows the parameter estimates to be tested in a factorial design. The factors are defined by the update predictability (PR), hierarchy (HY), and route (RO). For example, Figure 5.3c illustrates that the parameters for Sc-B in model 6 can be factorised with respect to high (dashed arrows; PUT/UPS→DPF connections) and low (solid arrows; Vix→PUT/IPS connections) hierarchies, or to PUT- (green arrows; Vix→PUT→DPF pathway) and IPS-routes (red arrows; Vix→IPS→DPF pathway).

The same factors then repeat across MAI-set and UPD-set. Likewise, the parameters for Xc-F in models 6 were treated according to Figure 5.3d. Analysis of variance (ANOVA) was used to test the main effect of PR, HY, and RO, as well as the interactions amongst these factors.

As for the parameter estimates for the recurrent connectivity, one-sample t tests were used to detect significant changes in coupling strengths in the four regions under Om and Dv. Bonferroni corrections were applied to control the false positive rate due to multiple comparisons (overall threshold $\alpha = 0.012$).

5.3. Results

5.3.1. Bayesian model comparison for family level inferences

Family-wise model comparisons revealed that the anticipatory sets (UPD-set and MAI-set) were more likely to modulate backward connections, rather than the forward connections. The Sc-B family showed a dominant exceedance probability (96.26%; right bar, Figure 5.4a), than the Sc-F family (3.74%; left bar, Figure 5.4a). Forward, but not backward, connections were modulated when the anticipatory sets were violated by surprising outcomes (Om or Dv). This was revealed in the inter-regional coupling Xc-F (exceedance probability, 98.56%; left bar, Figure 5.4b) over the Xc-B family (exceedance probability, 1.44%; right bar, Figure 5.4b). We also tested the hypothesis as to whether or not the surprises exerted modulatory effects over the inter-regional recurrent connectivity. Our result indicated that the models in which such recurrent modulations were present can better predict the observed BOLD responses. The model selection favoured the Xr^{+} family (exceedance

probability, 99.69%; right bar, Figure 5.4c) over the Xr^- family (exceedance probability, 0.31%; left bar, Figure 5.4c).

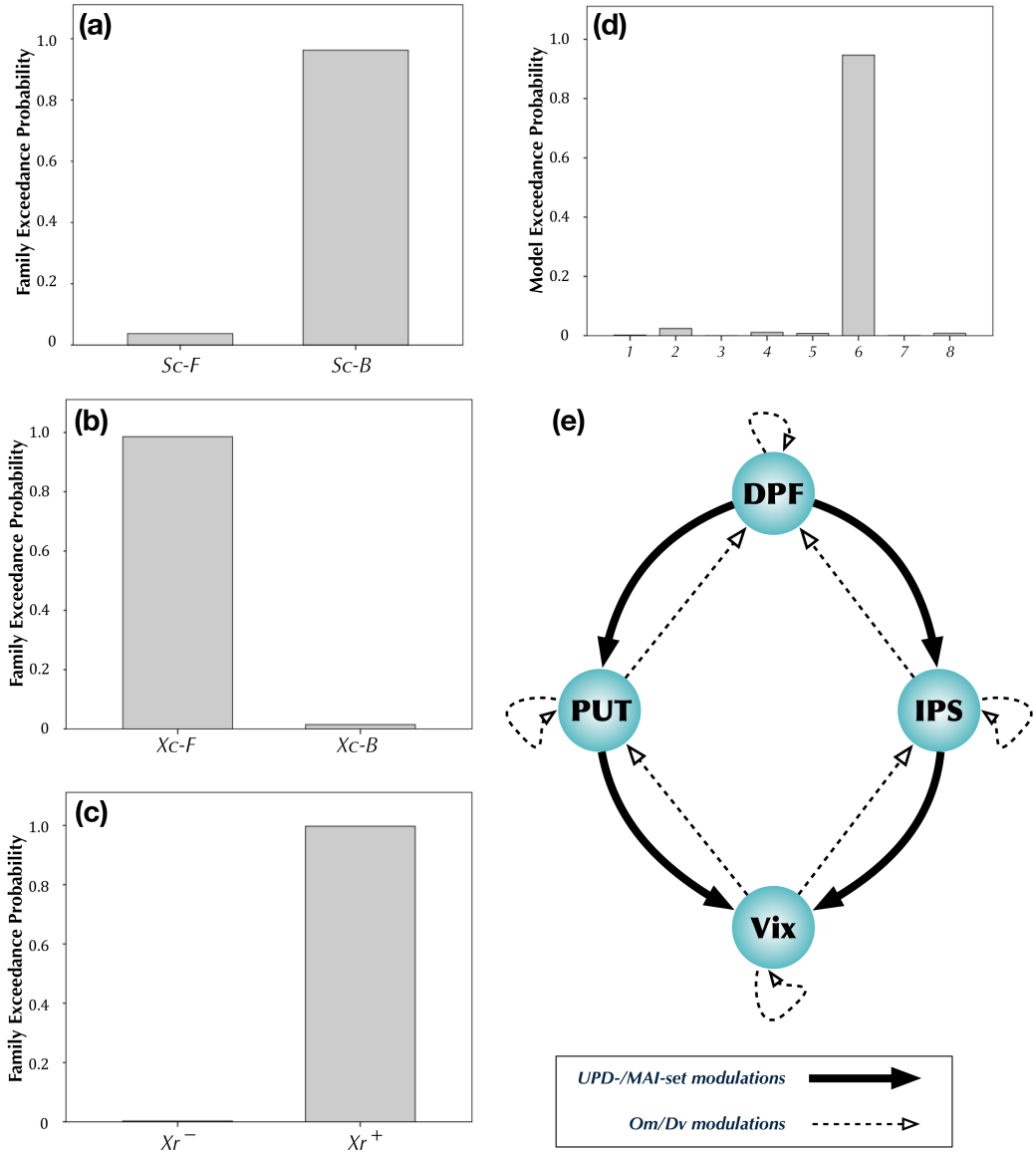

Figure 5.4 Bayesian model comparisons at family and model levels. ***a***, family-wise comparison for the directionality in inter-regional set modulations; Sc-F, forward; Sc-B, backward. ***b***, family-wise comparison for the directionality in inter-regional surprise modulations; Xc-F, forward; Xc-B, backward. ***c***, family-wise comparison for the presence of inter-regional surprise modulations; Xr^-, modulation absent; Xr^+, modulation present. ***d***, Bayesian model comparisons across individual models across all subjects. Model 6 is the optimal under random effects inference. ***e***, A schematic illustrating all modulatory effects in the

optimal model. The thick arrows indicate set modulations, whilst the dashed arrows indicated surprise modulations. For each arrow, the two levels of set/surprise effects entered concurrently.

5.3.2. Comparing individual models

Comparing individual models under the assumption of random effects revealed model 6 as the optimal model (exceedance probability, 94.64%; Figure 5.4d), in which the set modulations were exerted on backward connections, the surprise modulations on forward connections. In addition, the inter-regional recurrent connections were modulated by the surprise effects (Figure 5.4e).

5.3.3. Bayesian parameter averaging

To better illustrate the optimal model, Bayesian averaging of the posterior parameter estimates in model 6 was performed to summarise parameter estimates across subjects (Table 5.1).

Table 5.1 DCM parameter estimates of the intrinsic and modulatory connectivity derived from Bayesian averaging of the optimal models across all subjects. Figures in parentheses indicate standard deviations.

	intrinsic connectivity	**modulatory connectivity**			
		UPD-set	*MAI-set*	*Om*	*Dv*
PUT → Vix	0.2212 (0.0151)	0.0671 (0.0305)	-0.1802 (0.0299)		
IPS → Vix	-0.0776 (0.0126)	-0.0097 (0.0228)	0.2366 (0.0213)		
DPF → PUT	-0.4914 (0.0147)	0.0026 (0.0289)	-0.1099 (0.0276)		
DPF → IPS	-0.1543 (0.0214)	-0.0318 (0.0264)	0.1011 (0.0438)		
Vix → PUT	0.6280 (0.0256)			-1.7466 (0.1958)	-0.1521 (0.2049)
Vix → IPS	0.3438 (0.0321)			0.1862 (0.2545)	-0.0540 (0.1942)
PUT → DPF	0.2633 (0.0174)			0.3999 (0.2458)	-0.5229 (0.1754)
IPS → DPF	0.3420 (0.0184)			-0.0502 (0.1820)	0.2364 (0.1229)
PUT → IPS	0.6672 (0.0272)				
IPS → PUT	0.1009 (0.0118)				
Vix → Vix	-0.3673 (0.0391)			-0.4353 (0.2442)	-0.5044 (0.2952)
PUT → PUT	-0.5281 (0.0292)			-1.3929 (0.2473)	-0.7107 (0.2495)
IPS → IPS	-0.3380 (0.0326)			-0.8815 (0.3119)	-1.4069 (0.3361)
DPF → DPF	-0.1365 (0.0274)			-1.2048 (0.2636)	-1.2230 (0.2398)

5.3.4. Statistical analysis

Repeated-measure ANOVA for the inter-regional set modulations (Sc-B) revealed a significant PR x RO interaction ($F_{1,16}$ = 6.213; p = 0.024). No significant main effect was detected for PR ($F_{1,16}$ = 0.066; p = 0.801), HY ($F_{1,16}$ = 0.952; p = 0.351), and RO ($F_{1,16}$ = 2.435; p = 0.138), nor was the interaction for PR x HY ($F_{1,16}$ = 2.760; p = 0.116), HY x RO ($F_{1,16}$ = 0.446; p = 0.514), and PR x HY x RO ($F_{1,16}$ = 2.075; p = 0.169).

For inter-regional coupling parameters Xc-F, a significant main effect of HY was detected ($F_{1,16}$ = 5.082; p = 0.039). No significant main effect was detected for PR ($F_{1,16}$ = 0.867; p = 0.365), and RO ($F_{1,16}$ = 0.422; p = 0.525), nor was the interaction for PR x HY ($F_{1,16}$ = 0.929; p = 0.349), PR x RO ($F_{1,16}$ = 1.378; p = 0.258), HY x RO ($F_{1,16}$ = 3.983; p = 0.063), and PR x HY x RO ($F_{1,16}$ = 0.653; p = 0.431).

Student's *t*-tests for the Xr^+ parameters under Om were significant for the DPF recurrent connectivity (t_{16} = -2.919, p = 0.010) but not fore the Vix, PUT, or IPS recurrent connectivity (t_{16} = -1.401, p = 0.181; t_{16} = -1.996, p = 0.063; and t_{16} = -2.380, p = 0.030, respectively). None of the Xr^+ parameters under Dv was significant: Vix, t_{16} = -2.181, p = 0.045; PUT, t_{16} = -1.983, p = 0.065; IPS, t_{16} = -2.304, p = 0.035; DPF, t_{16} = -1.827, p = 0.087.

