Biologically plausible models of sequential action selection

Author: Jennifer Mary Lewis

Submitted for: PhD

Department of Psychology, University of Sheffield
Sheffield, S10 2TP, UK

Date of submission: September 2012

...it invariably delivered a cupful of liquid that was almost, but not quite,
entirely unlike tea.

Douglas Adams

The Hitchhiker's Guide to the Galaxy

Acknowledgements

It is difficult to express my gratitude to my supervisor, Professor Kevin Gurney, for the tremendous support he has offered during my PhD studies. His ever calm and patient guidance has been invaluable and I am immensely thankful to him for all his advice and reassurance over the last four years. Thanks are also due to my second supervisors: Dr. Ashvin Shah, for his extremely helpful thoroughness and attention to detail, and Dr. Jon Chambers, for the single most important piece of advice I have ever received.

There are too many friends who have offered tea and sympathy, or just a good distraction, to thank them all by name, though their contribution to my sanity is sincerely appreciated. A few, though, deserve a particular mention. Thanks to Alex and Javier, for the cheerful tech support, and to Martin, for his sage wisdom and (sometimes brutal, but always helpful) honesty. Very special thanks go to Dave, Donny, Mike, Louise, and of course, Sam. The last four years have been just cracking, and I could not have wished for a more fun or understanding cohort to experience them with.

Finally, my love and eternal gratitude are due to my wonderful family. Brilliant and bonkers in equal measure, their unfaltering belief in my abilities and their willingness to supply me with good wine, a shoulder to cry on, and a hefty shove in the right direction, have in no small way pulled me through the toughest times. Without them, this thesis would simply not exist.

Abstract

The performance of routine action sequences, such as tea-making, tooth-brushing, and driving to work, constitutes a significant proportion of human behaviour, and has received much attention in the cognitive psychology literature. However, an explanation of the cognitive processes underlying this routine sequential behaviour that takes neuroanatomical and neurophysiological principles into account is elusive.

It is widely accepted that the basal ganglia, an interconnected group of subcortical nuclei, are responsible for action selection. The basal ganglia are arranged in topographical 'loops' with cortex and thalamus, forming a series of recurrent circuits in limbic, associative and motor territories of the brain. Traditionally, these loops have been considered to be largely segregated, with little communication between adjacent territories. However, increasing neuroanatomical evidence suggests that these loops are interconnected by several means, forming a hierarchy through which limbic and associative information may influence action selection in motor territories. Furthermore, associative regions of this system appear to be heavily involved in organising routine action. This hierarchy of interconnected circuits is thus a likely neural substrate for the production of routine action sequences.

In this thesis, we build on existing computational models of action selection in the basal ganglia to develop a model of multiple basal ganglia-thalamocortical loops occupying associative and motor territories of the brain, which is applied to the tasks of tea- and coffee-making. Simulation results suggest a novel interpretation of cognitively focused accounts of sequential performance, and reconcile several important issues between two existing competing computational models of routine behaviour. Erroneous behaviour made by the model under conditions of disruption shows similar trends to those observed in studies of action slips and action disorganisation syndrome, and provides support for the hypothesis that damaged temporal order knowledge and action schemas underlie many of the errors typically performed by humans.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sequential behaviour

Human behaviour is frequently discussed in terms of 'action', but rarely is it defined precisely what is meant by this term. After Searle (1980), it is reasonable to consider an action as a unit of behaviour which, when performed correctly, brings about some intended or, perhaps more importantly, predictable consequence, however small. When defined in these terms, almost all human action may be regarded as essentially serial in nature; rarely are single actions performed completely in isolation from one another in a naturalistic environment. Accordingly, a great deal of research has concentrated on the learning and production of action sequences in humans and primates, from a wide range of fields including neuropsychology, neurophysiology, cognitive psychology, and computational neuroscience.