5.4. Discussion

This study aimed to characterise (1) the neural implementation of anticipatory set and (2) the information processing underlying the violation of the anticipated states in terms of their influences on effective connectivity. Here, the anticipatory set pertains to the maintenance of neuronal states that underpin flexible or stable working memory representations. In our experimental setting, this was induced by predictive cues that portend event cascades involving imminent updating or non-updating to information kept in working memory. We regarded the sustained neural activity associated with the anticipatory set as the implementation of prediction signals. Previously, we have demonstrated that such prediction signals mainly involve the parieto-occipital network and the striatum (see Chapter 3). However, another line of evidence (see Chapter 4) based on multivariate pattern analysis suggested otherwise – that, under surprising outcomes, neural responses reflecting pure prediction signals were associated with the fronto-parietal network. We argue that the ostensible contradiction between the two 'prediction signals' may be reconciled in light of the message passing scheme from theoretical neurobiology (Friston & Friston, 2005; Mumford, 1992; Rao & Ballard, 1999). That is, prediction

signals are encoded by top-down backward connections, whereas prediction errors signals are encoded by bottom-up forward connections. In other words, what we observed that appeared to be counter-intuitive may well speak to the directionality of information exchange within a neural network that integrates generative models with exogenous inputs. More importantly, the directionality of prediction/prediction error signal transmission may enable interpretations of neuronal plasticity that underlie representational meta-stability in working memory. Dynamic causal modelling (DCM) allows multiple plausible hypotheses to be motivated with regard to the statement above and to be tested using Bayesian model selection. Our result was consistent with the message passing scheme under the predictive coding framework in that the anticipatory set serves as top-down modulations on the backward connections, whilst the surprise (prediction error) reflects modulations on the forward and recurrent connections.

In cognitive neuroscience studies, it is considered that tasks requiring context-sensitive performance (e.g., those employing attentional or anticipatory set) are subject to top-down control. These contextual modulations are often trial-free, as opposed to evoked responses that are stimulus-bound, and speak to the effect of changes in membrane excitability and/or synaptic plasticity. In DCM analysis, a top-down modulation may be motivated in two equally plausible ways. One is via direct modulation on the backward connections. Alternatively, it may be expressed through modulations on the forward connections. Despite the plausibility, the two types of modulations can have quite distinct associations with neuronal innervations and therefore pertain to asymmetrical functional aspects. Given the current consensus that BOLD responses are more sensitive to presynaptic (driving/modulatory) activity, which is proportional neuronal spiking rates (Arthurs & Boniface, 2002;

Cardoso, Sirotin, Lima, Glushenkova, & Das, 2012; Friston, 2012), backward modulations may call for changes in spiking activity in the afferent neurons that have a strong ionotropic component or synaptic modifications that are metabotropically mediated in the dendritic tree of the lower area. Forward modulations, on the other hand, may reflect excitability of projection neurons targeting higher areas or shape the biophysical properties of the dendritic tree in the higher area. More importantly, cortico-cortical forward connections tend to terminate in the granular layer (L4), whereas backward connections originating from deep pyramidal cells tend to terminate outside L4 (i.e., L2/3 and L5/6; Felleman & Van Essen, 1991). Recent theoretical development in the canonical microcircuit and the free-energy principle (e.g., Bastos et al., 2012; Feldman & Friston, 2010; Friston, 2008) state that forward connections drive the L4 units reporting prediction errors, whilst backward connections signal the sensations about the world based on the underlying causes in encoded in a forward model (Friston & Friston, 2005). This lends an attentional role to the forward modulations because attention confers synaptic gain control over the prediction error units via nuancing the precision of the error signal (Feldman & Friston, 2010). A useful example to illustrate this is with the predominant forward modulations during load-dependent working memory performance (Dima, Joel, Jogia, & Frangou, 2013). Using the *n*-back paradigm, Dima et al. (2013) demonstrated that high *n*-back loads were associated with a tendency towards a lateralised forward parieto-prefrontal modulation. Their result can be interpreted under the generalised predictive coding framework (Friston, 2008; Friston et al., 2011), whereby an internal model is continuously inverted to update the causes – that generates imminent targets – from the inputs (Friston & Friston, 2005) with multiple instances of precision optimisation. We contrast the finding of Dima et al. with those

involved anticipatory processing. Rahnev et al. (2011) used predictive cues that portended the likely direction of moving stimuli and showed that employing prior expectations in perceptual decisions modulated both forward and backward connections between the prefrontal and sensory regions. Their result suggests that anticipatory processing involves a backward component, representing top-down prediction signals that modulate the motion-sensitive sensory areas.

Taken together, our result with regard to set modulation was in line with that of Rahnev et al. (2011), and was compatible with that of Dima et al. (2013) under the aforementioned theoretical framework. Although we did not model the concurrent forward/backward modulatory effect of the anticipatory set, we suspect that the forward modulation might be redundant in explaining our data, as compared with Rahnev et al. This is because upon the predictive cue our subjects lacked the recourse to utilise the prediction, which means there was no concrete representation on which predictive processing can be brought to bear. On the contrary, predictions employed in Rahnev et al. (2011) entailed concrete representations about how sensations will be caused. This allows the predictions to be reciprocated through intra-laminar or cortico-thalamic projections and back to higher areas to optimise the internal model. In other words, top-down modulations relating to perceptual set may have a role in exerting both driving and modulatory inputs, whereas the anticipatory set that serves to nuance cognitive meta-stability may have a predominantly modulatory role. We hope to elucidate this notion in future work.

Another crucial aspect of the forward set modulations is that the connections are differentially modulated in a context-sensitive manner. Specifically, our result suggests that anticipating stable or flexible working memory representations can be dissociated in terms of connectivity changes along the PUT-route or the IPS-route.

This was revealed by the significant PR x RO interaction. From Table 5.1 it is evident that the UPD-set had an enhancing influence on the DPF→PUT→Vix connection and a depressing influence on the DPF→IPS→Vix connection, whilst the reverse was the case for the MAI-set modulation. One can accordingly interpret that the DPF→PUT→Vix and DPF→IPS→Vix connections are two mutually antagonising functional circuits in the service of balancing cognitive flexibility and stability, respectively. Indeed, under the MAI-set, the brain must enable an efficient retrieval mechanism for the active representation (McElree, 2001) that may later come into the focus of attention. Empirical evidence has suggested that the parietal cortices mediate the selection of information outside of focus of attention (Bledowski et al., 2009) or the exclusion of irrelevant information (Vogel et al., 2005). Our result indicates that this mechanism may also be enabled in a preparatory sense. Along the same line, the specific selection mechanism may need to be downplayed under the UPD-set because the encoded information is potentially irrelevant before the realisation of an imminent update. This necessitates a higher degree of representational flexibility that implicates the gating mechanism via the basal ganglia (M. J. Frank et al., 2001; R. C. O'Reilly & Frank, 2006).

Our data were best explained by the model in which the violation of anticipatory set represents forward and local recurrent modulations. This finding is compatible with empirical findings from other domains (Garrido et al., 2008; e.g., Ouden et al., 2010), suggesting the predictive coding framework is a principled, unifying framework for understanding information processing in the brain. The predictive coding framework, i.e., hierarchical inference in the brain, states that the prediction error signals should take the form of forward (feedback) inputs and be minimised through recurrent interactions across levels of cortical hierarchy such that the most

probable cause of an input is derived (Friston, 2003; Friston & Friston, 2005). Changes in forward connectivity therefore conform to changes in the sensitivity of the unit reporting prediction error that is conveyed to higher areas and also speak to the relative influences between bottom-up prediction error and prediction error based on top-down prior expectations. Under this perspective, it may have been inferred in the brain that in the Om trials the cause of the sensations (i.e., upcoming target/probe stimulus) are solely due the variables encoded in the internal model of the world. On the other hand, under the Dv trials the internal model turned out to be improbable and had to incorporate external stimuli in order to update the internal model. This may partly explain our finding with regard to the main effect of HY and the significant DPF-DPF recurrent modulation: prediction error due to Om is predominantly endogenous and may rest on regulating higher level representations within association cortices or through lateral inhibition; whereas Dv had an additional level of prediction error that was stimulus-bound. However, it is unclear as to why the connectivity changes due to surprise did not exhibit a context-dependent dissociation, as one would have expected from the observation of set modulations. This calls for further studies to confirm this notion.

5.5. Conclusions

In summary, this study provided an integrative perspective of how anticipatory stability and flexibility, as well as their violations, modulate neuronal coupling within a working memory network. It provides evidence that working memory processes, as with perception, follow the framework of hierarchical inference, or generalised predictive coding. In other words, the brain not only represents the

environmental states that cause our sensations but also represents the likely fluctuations in those states, by implementing an 'anticipatory set'. The anticipatory set is synonymous with model predictions that pertain to backward connections from higher areas. We showed that the neuronal implementation of anticipatory set emerges as coupling between functionally specialised regions that differentially contribute to cognitive stability and flexibility. Consistent with the predictive coding framework, violating the anticipatory set reflects connectivity changes in forward and intra-regional coupling. However, the nature of neural computations underlying the surprise-related connectivity changes remains to be determined. Overall, our finding appeals to a novel yet complementary view of working memory function.

Chapter 6. General discussion and conclusions

This thesis started out questioning whether or not working memory follows the principle of hierarchical inference. That is, working memory function may be shaped by predictions and prediction errors. Although several previous studies have touched upon a relevant notion, they did not address the anticipatory flexibility of working memory representation, nor did they reveal the functional anatomy of prediction error responses. The original contributions of this thesis pertain to its methodology and empirical findings that addressed the aforementioned question. The novel experimental design used a cue-induced anticipatory set, not about stimulus identity, but about a likely event cascade that entails optimised cognitive flexibility and stability. Machine learning techniques and dynamic causal modelling were brought to bear to illustrate the multivariate nature and causal relationships in the working memory network. In Chapter 3, a key finding related the anticipatory set to the dopaminergic system. It showed how updating or maintaining of working memory contents may be mediated by anticipatory set through dopaminergic modulations. In Chapter 4, the violation of anticipatory set was examined with multivariate pattern analyses. It showed that prediction error responses comprise both endogenous (model) and exogenous (stimulus) components. This was reflected in the dissociable patterns of omission and deviation. Chapter 5 provided an integrated picture of prediction and prediction error in terms of their interactions with the working memory network. Using causal modelling and Bayesian model comparison, strong evidence indicated that prediction subserves backward modulations, whereas prediction error modulates forward and local recurrent connections. A crucial finding was also revealed that a connectivity-based mediation of representational flexibility

and stability may be attributable to the striatum and the inferior parietal cortex, respectively. Overall, this thesis provides the first evidence that working memory can be regarded as an instantiation of hierarchical inference. The following sections summarise several limitations of the current work and possible refinements to motivate future work.

6.1. Is the anticipatory set a non-specific modulation?

One obvious question relates to how specific the anticipatory set is in modulating working memory updating *per se*. Is the neuronal implementation of anticipatory processes targeting the time at which an update or maintenance takes place, or is it a fairly general nuance of neuronal dynamics, which influences the efficiency of information encoding? The latter speaks to the stability of the attractor network or the control of signal-to-noise ratio that entails shaping synaptic efficacy and lateral inhibition. The former, on the other hand, may require the exact timing of an update to be represented. It is one limitation of the current experimental design that the delay between successive cues was not jittered; therefore the aforementioned possibility cannot be ruled out. Nonetheless, it may be argued that the more concrete the idea is being anticipated, the more specific the anticipatory set is from an implementational aspect. In other words, concrete anticipation is about perceptual inference, about a single state that is expected. Anticipating whether or not to form a new binding of percepts may be a less concrete idea. Thus, the anticipatory set possibly defines a dynamic regime in which multiple tentative states may be coordinated to generate an integrated piece of information. This means the anticipatory set may be non-specific and may affect all information subsequently

represented. A metaphor to illustrate the difference an anticipatory set makes is, for example, placing stickers on a greased surface, as compared with on a paper surface. In one case, the stickers may be poorly secured but it is otherwise easier to re-order or replace some of them than in the other case. This tentative notion suggests possible differences in neural activity during encoding or retention between the two levels of anticipatory set, which calls for further empirical work.

6.2. To what extent is dopamine involved?

Following the question above, it is conceivable that dopaminergic modulation partly accounts for the functional role of anticipatory set. Indeed, elevated haemodynamic responses in the dopaminergic midbrain (SN/VTA) have been identified when subjects were implementing the anticipatory set for an imminent update (see Chapter 3). The midbrain activation was characterised by a sustained temporal profile, which was distinguished from the transient activation during an update. Although evidence has suggested midbrain (BOLD) activation reflects dopamine discharge (D'Ardenne et al., 2008) in its phasic mode (D'Ardenne et al., 2012), there is little evidence in association with tonic dopamine discharge. As reported in Chapter 3, the sustained midbrain activation was interpreted as reflecting tonic discharge. This is not entirely without physiological plausibility because (1) the fMRI signal reflects presynaptic activity (Friston, 2012) and (2) the tonic discharge of dopamine may be mediated by (prefrontal) glutamatergic afferents (Bilder et al., 2004; Grace, 1991). The problem remains with regard to which receptor subtype is implicated. Given the high binding affinity under relatively low extracellular concentration, the D2 receptor is a likely candidate. However, it is still unclear as to

whether the dopamine activity is restricted to interaction with auto-receptors or the concentration is high enough to interact with postsynaptic and extrasynaptic receptors. This may have a marked consequence in regulating the neuronal dynamic regime expressed through predominant D2 stimulations (Durstewitz & Seamans, 2008). One possible way to gain insight into this is to observe subjects' performance in the presence of distractors under an appropriate anticipatory set. Because the 'D2 state' entails cognitive flexibility and spontaneous representations (R. C. O'Reilly, 2006), if the anticipatory set induced a D2-state and had a non-specific modulatory effect (see above), then the subject's performance would be susceptible to irrelevant information in the environment. A relevant measure here is the intrusion rate (e.g., Artuso & Palladino, 2011). Alternatively, pharmacological interventions that induce D2 antagonism may be employed. Recent advances in combining drug administration and dynamic causal modelling as an assay of synaptic function (Moran, Symmonds, Stephan, Friston, & Dolan, 2011) would also shed light on this issue.

6.3. Towards a more comprehensive test of predictive coding

The current studies have drawn upon the predictive coding framework. The initial findings suggested that working memory follows hierarchical inferences in the brain, as revealed by backward modulations due to anticipatory set and forward modulations due to surprise.

The prefrontal cortex, the inferior parietal cortex, and the visual cortex are anatomically remote regions, therefore the inter-regional projections between these regions are more likely to follow the general pattern by Felleman and van Essen

(1991), thereby clearly defining their hierarchical relationship. The hierarchical position for the striatum is more elusive. Generally, the striatum receives cortical afferents mainly from layer V, the thalamus sends axons to cortical layer IV. This appears to make the striatum a hierarchically lower area to the other three mentioned earlier. However, the striatum also receives cortical afferents from supragranular layers (Steiner & Tseng, 2010). Additionally, the striatum receives converging inputs from nearly all cortical regions that are both hierarchically high and low. Overall, the cytoarchitecture of the striatum makes it more difficult to determine its hierarchical level.

Despite this limitation, the current evidence may be further strengthened by electro-/magnetophysiological measurements using EEG or MEG. The reason why M/EEG analysis may help characterising working memory processing as an instantiation of hierarchical inference is based on the findings with regard to lamina-specific neuronal synchronisation and spike-field coherence (Buffalo, Fries, Landman, Buschman, & Desimone, 2011; Roopun et al., 2008). Briefly, superficial layers of cortex are dominated by gamma frequencies, whereas deep layers show predominant alpha or beta frequencies. It is therefore useful to base inference of forward/backward connections on M/EEG data. Recent empirical evidence has implicated a functional dissociation between gamma and alpha oscillations in working memory performance: the alpha oscillation serves as an index to gate irrelevant information (Manza, Hau, & Leung, 2014) or as a preparatory set (Zanto, Chadick, & Gazzaley, 2014), whereas the gamma oscillation mediates successful execution or error detection in working memory performance (Yamamoto, Suh, Takeuchi, & Tonegawa, 2014). A potential problem might arise using an electromagnetophysiology approach to study neuronal activity in subcortical

structures. M/EEG is inherently of low sensitivity to subcortical generators, nevertheless, a model of deep brain activity may be applied to alleviate the limitation stated (e.g., Attal & Schwartz, 2013).

6.4. Synthetic model

Simulations with theoretical models of neural systems give insights into mechanistic principles; also, they predict the system's behaviour under aberrant parameters that can simulate neurological disorders (e.g., Friston et al., 2012; Humphries, Stewart, & Gurney, 2006). These models can be realised at different scales, from single neurons (e.g., leaky integrate-an-fire model; Brunel, 2000; Humphries et al., 2006) to neuronal ensembles (Friston et al., 2012; e.g., Pinotsis & Friston, 2011). Our understanding of working memory processing has benefitted from attractor models of neuronal firing pertaining to delayed-response (Amit et al., 1997) or neuromodulations (Durstewitz et al., 2000; Durstewitz & Seamans, 2008). Contrary to many models of neural mechanisms that hold an implicit assumption of steady-state or periodic network dynamics, models assuming transient states arguably provide better accounts for network behaviour (Rabinovich et al., 2008). A nice example of transient population dynamics is with the winner-less competition – or the predator-prey relationship – in which no stable equilibrium is reached. Winner-less competition can be implemented using the Lotka-Volterra equation (Hoppensteadt, 2006). If one wishes to model working memory processes at a population level, then the Lotka-Volterra model would be a suitable approach. This is because working memory representations are transiently stable, i.e., they can achieve stability and flexibility concurrently and selectively. Recent modelling work

based on winner-less competition (or stable heteroclinic sequence) revealed that an effective limit in capacity (cf. magic number 7; Jensen & Lisman, 1996) naturally emerges in working memory (Bick & Rabinovich, 2009). Bick and Rabinovich's (2009) work has a profound implication for dopaminergic modulation, as dopamine is implicated in representational stability (Durstewitz & Seamans, 2008) and capacity limits (Cools et al., 2008). It is foreseeable that such models can be refined to include descriptions of dopaminergic or anticipatory modulations. Also, multiple layers of stable heteroclinic sequences/cycles (Bick et al., 2010) may be devised to allow characterisation of slow dynamics of the set-maintenance network (cf. task-set control of the cingulo-opercular network; Chapter 4). Updating in working memory, on the other hand, may be realised with the heteroclinic binding model (Rabinovich, Afraimovich, & Varona, 2010). Crucially, one would hope to take a probabilistic perspective on neuronal states (see variable-precision models; Ma et al., 2014) and to bring the above framework into the formalisation of free-energy minimisation (Friston et al., 2006; Friston, 2008).

6.5. The issue with working memory capacity

Little has been addressed in this thesis with regard to the contribution of the measure of working memory capacity to individual working memory performance. Working memory capacity varies across individuals and may speak to the intrinsic heterogeneity of neurochemistry in the brain (Cools et al., 2008; Cools & Robbins, 2004). One preliminary finding not reported in this thesis is the correlation between individual working memory capacity and inverse efficiency scores (IES; Bruyer & Brysbaert, 2013). The IES was first proposed by Townsend and Ashby (Townsend &

Ashby, 1978; 1983) as an attempt to combine reaction time and error rate into a single measure. IES was taken as a measure of average 'energy consumption' over time. In other words, it treats an individual as less 'energetic' in the course of performing a (mentally) resource-intensive task, thus reflecting a higher IES. The finding stated revealed that individual working memory capacity predicts IES in all four conditions (Spearman's correlation; $p < 0.001$): the higher the capacity, the lower the IES. This implies that subjects with higher capacity require less effort or are more efficient at processing relevant information. A consistent interpretation can be found in Vogel et al. (Vogel et al., 2005). More recently, probabilistic characterisation of precision-based memory representation lends a new perspective on the role of memory capacity and its underlying neural substrates (see Ma & Jazayeri, 2014 for review).

6.6. Conclusions

In summary, this thesis has provided a more comprehensive understanding of working memory processing. Specifically, it states that anticipatory processing is also a determinant of working memory performance and information processing in the brain. A likely interpretation follows that the principle of hierarchical inference is applied to working memory as well. It is possible that sensory processing and higher cognition may employ a unified computational principle. This is a notion that deserves intensive explorations in due course.

Bibliography

Albin, R. L., Young, A. B., & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, *12*(10), 366–375.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*(1), 357–381. doi:10.1146/annurev.ne.09.030186.002041

Almeida, J. R. C. de, Kronhaus, D. M., Sibille, E. L., Langenecker, S. A., Versace, A., Labarbara, E. J., & Phillips, M. L. (2011). Abnormal left-sided orbitomedial prefrontal cortical-amygdala connectivity during happy and fear face processing: a potential neural mechanism of female MDD. *Frontiers in Psychiatry*, *2*, 69. doi:10.3389/fpsyt.2011.00069

Almeida, J. R. C. de, Versace, A., Mechelli, A., Hassel, S., Quevedo, K., Kupfer, D. J., & Phillips, M. L. (2009). Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. *Biological Psychiatry*, *66*(5), 451–459. doi:10.1016/j.biopsych.2009.03.024

Amit, D. J., Fusi, S., & Yakovlev, V. (1997). Paradigmatic working memory (attractor) cell in IT cortex. *Neural Computation*, (5).