1.1.1 Types of action sequence

Though behaviour may be considered intrinsically sequential, it is far from plausible that all sequential action is mediated by the same neural or cognitive mechanisms (Courtney, 2004; Fuster, 2001; Rhodes, Bullock, Verwey, Averbeck, & Page, 2004). It is therefore important to recognise the vast variety of types of sequence that comprise the repertoire of human performance, and to note that these often have strikingly different demands and are likely to be mediated by different cognitive and neural mechanisms. For instance, actions and the sequences they compose may be goal directed or sensory evoked (Dickinson, 1985;

Redgrave et al., 2010), behaviour types which have long been theorised to rely on different functional and structural substrates (Balleine, Liljeholm, & Ostlund, 2009; Bornstein & Daw, 2011; Daw, Niv, & Dayan, 2005; Redgrave et al., 2010; Seger & Spiering, 2011; Tricomi, Balleine, & O'Doherty, 2009; Yin et al., 2009). Equally, innate sequences have been suggested to rely on intrinsically different mechanisms from those mediating acquired sequences (Berns & Sejnowski, 1998; Cromwell & Berridge, 1996; Graybiel, 2008). Sequences may be implicitly or explicitly learned with implications for the subsequent neural substrate (Ashe, Lungu, Basford, & Lu, 2006; Bayley, Frascino, & Squire, 2005), and evidence suggests that even practice structure may affect the resulting underlying structural and functional mechanisms mediating the performance of a task (Kantak, Sullivan, Fisher, Knowlton, & Winstein, 2010). It is thus important for any study of sequential behaviour to define which type of sequence it intends to address.

While it is beyond the scope of the current discussion to attempt to categorise all sequence types (though see Rhodes et al.(2004) for an excellent overview of several sequential paradigms and their implications for understanding sequential performance), as an illustration of the diversity of sequential processing, here we outline two delineable categories which appear to have differing demands, and accordingly, may rely on different functional mechanisms.

Immediate Serial Recall

A great deal of research examining the nature of sequence production in humans has focused on the *immediate serial recall* (ISR) task (Botvinick & Plaut, 2006a; Rhodes et al., 2004; Tan & Ward, 2008). This involves the brief retention of short lists of items - typically digits or words - and their immediate recall in the same order as their presentation. This paradigm has been used to study sequential performance largely due to its simple application in experiments and its convenience, given that participants are equally naive, having no prior experience of the specific sequences to be remembered and reproduced.

Performance on this task imposes particular demands upon working memory, notably the retention of individual items and their order which, critically, are unfamiliar. Moreover, as there are no requirements for storage to or retrieval from long term memory, it is reasonable to suppose that the internal representations underlying this type of task are dynamic (Botvinick & Plaut, 2006a; Duncan, 2001). Additionally, as there is no external physical constraint on temporal ordering, nor cues resulting from one item directly indicating the identity of the next, the participant must rely entirely on internal factors for generating the correct sequence; the ‘load’ on working memory is thus brief, but relatively high. ISR tasks consequently appear to rely on rehearsal techniques (Tan & Ward, 2008), and unsurprisingly, participants tend to display primacy and recency effects (Henson, 1998).

Routine action

‘Routine’ action describes a qualitatively different form of sequential performance. Commonly described as *activities of daily living* (ADLs), this category comprises sequences of actions which are familiar and well learned, but which often incorporate a degree of flexibility in their temporal ordering, thus standing in contrast to ISR or similar tasks. For instance, a commonly cited example of an ADL is toothbrushing; here, it is equally valid to rinse a toothbrush before brushing one’s teeth, or not. Equally, one may or may not choose to rinse a toothbrush after brushing one’s teeth. ‘Rinsing’ is therefore a highly flexible component of the toothbrushing sequence.

Despite this flexibility, ADLs tend to consist of stereotypical sequences of actions taking roughly the same form with each performance. However, given that such tasks are generally performed at a different temporal level of description than ISR - on the order of minutes or hours rather than seconds - dynamic environmental factors often demand different performance, such as when interruptions occur, or due to some extraneous factor a typically performed action might need to be omitted, or an unusual one added. The ability to vary the precise temporal order of a sequence of actions is thus not optional for routine performance. Furthermore, ADLs require the retrieval from long-term memory of the required steps for task completion (Courtney, 2004), and their temporal order; ADLs therefore impose demands upon storage and retrieval processes that are distinct from those required for ISR. Thus, the nature of the underlying representations for the performance of routine sequences may also be different from those for ISR, possibly even taking a consistent neural form during each execution, in the manner of action 'schemas' (see section 1.2.2).

1.2 Focus of the thesis: routine action

From the above, it is clear that all sequential action may by no means be accounted for by a single mechanism, and that different sequential tasks impose significantly different demands on cognitive processing. It is possible, even likely, that such different tasks, while each 'sequential', have fundamentally different functional and neural underpinnings. In this thesis, therefore, we do not claim to be investigating the principles underlying sequential action as a whole. Rather, we take a specific focus on the mechanisms of routine action. In doing so, however, it is pertinent to consider the implications of any insights for various types of sequential process, and whether any principles may indeed be applied to sequential processing in general.