Anderson, D. R. (2008). *Model Based Inference in the Life Sciences: A Primer on Evidence*. New York, NY: Springer New York. doi:10.1007/978-0-387-74075-1

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage*, *13*(5), 903–919. doi:10.1006/nimg.2001.0746

Aron, A. R. (2007). The neural basis of inhibition in cognitive control. *The Neuroscientist*, *13*(3), 214–228. doi:10.1177/1073858407299288

Arthurs, O. J., & Boniface, S. (2002). How well do we understand the neural origins of the fMRI BOLD signal? *Trends in Neurosciences*, *25*(1), 27–31.

Artuso, C., & Palladino, P. (2011). Content-context binding in verbal working memory updating: on-line and off-line effects. *Acta Psychologica*, *136*(3), 363–369. doi:10.1016/j.actpsy.2011.01.001

Artuso, C., & Palladino, P. (2014). Binding and content updating in working memory tasks. *British Journal of Psychology*, *105*(2), 226–242. doi:10.1111/bjop.12024

Ashburner, J. (2012). SPM: a history. *NeuroImage*, *62*(2), 791–800. doi:10.1016/j.neuroimage.2011.10.025

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence, *The psychology of learning and motivation: advances in research and theory* (Vol. 2, pp. 89–195). New York: Academic Press.

Attal, Y., & Schwartz, D. (2013). Assessment of subcortical source localization using deep brain activity imaging model with minimum norm operators: a MEG study. *PloS One*, *8*(3), e59856. doi:10.1371/journal.pone.0059856

Baddeley, A. (1992). Working memory. *Science (New York, N.Y.)*, *255*(5044), 556–559. doi:10.1126/science.1736359

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29. doi:10.1146/annurev-psych-120710-100422

Baddeley, A. D. (1986). *Working memory*. New York : Oxford University Press.

Baddeley, A. D. (2007). *Working memory, thought, and action*. New York: Oxford University Press.

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(2), 176–189. doi:10.1016/S0022-5371(70)80048-2

Baier, B., Karnath, H.-O., Dieterich, M., Birklein, F., Heinze, C., & Müller, N. G. (2010). Keeping memory clear and stable--the contribution of human basal ganglia and prefrontal cortex to working memory. *Journal of Neuroscience*, *30*(29), 9788–9792. doi:10.1523/JNEUROSCI.1513-10.2010

Bamford, N. S., Zhang, H., Schmitz, Y., Wu, N.-P., Cepeda, C., Levine, M. S., et al. (2004). Heterosynaptic dopamine neurotransmission selects sets of corticostriatal terminals. *Neuron*, *42*(4), 653–663.

Bar-Gad, I., & Bergman, H. (2001). Stepping out of the box: information processing in the neural networks of the basal ganglia. *Current Opinion in Neurobiology*, *11*(6), 689–695.

Barch, D. M., Braver, T. S., Nystrom, L. E., Forman, S. D., Noll, D. C., & Cohen, J. D. (1997). Dissociating working memory from task difficulty in human prefrontal cortex. *Neuropsychologia*, *35*(10), 1373–1380.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. doi:10.1016/j.neuron.2012.10.038

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science (New York, N.Y.)*, *321*(5890), 851–854. doi:10.1126/science.1158023

Bányai, M., Diwadkar, V. A., & Erdi, P. (2011). Model-based dynamical analysis of functional disconnection in schizophrenia. *NeuroImage*, *58*(3), 870–877. doi:10.1016/j.neuroimage.2011.06.046

Bäckman, L., Nyberg, L., Soveri, A., Johansson, J., Andersson, M., Dahlin, E., et al. (2011). Effects of working-memory training on striatal dopamine release. *Science (New York, N.Y.)*, *333*(6043), 718. doi:10.1126/science.1204978

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1221. doi:10.1038/nn1954

Bellander, M., Bäckman, L., Liu, T., Schjeide, B.-M. M., Bertram, L., Schmiedek, F., et al. (2014). Lower baseline performance but greater plasticity of working memory for carriers of the val allele of the comt val158met polymorphism. *Neuropsychology*. doi:10.1037/neu0000088

Benoit-Marand, M., Borrelli, E., & Gonon, F. (2001). Inhibition of dopamine release via presynaptic D2 receptors: time course and functional characteristics in vivo. *Journal of Neuroscience*, *21*(23), 9134–9141.

Bick, C., & Rabinovich, M. I. (2009). Dynamical origin of the effective storage capacity in the brain's working memory. *Physical Review Letters*. doi:10.1103/PhysRevLett.103.218101

Bick, C., Rabinovich, M. I., & Rabinovich, M. I. (2010). On the occurrence of stable heteroclinic channels in Lotka–Volterra models. *Dynamical Systems*, *25*(1), 97–110. doi:10.1080/14689360903322227

Bilder, R. M., Volavka, J., Lachman, H. M., & Grace, A. A. (2004). The catechol-O-methyltransferase polymorphism: relations to the tonic-phasic dopamine hypothesis and neuropsychiatric phenotypes. *Neuropsychopharmacology*, *29*(11), 1943–1961. doi:10.1038/sj.npp.1300542

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York : Springer.

Bledowski, C., Rahm, B., & Rowe, J. B. (2009). What “works” in working memory? Separate systems for selection and updating of critical information. *Journal of Neuroscience*, *29*(43), 13735–13741. doi:10.1523/JNEUROSCI.2547-09.2009

Bollinger, J., Rubens, M. T., Zanto, T. P., & Gazzaley, A. (2010). Expectation-driven changes in cortical functional connectivity influence working memory and long-term memory performance. *Journal of Neuroscience*, *30*(43), 14399–14410. doi:10.1523/JNEUROSCI.1547-10.2010

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, *16*(13), 4207–4221.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell & J. Driver, *Attention and performance XVIII Control of cognitive processes* (pp. 713–737).
Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. doi:10.1016/j.neuron.2009.06.009
Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21. doi:10.1080/17470215808416249
Brunel, N. (2000). Dynamics of networks of randomly connected excitatory and inhibitory spiking neurons. *Journal of Physiology-Paris*, *94*(5-6), 445–463.
Bruyer, R., & Brysbaert, M. (2013). Combining Speed and Accuracy in Cognitive Psychology: Is the Inverse Efficiency Score (IES) a Better Dependent Variable than the Mean Reaction Time (RT) and the Percentage Of Errors (PE)? *Psychologica Belgica*, *51*(1). doi:10.5334/pb-51-1-5
Bubic, A., Cramon, von, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*, 1–15. doi:10.3389/fnhum.2010.00025
Bubic, A., Cramon, von, D. Y., Jacobsen, T., Schröger, E., & Schubotz, R. I. (2009). Violation of expectation: neural correlates reflect bases of prediction. *Journal of Cognitive Neuroscience*, *21*(1), 155–168. doi:10.1162/jocn.2009.21013
Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J., & Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11262–11267. doi:10.1073/pnas.1011284108
Bunting, M., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *The Quarterly Journal of Experimental Psychology*, *59*(10), 1691–1700. doi:10.1080/17470210600848402
Buxton, R. B., Uludağ, K., Dubowitz, D. J., & Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*, *23 Suppl 1*, S220–33. doi:10.1016/j.neuroimage.2004.07.013
Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, *39*(6), 855–864.
Camperi, M., & Wang, X. J. (1998). A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. *Journal of Computational Neuroscience*, *5*(4), 383–405.
Camps, M., Cortés, R., Gueye, B., Probst, A., & Palacios, J. M. (1989). Dopamine receptors in human brain: Autoradiographic distribution of D2 sites. *Neuroscience*, *28*(2), 275–290. doi:10.1016/0306-4522(89)90179-6
Cardoso, M. M. B., Sirotin, Y. B., Lima, B., Glushenkova, E., & Das, A. (2012). The neuroimaging signal is a linear sum of neurally distinct stimulus- and task-related components. *Nature Neuroscience*, *15*(9), 1298–1306. doi:10.1038/nn.3170
Carpenter, B. E., & Doran, R. W. (Eds.). (1986). *A.M. Turing's ACE report of 1946 and other papers*. MIT Press.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. doi:10.1613/jair.953
Chen, C. C., Henson, R. N., Stephan, K. E., Kilner, J. M., & Friston, K. J. (2009). Forward and backward connections in the brain: a DCM study of functional asymmetries. *NeuroImage*, *45*(2), 453–462. doi:10.1016/j.neuroimage.2008.12.041
Chen, T., & Li, D. (2007). The roles of working memory updating and processing speed in mediating age-related differences in fluid intelligence. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *14*(6), 631–646. doi:10.1080/13825580600987660
Chu, C. Y. C. (2009, May 1). *Pattern recognition and machine learning for magnetic resonance images with kernel methods*.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: recent developments and current challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *351*(1346), 1515–1527. doi:10.1098/rstb.1996.0138

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), 604–608. doi:10.1038/386604a0

Collette, F., Van der Linden, M., Laureys, S., Delfiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2005). Exploring the unity and diversity of the neural substrates of executive functioning. *Human Brain Mapping*, *25*(4), 409–423. doi:10.1002/hbm.20118

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923. doi:10.1093/cercor/10.9.910

Cools, R., & D'Esposito, M. (2011). Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biological Psychiatry*, *69*(12), e113–25. doi:10.1016/j.biopsych.2011.03.028

Cools, R., & Robbins, T. W. (2004). Chemistry of the adaptive mind. *Philosophical Transactions. Series a, Mathematical, Physical, and Engineering Sciences*, *362*(1825), 2871–2888. doi:10.1098/rsta.2004.1468

Cools, R., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2008). Working memory capacity predicts dopamine synthesis capacity in the human striatum. *Journal of Neuroscience*, *28*(5), 1208–1212. doi:10.1523/JNEUROSCI.4475-07.2008

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215. doi:10.1038/nrn755

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature*, *386*(6625), 608–611. doi:10.1038/386608a0

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191.

Cowan, N. (2005). Capacity limits for unstructured materials. In *Working memory capacity*. Hove: Psychology Press.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In *Essence of Memory* (Vol. 169, pp. 323–338). Elsevier. doi:10.1016/S0079-6123(07)00020-9

Cowan, R. L., & Wilson, C. J. (1994). Spontaneous firing patterns and axonal projections of single corticostriatal neurons in the rat medial agranular cortex. *Journal of Neurophysiology*, *71*(1), 17–32.

Crofts, H. S., Dalley, J. W., Collins, P., Van Denderen, J. C., Everitt, B. J., Robbins, T. W., & Roberts, A. C. (2001). Differential effects of 6-OHDA lesions of the frontal cortex and caudate nucleus on the ability to acquire an attentional set. *Cerebral Cortex*, *11*(11), 1015–1026.

D'Ardenne, K., Eshel, N., Luka, J., Lenartowicz, A., Nystrom, L. E., & Cohen, J. D. (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(49), 19900–19909. doi:10.1073/pnas.1116727109

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science (New York, N.Y.)*, *319*(5867), 1264–1267. doi:10.1126/science.1150605

D'Esposito, M., Postle, B. R., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Experimental Brain Research*, *133*(1), 3–11. doi:10.1007/s002210000395

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Deserno, L., Sterzer, P., Wüstenberg, T., Heinz, A., & Schlagenhauf, F. (2012). Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. *Journal of Neuroscience*, *32*(1), 12–20. doi:10.1523/JNEUROSCI.3405-11.2012

Desrochers, T. M., & Badre, D. (2012). Finding parallels in fronto-striatal organization. *Trends in Cognitive Sciences*, *16*(8), 407–408. doi:10.1016/j.tics.2012.06.009

Desseilles, M., Schwartz, S., Dang-Vu, T. T., Sterpenich, V., Ansseau, M., Maquet, P., & Phillips, C. (2011). Depression alters "top-down" visual attention: a dynamic causal modeling comparison between depressed and healthy subjects. *NeuroImage*, *54*(2), 1662–1668. doi:10.1016/j.neuroimage.2010.08.061

Dima, D., Joel, D., Jogia, J., & Frangou, S. (2013). Dynamic causal modeling of load-dependent modulation of effective connectivity within the verbal working memory network. *Human Brain Mapping*, *35*(7), 3025–3035. doi:10.1002/hbm.22382

Diwadkar, V. A., Wadehra, S., Pruitt, P., Keshavan, M. S., Rajan, U., Zajac-Benitez, C., & Eickhoff, S. B. (2012). Disordered corticolimbic interactions during affective processing in children and adolescents at risk for schizophrenia revealed by functional magnetic resonance imaging and dynamic causal modeling. *Archives of General Psychiatry*, *69*(3), 231–242. doi:10.1001/archgenpsychiatry.2011.1349

Donaldson, D. I. (2004). Parsing brain activity with fMRI and mixed designs: what kind of a state is neuroimaging in? *Trends in Neurosciences*, *27*(8), 442–444. doi:10.1016/j.tins.2004.06.001

Dosenbach, N. U. F., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends in Cognitive Sciences*, *12*(3), 99–105. doi:10.1016/j.tics.2008.01.001

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., et al. (2006). A core system for the implementation of task sets. *Neuron*, *50*(5), 799–812. doi:10.1016/j.neuron.2006.04.031

Doya, K. (2007). *Bayesian brain : probabilistic approaches to neural coding*. Cambridge, Mass. : MIT Press.

Draganski, B., Kherif, F., Klöppel, S., Cook, P. A., Alexander, D. C., Parker, G. J. M., et al. (2008). Evidence for segregated and integrative connectivity patterns in the human Basal Ganglia. *Journal of Neuroscience*, *28*(28), 7143–7152. doi:10.1523/JNEUROSCI.1486-08.2008

Dreher, J.-C., & Burnod, Y. (2002). An integrative theory of the phasic and tonic modes of dopamine modulation in the prefrontal cortex. *Neural Networks : the Official Journal of the International Neural Network Society*, *15*(4-6), 583–602.

Dreyer, J. K., Herrik, K. F., Berg, R. W., & Hounsgaard, J. D. (2010). Influence of phasic and tonic dopamine release on receptor activation. *Journal of Neuroscience*, *30*(42), 14273–14283. doi:10.1523/JNEUROSCI.1894-10.2010

Durstewitz, D., & Seamans, J. K. (2008). The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. *Biological Psychiatry*, *64*(9), 739–749. doi:10.1016/j.biopsych.2008.05.015

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

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3 Suppl*, 1184–1191. doi:10.1038/81460

Düzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B. H., & Tobler, P. N. (2009). Functional imaging of the human dopaminergic midbrain. *Trends in Neurosciences*, *32*(6), 321–328. doi:10.1016/j.tins.2009.02.005

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., & Chee, A. E. H. (2010). The components of working memory updating: an experimental decomposition and individual differences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(1), 170–189. doi:10.1037/a0017891

Eshel, N., Tian, J., & Uchida, N. (2013). Opening the black box: dopamine, predictions, and

learning. *Trends in Cognitive Sciences*, *17*(9), 430–431. doi:10.1016/j.tics.2013.06.010
Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215. doi:10.3389/fnhum.2010.00215
Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.
Fincham, J. M., Carter, C. S., van Veen, V., Stenger, V. A., & Anderson, J. R. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proceedings of the National Academy of Sciences*, *99*(5), 3346–3351. doi:10.1073/pnas.052703399
Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, *299*(5614), 1898–1902. doi:10.1126/science.1077349
Fitzgerald, T. H. B., Friston, K. J., & Dolan, R. J. (2012). Action-specific value signals in reward-related regions of the human brain. *Journal of Neuroscience*, *32*(46), 16417–23a. doi:10.1523/JNEUROSCI.3254-12.2012
Floresco, S. B., & Phillips, A. G. (2001). Delay-dependent modulation of memory retrieval by infusion of a dopamine D1 agonist into the rat medial prefrontal cortex. *Behavioral Neuroscience*, *115*(4), 934–939.
Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral Cortex*, *22*(3), 509–526. doi:10.1093/cercor/bhr114
Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience*, *120*(3), 497–517. doi:10.1037/0735-7044.120.3.497
Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective & Behavioral Neuroscience*, *1*(2), 137–160.
Friston, K. (2003). Learning and inference in the brain. *Neural Networks : the Official Journal of the International Neural Network Society*, *16*(9), 1325–1352. doi:10.1016/j.neunet.2003.06.005
Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11), e1000211. doi:10.1371/journal.pcbi.1000211
Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi:10.1016/j.tics.2009.04.005
Friston, K. J. (2012). What does functional MRI measure? Two complementary perspectives. *Trends in Cognitive Sciences*, *16*(10), 491–492. doi:10.1016/j.tics.2012.08.005
Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458. doi:10.1007/s11229-007-9237-y
Friston, K. J., Friston, K., Kiebel, S., & Kiebel, S. J. (2009). Cortical circuits for perceptual inference. *Neural Networks*, *22*(8), 1093–1104. doi:10.1016/j.neunet.2009.07.023
Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J. T., & Penny, W. D. (2004). *Human brain function*. Academic Press.
Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.
Friston, K. J., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1-2), 137–160. doi:10.1007/s00422-011-0424-z
Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, *12*(4), 466–477. doi:10.1006/nimg.2000.0630
Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, *8*(1), e1002327. doi:10.1371/journal.pcbi.1002327
Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. doi:10.1098/rstb.2005.1622

Friston, K., Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1-3), 70–87. doi:10.1016/j.jphysparis.2006.10.001

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., et al. (2008). Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, *39*(3), 1104–1120. doi:10.1016/j.neuroimage.2007.09.048

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234. doi:10.1016/j.neuroimage.2006.08.035

Frith, C. D. (2007). *Making up the mind : how the brain creates our mental world*. Malden, MA : Blackwell Pub.

Fuster, J. (2008). *The Prefrontal Cortex*. Elsevier.

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*(1), 61–78.

Fuster, J. M., & Alexander, G. E. (1973). Firing changes in cells of the nucleus medialis dorsalis associated with delayed response behavior. *Brain Research*, *61*, 79–91.

Garavan, H. (1998). Serial attention within working memory. *Memory & Cognition*, *26*(2), 263–276.

Garrido, M. I., Dolan, R. J., & Sahani, M. (2011). Surprise leads to noisier perceptual decisions., *2*(2), 112–120. doi:10.1068/i0411

Garrido, M. I., Friston, K. J., Kiebel, S. J., Stephan, K. E., Baldeweg, T., & Kilner, J. M. (2008). The functional anatomy of the MMN: a DCM study of the roving paradigm. *NeuroImage*, *42*(2), 936–944. doi:10.1016/j.neuroimage.2008.05.018

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *120*(3), 453–463. doi:10.1016/j.clinph.2008.11.029

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. doi:10.1016/j.tics.2011.11.014

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485. doi:10.1016/0896-6273(95)90304-6

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13473–13480.

Goldman-Rakic, P. S., Lidow, M. S., Smiley, J. F., & Williams, M. S. (1992). The anatomy of dopamine in monkey and human prefrontal cortex. In *Advances in neuroscience and Schizophrenia* (pp. 163–177). Vienna: Springer Vienna. doi:10.1007/978-3-7091-9211-5_8

Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. *Towards an Executive Without a Homunculus: Computational Models of the Prefrontal Cortex/Basal Ganglia System.*, *18*, 330–341.

Gordon, E. M., Stollstorff, M., & Vaidya, C. J. (2012). Using spatial multiple regression to identify intrinsic connectivity networks involved in working memory performance. *Human Brain Mapping*, *33*(7), 1536–1552. doi:10.1002/hbm.21306

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience*, *31*(23), 8502–8511. doi:10.1523/JNEUROSCI.0208-11.2011

Goto, Y., Otani, S., & Grace, A. A. (2007). The Yin and Yang of dopamine release: a new perspective. *Neuropharmacology*, *53*(5), 583–587. doi:10.1016/j.neuropharm.2007.07.007

Grace, A. A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience*, *41*(1), 1–24. doi:10.1016/0306-4522(91)90196-U

Granon, S., Passetti, F., Thomas, K. L., Dalley, J. W., Everitt, B. J., & Robbins, T. W.