A focus on routine action, rather than the more commonly studied ISR, is desirable for several reasons. Firstly, ADLs are by definition 'naturalistic' action, which cannot always be said for experimental tasks designed for investigating sequencing. Thus, an understanding of the responsible mechanisms has potentially greater implications for understanding human behaviour in general than for more artificially constructed laboratory tasks (Ryou & Wilson, 2004). Several cognitive accounts of routine performance do exist, including computational models (Botvinick & Plaut, 2004; Cooper & Shallice, 2000; see below). However, rarely have these accounts reconciled the cognitive explanations they invoke with research examining the precise neural underpinnings of the learning and performance of either goal-directed or habitual sequences (Balleine et al., 2009; Dezfouli & Balleine, 2012; Daw et al., 2005; Redgrave et al., 2010; Yin & Knowlton, 2006), possibly since routine action may occupy some intermediate state between truly goal-directed and habitual performance that does not lend itself easily to interpretation as one or the other (see section 1.3.5). Further investigation in this area also has important implications for an existing debate over the functional basis of routine action arising from this modelling work (Botvinick & Plaut, 2006b, 2006c; Cooper & Shallice, 2006a, 2006b).

1.2.1 Error behaviour

It is widely acknowledged that the study of typical errors made during performance can reveal much about the organisation of information for action (Henson, 1996; Rhodes et al., 2004); indeed, typical error types are believed to stem directly from the very cognitive processes which mediate correct performance (Reason, 1990). Error commission is regarded as a normal feature of healthy performance, and is a fruitful area for study. Unfortunately, while not impossible (Botvinick & Bylsma, 2005), it is difficult to study performance of ADLs in an experimental setting. As a result, there is little controlled data in this area. In an influential series of diary studies however, Reason (1979, 1984, 1990) obtained from healthy participants detailed descriptions of errors they made during their daily lives, with a particular focus on actions which were discordant with the original plan or intention, described as action 'slips' (Norman, 1981). Analysis of these errors revealed several common features in the types of error committed. Errors tended to occur during well practised tasks, and at obvious 'branch points' within a sequence where, depending on the overall context of the sequence, multiple subsequent actions were valid (Reason, 1979, 1984). Errors tended to consist of the intrusion of actions which belonged to sequences similar in nature to the intended one (Norman, 1981), and that were frequently - and often recently - performed (Reason, 1984). Where this was the case, errors generally took the form of well formed action sequences or subsequences (Reason, 1979). For example, this description was given by a participant of Reason's 1979 diary study:

> 'I have two mirrors on my dressing table. One I use for making up and brushing my hair, the other for inserting and removing my contact lenses. On this occasion I intended to brush my hair, but sat down in front of the wrong mirror, and removed my contact lenses instead' (Reason, 1979; p71).

Errors also consisted to a great extent of omitted actions, the usage of erroneous objects, and repetitions of a previously performed action (Reason, 1984).

Further insights may be gained from neuropsychology, where brain injury may result in qualitatively or quantitatively different error patterns from those observed in healthy controls, allowing further insights to be gained at a cognitive level, and also, to a limited extent, pointing to the neural basis of particular cognitive processes. Damage to the frontal lobes frequently has a disruptive effect on the performance of ADLs (Humphreys & Forde, 1998; Luria, 1965; Schwartz et al., 1998). Such damage generally affects not only the frequency of error commission on such tasks, but also the relative rates of the types of error that are observed. However, striking similarities in the error commission of patients with differing brain injuries have been reported. Injury to either hemisphere of the frontal lobes, for instance, has been shown to result in a stereotypical profile of error now termed ‘action disorganisation syndrome’ (ADS) (Buxbaum, Schwartz, & Montgomery, 1998; Schwartz et al., 1999). Patients displaying ADS have been shown to have impairments over and above those of so called ‘executive’ dysfunction (Humphreys & Forde, 1998) which is also associated with frontal lobe injury, and manifests on tests of abstract reasoning, set switching, response inhibition and directed attention (Jurado & Rosselli, 2007). Rather, or in addition, ADS sufferers tend to have specific problems composing naturalistic action sequences that are common examples of ADLs.