(2000). Enhanced and impaired attentional performance after infusion of D1 dopaminergic receptor agents into rat prefrontal cortex. *Journal of Neuroscience*, *20*(3), 1208–1215.
Grubb, R. L., Raichle, M. E., Eichling, J. O., & Ter-Pogossian, M. M. (1974). The effects of changes in PaCO2 on cerebral blood volume, blood flow, and vascular mean transit time. *Stroke; a Journal of Cerebral Circulation*, *5*(5), 630–639.
Gruber, O., Kleinschmidt, A., Binkofski, F., Steinmetz, H., & Cramon, von, D. Y. (2000). Cerebral correlates of working memory for temporal information. *Neuroreport*, *11*(8), 1689–1693.
Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *NeuroImage*, *62*(2), 852–855. doi:10.1016/j.neuroimage.2012.03.016
Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: making working memory work. *Neuroscience*, *139*(1), 105–118. doi:10.1016/j.neuroscience.2005.04.067
Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1601–1613. doi:10.1098/rstb.2007.2055
Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory*. New York: Wiley.
Helms, G., Draganski, B., Frackowiak, R., Ashburner, J., & Weiskopf, N. (2009). Improved segmentation of deep brain grey matter structures using magnetization transfer (MT) parameter maps. *NeuroImage*, *47*(1), 194–198. doi:10.1016/j.neuroimage.2009.03.053
Henson, R. N., Büchel, C., Josephs, O., & Friston, K. J. (1999). The slice-timing problem in event-related fMRI. *NeuroImage*, *25*, 125.
Hikosaka, O., & Isoda, M. (2010). Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms. *Trends in Cognitive Sciences*, *14*(4), 154–161. doi:10.1016/j.tics.2010.01.006
Hitch, G. J., & Baddeley, A. D. (1976). Verbal reasoning and working memory. *Quarterly Journal of Experimental Psychology*, *28*(4), 603–621. doi:10.1080/14640747608400587
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, *36*(3), 1171–1220. doi:10.1214/009053607000000677
Hong, S. (2013). Dopamine system: manager of neural pathways. *Frontiers in Human Neuroscience*, *7*, 854. doi:10.3389/fnhum.2013.00854
Hoppensteadt, F. (2006). Predator-prey model. *Scholarpedia*.
Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.
Humphries, M. D., Stewart, R. D., & Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *Journal of Neuroscience*, *26*(50), 12921–12942. doi:10.1523/JNEUROSCI.3486-06.2006
Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. *NeuroImage*, *16*(1), 217–240. doi:10.1006/nimg.2001.1054
Hutton, C., Hutton, C., Josephs, O., Josephs, O., Stadler, J., Featherstone, E., et al. (2011). The impact of physiological noise correction on fMRI at 7T. *NeuroImage*, *57*(1), 101–112. doi:10.1016/j.neuroimage.2011.04.018
Iversen, L. L. (1973). Catecholamine uptake processes. *British Medical Bulletin*, *29*(2), 130–135.
Iversen, L. L. (2010). *Dopamine Handbook*. Oxford University Press.
James, W. (1890). *The principles of psychology*. New York: Henry Holt.
Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449.
Jensen, O., & Lisman, J. E. (1996). Novel lists of 7 +/- 2 known items can be reliably stored

in an oscillatory short-term memory network: interaction with long-term memory. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *3*(2-3), 257–263.
Jezzard, P., & Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic Resonance in Medicine*, *34*(1), 65–73. doi:10.1002/mrm.1910340111
Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davies, *Varieties of attention* (pp. 29–61).
Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.
Karreman, M., & Moghaddam, B. (1996). The prefrontal cortex regulates the basal release of dopamine in the limbic striatum: an effect mediated by ventral tegmental area. *Journal of Neurochemistry*, *66*(2), 589–598.
Kasess, C. H., Stephan, K. E., Weissenbacher, A., Pezawas, L., Moser, E., & Windischberger, C. (2010). Multi-subject analyses with dynamic causal modeling. *NeuroImage*, *49*(4), 3065–3074. doi:10.1016/j.neuroimage.2009.11.037
Kessler, Y., & Meiran, N. (2006). All updateable objects in working memory are updated whenever any of them are modified: evidence from the memory updating paradigm. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(3), 570–585. doi:10.1037/0278-7393.32.3.570
Kessler, Y., & Meiran, N. (2008). Two dissociable updating processes in working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(6), 1339–1348. doi:10.1037/a0013078
Kimberg, D. Y., D'Esposito, M., & Farah, M. J. (1997). Effects of bromocriptine on human subjects depend on working memory capacity. *Neuroreport*, *8*(16), 3581–3585.
Knudsen, E. I. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, *30*, 57–78. doi:10.1146/annurev.neuro.30.051606.094256
Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*(6), 229–235. doi:10.1016/j.tics.2007.04.005
Kreitzer, A. C., & Malenka, R. C. (2008). Striatal plasticity and basal ganglia circuit function. *Neuron*, *60*(4), 543–554. doi:10.1016/j.neuron.2008.11.005
Kuhl, B. A., Bainbridge, W. A., & Chun, M. M. (2012). Neural reactivation reveals mechanisms for updating memory. *Journal of Neuroscience*, *32*(10), 3453–3461. doi:10.1523/JNEUROSCI.5846-11.2012
Lachman, H. M., Papolos, D. F., Saito, T., Yu, Y. M., Szumlanski, C. L., & Weinshilboum, R. M. (1996). Human catechol-O-methyltransferase pharmacogenetics: description of a functional polymorphism and its potential application to neuropsychiatric disorders. *Pharmacogenetics*, *6*(3), 243–250.
Lalchandani, R. R., van der Goes, M.-S., Partridge, J. G., & Vicini, S. (2013). Dopamine D2 receptors regulate collateral inhibition between striatal medium spiny neurons. *Journal of Neuroscience*, *33*(35), 14075–14086. doi:10.1523/JNEUROSCI.0692-13.2013
Lavin, A., & Grace, A. A. (2001). Stimulation of D1-type dopamine receptors enhances excitability in prefrontal cortical pyramidal neurons in a state-dependent manner. *Neuroscience*, *104*(2), 335–346.
Lenartowicz, A., Escobedo-Quiroz, R., & Cohen, J. D. (2010). Updating of context in working memory: an event-related potential study. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 298–315. doi:10.3758/CABN.10.2.298
Levey, A. I., Hersch, S. M., Rye, D. B., Sunahara, R. K., Niznik, H. B., Kitt, C. A., et al. (1993). Localization of D1 and D2 dopamine receptors in brain with subtype-specific antibodies. *Proceedings of the National Academy of Sciences*, *90*(19), 8861–8865.
Lewis, D. A., Melchitzky, D. S., Sesack, S. R., Whitehead, R. E., Auh, S., & Sampson, A. (2001). Dopamine transporter immunoreactivity in monkey cerebral cortex: regional, laminar, and ultrastructural localization. *The Journal of Comparative Neurology*, *432*(1), 119–136.
Lewis-Peacock, J. A., & Postle, B. R. (2008). Temporary activation of long-term memory

supports working memory. *Journal of Neuroscience*, *28*(35), 8765–8771. doi:10.1523/JNEUROSCI.1953-08.2008
Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. doi:10.1162/jocn_a_00140
Li, B., Daunizeau, J., Stephan, K. E., Penny, W., Hu, D., & Friston, K. (2011a). Generalised filtering and stochastic DCM for fMRI. *NeuroImage*, *58*(2), 442–457. doi:10.1016/j.neuroimage.2011.01.085
Li, X., Large, C. H., Ricci, R., Taylor, J. J., Nahas, Z., Bohning, D. E., et al. (2011b). Using interleaved transcranial magnetic stimulation/functional magnetic resonance imaging (fMRI) and dynamic causal modeling to understand the discrete circuit specific changes of medications: lamotrigine and valproic acid changes in motor or prefrontal effective connectivity. *Psychiatry Research*, *194*(2), 141–148. doi:10.1016/j.pscychresns.2011.04.012
Lidow, M. S., Goldman-Rakic, P. S., Gallager, D. W., & Rakic, P. (1991). Distribution of dopaminergic receptors in the primate cerebral cortex: quantitative autoradiographic analysis using [3H]raclopride, [3H]spiperone and [3H]SCH23390. *Neuroscience*, *40*(3), 657–671.
Linden, D. E. J. (2007). The working memory networks of the human brain. *The Neuroscientist*, *13*(3), 257–267. doi:10.1177/1073858406298480
Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., et al. (2011). EEG and MEG data analysis in SPM8. *Computational Intelligence and Neuroscience*, *2011*(4), 852961–32. doi:10.1155/2011/852961
Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220. doi:10.1146/annurev-neuro-071013-014017
Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. doi:10.1038/nn.3655
MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK ; New York : Cambridge University Press.
Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel pattern analysis for FMRI data: a review. *Computational and Mathematical Methods in Medicine*, *2012*, 961257. doi:10.1155/2012/961257
Maldjian, J. A., Laurienti, P. J., & Burdette, J. H. (2004). Precentral gyrus discrepancy in electronic versions of the Talairach atlas. *NeuroImage*, *21*(1), 450–455.
Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*(3), 1233–1239.
Mandeville, J. B., Marota, J. J., Ayata, C., Moskowitz, M. A., Weisskoff, R. M., & Rosen, B. R. (1999). MRI measurement of the temporal evolution of relative CMRO(2) during rat forepaw stimulation. *Magnetic Resonance in Medicine*, *42*(5), 944–951.
Manza, P., Hau, C. L. V., & Leung, H.-C. (2014). Alpha power gates relevant information during working memory updating. *Journal of Neuroscience*, *34*(17), 5998–6002. doi:10.1523/JNEUROSCI.4641-13.2014
Markett, S., Reuter, M., Montag, C., Voigt, G., Lachmann, B., Rudorf, S., et al. (2013). Assessing the function of the fronto-parietal attention network: Insights from resting-state fMRI and the attentional network test. *Human Brain Mapping*, *35*(4), 1700–1709. doi:10.1002/hbm.22285
Marklund, P., Larsson, A., Elgh, E., Linder, J., Riklund, K. A., Forsgren, L., & Nyberg, L. (2009). Temporal dynamics of basal ganglia under-recruitment in Parkinson's disease: transient caudate abnormalities during updating of working memory. *Brain*, *132*(Pt 2), 336–346. doi:10.1093/brain/awn309
Markov, N. T., & Kennedy, H. (2013). The importance of being hierarchical. *Current Opinion in Neurobiology*, *23*(2), 187–194. doi:10.1016/j.conb.2012.12.008
Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39.

doi:10.3389/fnhum.2011.00039
Mattay, V. S., Goldberg, T. E., Fera, F., Hariri, A. R., Tessitore, A., Egan, M. F., et al. (2003). Catechol O-methyltransferase val158-met genotype and individual variation in the brain response to amphetamine. *Proceedings of the National Academy of Sciences*, *100*(10), 6186–6191. doi:10.1073/pnas.0931309100
Mayer, J. S., Bittner, R. A., Nikolić, D., Bledowski, C., Goebel, R., & Linden, D. E. J. (2007). Common neural substrates for visual working memory and attention. *NeuroImage*, *36*(2), 441–453. doi:10.1016/j.neuroimage.2007.03.007
McElree, B. (1998). Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory and Language*, *38*(2), 225–252. doi:10.1006/jmla.1997.2545
McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(3), 817–835.
McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(1), 103–107. doi:10.1038/nn2024
Mehta, M. A., Owen, A. M., Sahakian, B. J., Mavaddat, N., Pickard, J. D., & Robbins, T. W. (2000). Methylphenidate enhances working memory by modulating discrete frontal and parietal lobe regions in the human brain. *Journal of Neuroscience*, *20*(6), RC65.
Mercuri, N. B., Saiardi, A., Bonci, A., Picetti, R., Calabresi, P., Bernardi, G., & Borrelli, E. (1997). Loss of autoreceptor function in dopaminergic neurons from dopamine D2 receptor deficient mice. *Neuroscience*, *79*(2), 323–327.
Mier, D., Kirsch, P., & Meyer-Lindenberg, A. (2010). Neural substrates of pleiotropic action of genetic variation in COMT: a meta-analysis. *Molecular Psychiatry*, *15*(9), 918–927. doi:10.1038/mp.2009.36
Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*(16), 5154–5167.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. doi:10.1146/annurev.neuro.24.1.167
Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425.
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. doi:10.1006/cogp.1999.0734
Moghaddam, B., & Bunney, B. S. (1993). Depolarization inactivation of dopamine neurons: terminal release characteristics. *Synapse (New York, N.Y.)*, *14*(3), 195–200. doi:10.1002/syn.890140302
Moran, R. J., Symmonds, M., Stephan, K. E., Friston, K. J., & Dolan, R. J. (2011). An in vivo assay of synaptic function mediating human cognition. *Current Biology : CB*, *21*(15), 1320–1325. doi:10.1016/j.cub.2011.06.053
Morris, N., & Jones, D. M. (1990). Memory updating in working memory: the role of the central executive. *British Journal of Psychology*, *81*(2), 111–121. doi:10.1111/j.2044-8295.1990.tb02349.x
Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, *66*(3), 241–251.
Murty, V. P., Sambataro, F., Radulescu, E., Altamura, M., Iudicello, J., Zoltick, B., et al. (2011). Selective updating of working memory content modulates meso-cortico-striatal activity. *NeuroImage*, *57*(3), 1264–1272. doi:10.1016/j.neuroimage.2011.05.006
Murty, V. P., Shermohammed, M., Smith, D. V., Carter, R. M., Huettel, S. A., & Adcock, R. A. (2014). Resting state networks distinguish human ventral tegmental area from substantia nigra. *NeuroImage*, *100*, 580–589. doi:10.1016/j.neuroimage.2014.06.047
Nee, D. E., & Brown, J. W. (2012). Rostral-caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory. *NeuroImage*, *63*(3), 1285–1294. doi:10.1016/j.neuroimage.2012.08.034

Nee, D. E., & Brown, J. W. (2013). Dissociable frontal-striatal and frontal-parietal networks involved in updating hierarchical contexts in working memory. *Cerebral Cortex*, *23*(9), 2146–2158. doi:10.1093/cercor/bhs194

Neufang, S., Akhrif, A., Riedl, V., Förstl, H., Kurz, A., Zimmer, C., et al. (2011). Disconnection of frontal and parietal areas contributes to impaired attention in very early Alzheimer's disease. *Journal of Alzheimer's Disease : JAD*, *25*(2), 309–321. doi:10.3233/JAD-2011-102154

Nicola, S. M., Surmeier, J., & Malenka, R. C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annual Review of Neuroscience*, *23*, 185–215. doi:10.1146/annurev.neuro.23.1.185

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. doi:10.1007/s00213-006-0502-4

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. doi:10.1016/j.tics.2006.07.005

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337. doi:10.1016/S0896-6273(03)00169-7

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*(1), 35–53. doi:10.1196/annals.1390.022

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, N.Y.)*, *304*(5669), 452–454. doi:10.1126/science.1094285

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(38), E3660–9. doi:10.1073/pnas.1305373110

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science (New York, N.Y.)*, *314*(5796), 91–94. doi:10.1126/science.1127242

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(2), 283–328. doi:10.1162/089976606775093909

O'Reilly, R. C., Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1999). A biologically based computational model of working memory. In A. Miyake & P. Shah, *Models of Working Memory* (pp. 375–411). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139174909.014

Oberauer, K. (2003). Selective attention to elements in working memory. *Experimental Psychology*, *50*(4), 257–269. doi:10.1027//1618-3169.50.4.257

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(3), 411–421. doi:10.1037//0278-7393.28.3.411

Oosterwijk, S., Lindquist, K. A., Anderson, E., Dautoff, R., Moriguchi, Y., & Barrett, L. F. (2012). States of mind: emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, *62*(3), 2110–2128. doi:10.1016/j.neuroimage.2012.05.079

Ouden, den, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience*, *30*(9), 3210–3219. doi:10.1523/JNEUROSCI.4458-09.2010

Ouden, den, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*(5), 1175–1185. doi:10.1093/cercor/bhn161

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working

memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59. doi:10.1002/hbm.20131
Owen, A. M., Roberts, A. C., Hodges, J. R., Summers, B. A., Polkey, C. E., & Robbins, T. W. (1993). Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. *Brain*, *116 (Pt 5)*, 1159–1175.
Palladino, P., & Jarrold, C. (2008). Do updating tasks involve updating? Evidence from comparisons with immediate serial recall. *The Quarterly Journal of Experimental Psychology*, *61*(3), 392–399. doi:10.1080/17470210701664989
Palladino, P., Cornoldi, C., De Beni, R., & Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Memory & Cognition*, *29*(2), 344–354.
Passamonti, L., Crockett, M. J., Apergis-Schoute, A. M., Clark, L., Rowe, J. B., Calder, A. J., & Robbins, T. W. (2012). Effects of acute tryptophan depletion on prefrontal-amygdala connectivity while viewing facial signals of aggression. *Biological Psychiatry*, *71*(1), 36–43. doi:10.1016/j.biopsych.2011.07.033
Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, *59*(1), 319–330. doi:10.1016/j.neuroimage.2011.07.039
Penny, W. D., Kilner, J., & Blankenburg, F. (2007). Robust Bayesian General Linear Models. *NeuroImage*, *36*(3), 661–671. doi:10.1016/j.neuroimage.2007.01.058
Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, *6*(3), e1000709. doi:10.1371/journal.pcbi.1000709
Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, *22*(3), 1157–1172. doi:10.1016/j.neuroimage.2004.03.026
Penny, W., Kiebel, S., & Friston, K. (2003). Variational Bayesian inference for fMRI time series. *NeuroImage*, *19*(3), 727–741.
Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.
Phillips, A. G., Ahn, S., & Floresco, S. B. (2004). Magnitude of dopamine release in medial prefrontal cortex predicts accuracy of memory on a delayed response task. *Journal of Neuroscience*, *24*(2), 547–553. doi:10.1523/JNEUROSCI.4653-03.2004
Pinotsis, D. A., & Friston, K. J. (2011). Neural fields, spectral responses and lateral connections. *NeuroImage*, *55*(1), 39–48. doi:10.1016/j.neuroimage.2010.11.081
Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425.
Platt, J. C. (1998). Sequential Minimal Optimization: a fast algorithm for training support vector Machines. In B. Schölkopf, C. J. C. Burges, & A. J. Smola, *Advances in kernel methods support vector learning*. unknown.
Podell, J. E., Sambataro, F., Murty, V. P., Emery, M. R., Tong, Y., Das, S., et al. (2012). Neurophysiological correlates of age-related changes in working memory updating. *NeuroImage*, *62*(3), 2151–2160. doi:10.1016/j.neuroimage.2012.05.066
Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. doi:10.1080/00335558008248231
Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38. doi:10.1016/j.neuroscience.2005.06.005
Rabinovich, M. I., Afraimovich, V. S., & Varona, P. (2010). Heteroclinic binding. *Dynamical Systems*, *25*(3), 433–442. doi:10.1080/14689367.2010.515396
Rabinovich, M., Huerta, R., & Laurent, G. (2008). Transient dynamics for neural processing. *Science (New York, N.Y.)*, *321*(5885), 48–50. doi:10.1126/science.1155564
Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163. doi:10.2307/271063?ref=no-x-route:ad785a588e8bc5314fd974513f9f83e7
Rahnev, D., Rahnev, D., Lau, H., Lau, H., de Lange, F. P., & de Lange, F. P. (2011). Prior Expectation Modulates the Interaction between Sensory and Prefrontal Regions in the Human Brain. *Journal of Neuroscience*, *31*(29), 10741–10748. doi:10.1523/JNEUROSCI.1478-11.2011

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. doi:10.1038/4580

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. doi:10.1037/0033-2909.114.3.510

Redgrave, P., Gurney, K., Gurney, K. N., & Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, *58*(2), 322–339. doi:10.1016/j.brainresrev.2007.10.007

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., et al. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews. Neuroscience*, *11*(11), 760–772. doi:10.1038/nrn2915

Repovš, G., & Barch, D. M. (2012). Working memory related brain network connectivity in individuals with schizophrenia and their siblings. *Frontiers in Human Neuroscience*, *6*, 137. doi:10.3389/fnhum.2012.00137

Rice, M. E., & Cragg, S. J. (2008). Dopamine spillover after quantal release: rethinking dopamine transmission in the nigrostriatal pathway. *Brain Research Reviews*, *58*(2), 303–313. doi:10.1016/j.brainresrev.2008.02.004

Richfield, E. K., Penney, J. B., & Young, A. B. (1989). Anatomical and affinity state comparisons between dopamine D1 and D2 receptors in the rat central nervous system. *Neuroscience*, *30*(3), 767–777.

Richfield, E. K., Young, A. B., & Penney, J. B. (1987). Comparative distribution of dopamine D-1 and D-2 receptors in the basal ganglia of turtles, pigeons, rats, cats, and monkeys. *The Journal of Comparative Neurology*, *262*(3), 446–463. doi:10.1002/cne.902620308

Roopun, A. K., Kramer, M. A., Carracedo, L. M., Kaiser, M., Davies, C. H., Traub, R. D., et al. (2008). Period concatenation underlies interactions between gamma and beta rhythms in neocortex. *Frontiers in Cellular Neuroscience*, *2*, 1. doi:10.3389/neuro.03.001.2008

Rorden, C., Karnath, H.-O., & Bonilha, L. (2007). Improving lesion-symptom mapping. *Journal of Cognitive Neuroscience*, *19*(7), 1081–1088. doi:10.1162/jocn.2007.19.7.1081

Salmon, E., Van der Linden, M., Collette, F., Delfiore, G., Maquet, P., Degueldre, C., et al. (1996). Regional brain activity during working memory tasks. *Brain*, *119 (Pt 5)*, 1617–1625.

SanMiguel, I., Saupe, K., & Schröger, E. (2013). I know what is missing here: electrophysiological prediction error signals elicited by omissions of predicted ”what” but not ”when.” *Frontiers in Human Neuroscience*, *7*, 407. doi:10.3389/fnhum.2013.00407

Santangelo, V., & Macaluso, E. (2013). The contribution of working memory to divided attention. *Human Brain Mapping*, *34*(1), 158–175. doi:10.1002/hbm.21430

Schlösser, R. G. M., Wagner, G., Schachtzabel, C., Peikert, G., Koch, K., Reichenbach, J. R., & Sauer, H. (2010). Fronto-cingulate effective connectivity in obsessive compulsive disorder: a study with fMRI and dynamic causal modeling. *Human Brain Mapping*, *31*(12), 1834–1850. doi:10.1002/hbm.20980

Schmidt, B. K., Vogel, E. K., & Woodman, G. F. (2002). Voluntary and automatic attentional control of visual working memory. *Perception & …*.

Schmiedek, F., Hildebrandt, A., Lövden, M., Lindenberger, U., & Wilhelm, O. (2009). Complex span versus updating tasks of working memory: the gap is not that deep. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 1089–1096. doi:10.1037/a0015730

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., et al. (2013). PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*(3), 319–337. doi:10.1007/s12021-013-9178-1

Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual Review of Neuroscience*, *30*, 259–288. doi:10.1146/annurev.neuro.28.061604.135722

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473–500. doi:10.1146/annurev.neuro.23.1.473
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, *275*(5306), 1593–1599.
Sesack, S. R., Aoki, C., & Pickel, V. M. (1994). Ultrastructural localization of D2 receptor-like immunoreactivity in midbrain dopamine neurons and their striatal targets. *Journal of Neuroscience*, *14*(1), 88–106.
Sesack, S. R., Hawrylak, V. A., Matus, C., Guido, M. A., & Levey, A. I. (1998). Dopamine axon varicosities in the prelimbic division of the rat prefrontal cortex exhibit sparse immunoreactivity for the dopamine transporter. *Journal of Neuroscience*, *18*(7), 2697–2708.
Sestieri, C., Corbetta, M., Spadone, S., Romani, G. L., & Shulman, G. L. (2014). Domain-general signals in the cingulo-opercular network for visuospatial attention and episodic memory. *Journal of Cognitive Neuroscience*, *26*(3), 551–568. doi:10.1162/jocn_a_00504
Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: a neuropsychological study. *Quarterly Journal of Experimental Psychology*, *22*(2), 261–273. doi:10.1080/00335557043000203
Sharp, T., Zetterström, T., & Ungerstedt, U. (1986). An in vivo study of dopamine release and metabolism in rat brain regions using intracerebral dialysis. *Journal of Neurochemistry*, *47*(1), 113–122.
Slater, M. (2013). Lagrange multipliers revisited. In *Traces and Emergence of Nonlinear Programming* (pp. 293–306). Basel: Springer Basel. doi:10.1007/978-3-0348-0439-4_14.
Smiley, J. F., Levey, A. I., Ciliax, B. J., & Goldman-Rakic, P. S. (1994). D1 dopamine receptor immunoreactivity in human and monkey cerebral cortex: predominant and extrasynaptic localization in dendritic spines. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(12), 5720–5724.
Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. doi:10.1016/j.tics.2013.12.001
Steiner, H., & Tseng, K. Y. (2010). *Handbook of Basal Ganglia Structure and Function*. Academic Press.
Stelzel, C., Fiebach, C. J., Cools, R., Tafazoli, S., & D'Esposito, M. (2013). Dissociable fronto-striatal effects of dopamine D2 receptor stimulation on cognitive versus motor flexibility. *Cortex*, *49*(10), 2799–2811. doi:10.1016/j.cortex.2013.04.002
Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., Ouden, den, H. E. M., Breakspear, M., & Friston, K. J. (2008). Nonlinear dynamic causal models for fMRI. *NeuroImage*, *42*(2), 649–662. doi:10.1016/j.neuroimage.2008.04.262
Stephan, K. E., Marshall, J. C., Penny, W. D., Friston, K. J., & Fink, G. R. (2007a). Interhemispheric integration of visual processing during task-driven lateralization. *Journal of Neuroscience*, *27*(13), 3512–3522. doi:10.1523/JNEUROSCI.4766-06.2007
Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009a). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. doi:10.1016/j.neuroimage.2009.03.025
Stephan, K. E., Penny, W. D., Moran, R. J., Ouden, den, H. E. M., Daunizeau, J., & Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *NeuroImage*, *49*(4), 3099–3109. doi:10.1016/j.neuroimage.2009.11.015
Stephan, K. E., Tittgemeyer, M., Knösche, T. R., Moran, R. J., & Friston, K. J. (2009b). Tractography-based priors for dynamic causal models. *NeuroImage*, *47*(4), 1628–1638. doi:10.1016/j.neuroimage.2009.05.096
Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., & Friston, K. J. (2007b). Comparing hemodynamic models with DCM. *NeuroImage*, *38*(3), 387–401. doi:10.1016/j.neuroimage.2007.07.040
Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006).

Predictive codes for forthcoming perception in the frontal cortex. *Science (New York, N.Y.)*, *314*(5803), 1311–1314. doi:10.1126/science.1132028
Surmeier, D. J., Song, W. J., & Yan, Z. (1996). Coordinated expression of dopamine receptors in neostriatal medium spiny neurons. *Journal of Neuroscience*, *16*(20), 6579–6591.
Svensén, M., Kruggel, F., & Benali, H. (2002). ICA of fMRI group study data. *NeuroImage*, *16*(3 Pt 1), 551–563.
Tan, H.-Y., Chen, Q., Goldberg, T. E., Mattay, V. S., Meyer-Lindenberg, A., Weinberger, D. R., & Callicott, J. H. (2007). Catechol-O-methyltransferase Val158Met modulation of prefrontal-parietal-striatal brain systems during arithmetic and temporal transformations in working memory. *Journal of Neuroscience*, *27*(49), 13393–13401. doi:10.1523/JNEUROSCI.4041-07.2007
Taverna, S., Ilijic, E., & Surmeier, D. J. (2008). Recurrent collateral connections of striatal medium spiny neurons are disrupted in models of Parkinson's disease. *Journal of Neuroscience*, *28*(21), 5504–5512. doi:10.1523/JNEUROSCI.5493-07.2008
Tehan, G., & Humphreys, M. S. (1995). Transient phonemic codes and immunity to proactive interference. *Memory & Cognition*, *23*(2), 181–191.
Tehan, G., & Humphreys, M. S. (1996). Cuing effects in short-term recall. *Memory & Cognition*, *24*(6), 719–732.
Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle, *Cognitive theory* (Vol. 3, pp. 200–239).
Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge [Cambridgeshire] ; New York : Cambridge University Press.
Trantham-Davidson, H., Neely, L. C., Lavin, A., & Seamans, J. K. (2004). Mechanisms underlying differential D1 versus D2 dopamine receptor regulation of inhibition in prefrontal cortex. *Journal of Neuroscience*, *24*(47), 10652–10659. doi:10.1523/JNEUROSCI.3179-04.2004
Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.
Tseng, K. Y., & O'Donnell, P. (2004). Dopamine-glutamate interactions controlling prefrontal cortical pyramidal cell excitability involve multiple signaling mechanisms. *Journal of Neuroscience*, *24*(22), 5131–5139. doi:10.1523/JNEUROSCI.1021-04.2004
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. doi:10.1006/nimg.2001.0978
van Leeuwen, T. M., Ouden, den, H. E. M., & Hagoort, P. (2011). Effective connectivity determines the nature of subjective experience in grapheme-color synesthesia. *Journal of Neuroscience*, *31*(27), 9879–9884. doi:10.1523/JNEUROSCI.0569-11.2011
van Schouwenburg, M. R., Ouden, den, H. E. M., & Cools, R. (2010). The human basal ganglia modulate frontal-posterior connectivity during attention shifting. *Journal of Neuroscience*, *30*(29), 9910–9918. doi:10.1523/JNEUROSCI.1111-10.2010
Veltman, D. J., Rombouts, S. A. R. B., & Dolan, R. J. (2003). Maintenance versus manipulation in verbal working memory revisited: an fMRI study. *NeuroImage*, *18*(2), 247–256. doi:10.1016/S1053-8119(02)00049-6
Vijayraghavan, S., Wang, M., Birnbaum, S. G., Williams, G. V., & Arnsten, A. F. T. (2007). Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. *Nature Neuroscience*, *10*(3), 376–384. doi:10.1038/nn1846
Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500–503. doi:10.1038/nature04171
Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences of the United States of America*,

108(51), 20754–20759. doi:10.1073/pnas.1117807108
Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: a meta-analysis. *Cognitive, Affective & Behavioral Neuroscience*, *3*(4), 255–274.
Wang, J., & O'Donnell, P. (2001). D(1) dopamine receptors potentiate nmda-mediated excitability increase in layer V prefrontal cortical pyramidal neurons. *Cerebral Cortex*, *11*(5), 452–462.
Weinshilboum, R. M., Otterness, D. M., & Szumlanski, C. L. (1999). Methylation pharmacogenetics: catechol O-methyltransferase, thiopurine methyltransferase, and histamine N-methyltransferase. *Annual Review of Pharmacology and Toxicology*, *39*(1), 19–52. doi:10.1146/annurev.pharmtox.39.1.19
Williams, G. V., & Goldman-Rakic, P. S. (1995). Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature*, *376*(6541), 572–575. doi:10.1038/376572a0
Worsley, K. J., Cao, J., Paus, T., Petrides, M., & Evans, A. C. (1998). Applications of random field theory to functional connectivity. *Human Brain Mapping*, *6*(5-6), 364–367.
Yamamoto, J., Suh, J., Takeuchi, D., & Tonegawa, S. (2014). Successful execution of working memory linked to synchronized high-frequency gamma oscillations. *Cell*, *157*(4), 845–857. doi:10.1016/j.cell.2014.04.009
Yang, C. R., & Seamans, J. K. (1996). Dopamine D1 receptor actions in layers V-VI rat prefrontal cortex neurons in vitro: modulation of dendritic-somatic signal integration. *Journal of Neuroscience*, *16*(5), 1922–1935.
Yu, Y., Fitzgerald, T. H. B., & Friston, K. J. (2013). Working memory and anticipatory set modulate midbrain and putamen activity. *Journal of Neuroscience*, *33*(35), 14040–14047. doi:10.1523/JNEUROSCI.1176-13.2013
Zahrt, J., Taylor, J. R., Mathew, R. G., & Arnsten, A. F. (1997). Supranormal stimulation of D1 dopamine receptors in the rodent prefrontal cortex impairs spatial working memory performance. *Journal of Neuroscience*, *17*(21), 8528–8535.
Zanto, T. P., Chadick, J. Z., & Gazzaley, A. (2014). Anticipatory alpha phase influences visual working memory performance. *NeuroImage*, *85*, 794–802. doi:10.1016/j.neuroimage.2013.07.048
Zheng, T., & Wilson, C. J. (2002). Corticostriatal combinatorics: the implications of corticostriatal axonal arborizations. *Journal of Neurophysiology*, *87*(2), 1007–1017.
Zink, C. F., Pagnoni, G., Martin, M. E., Dhamala, M., & Berns, G. S. (2003). Human striatal response to salient nonrewarding stimuli. *Journal of Neuroscience*, *23*(22), 8092–8097.