Consistencies that have been observed across patients include a general decline in the coherence of behaviour with increasing disorder severity (Schwartz, Reed, Montgomery, Palmer, & Mayer, 1991; Schwartz et al., 1998). In particular, single actions are increasingly performed outside the boundaries of distinct subsequences, often manifesting as ‘toying’ behaviour, a term describing the repeated picking up and putting down of objects without their functional use towards any goal or subgoal. A particularly strong finding is the tendency to omit actions or subtasks in greater frequency than other types of error (Humphreys & Forde, 1998; Schwartz & Buxbaum, 1997; Schwartz et al., 1998). Omissions also tend to increase in relative frequency with increasing severity or in the presence of distractor objects (Morady & Humphreys, 2009; Schwartz et al., 1998), though not necessarily as a general function of load or task difficulty (Forde & Humphreys, 2002). Object substitutions, whereby an action is completed with an erroneous object, are also common (Schwartz et al., 1991, 1995).

Despite these notable commonalities however, significant variability is observed between patients and tasks (Humphreys & Forde, 1998; Schwartz et al., 1991), and even between different instances of the same task (Schwartz et al., 1995), though different *types* of errors tend to manifest during different types of task (Forde & Humphreys, 2002; Schwartz et al., 1991). In particular, a dissociation between ‘recurrent’ and ‘continuous’ perseveration types has also been explicitly observed (Forde & Humphreys, 2002; Humphreys & Forde, 1998), where these terms distinguish the repetition of an action or subsequence after one or more intervening actions (recurrent perseveration), from the immediate repetition of an action (continuous perseveration) (Sandson & Albert, 1984). This variability implies that a number of processes are required for the performance of routine sequences which may be selectively disrupted to a greater or lesser extent; the question remains, however, as to whether a single particular deficit is responsible for the common findings across patients (see chapter 5).

1.2.2 Interpretation of sequencing mechanisms

The vast majority of explanations of the mechanisms underlying routine sequential action, and the typical errors made during its performance, are focused at the cognitive level. Early interpretations of sequencing behaviour in general focused on ‘associative chaining’ accounts. These suggested a means of sequencing whereby the execution of one action in a sequence automatically triggers the next, either by internal associations between actions or in a stimulus-response fashion. Lashey (1951) famously criticised this suggestion, pointing out its omission of contextual information that might be vital for the production of the correct sequence. More recently, and particularly for certain types of sequence (most notably, serial recall tasks such as ISR), it has been shown that this model is indeed insufficient, due largely to its inability to account for error data on such tasks. For example, error patterns demonstrate a tendency to switch the positions of two successive items in a list; a finding which chaining based models of sequencing have failed to reproduce (Henson, 1996).

In contrast to this view however, some authors suggest that well learned tasks may indeed be organised as ‘chains’, particularly for simple, automated sequences with little overall structure other than linear ordering, and where sequence elements are encoded in a context-sensitive fashion (Fuster, 2001; Reason, 1979; Wickelgren, 1969), though it is likely that such routine tasks are far more simple than ADLs which require a certain degree of planning, however routine. For instance, ADLs appear to be organised hierarchically and may require representation at several levels of a conceptual action hierarchy (Fuster, 2001; Reason, 1979) which a simple chaining account may struggle to capture. Indeed, an intrinsic hierarchical structure of action is often assumed for routine tasks (Badre, 2008; Botvinick, 2007; Cooper & Shallice, 2006a), and forms the basis for many accounts of action sequencing (see figure 1.1). Lower levels of the action hierarchy correspond to 'composite actions', which may be executed as part of several sequences or subsequences. Progressively higher levels of the action hierarchy account for more temporally extended segments of action, accounting for increasingly more distal levels of structure. An overall task, for example, might be composed of multiple subtasks, each requiring the execution of several composite actions.

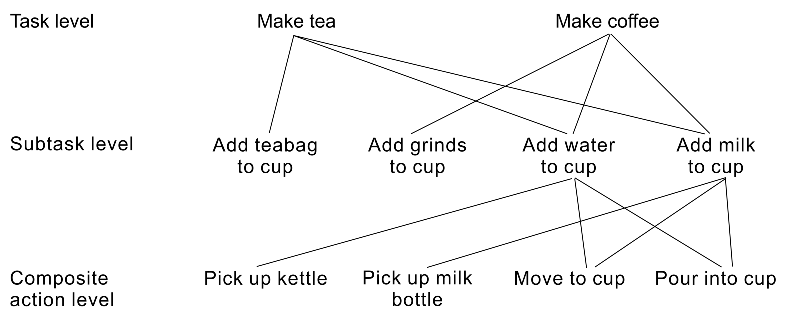

Figure 1.1: A three-level conceptual 'action hierarchy', using commonly cited examples of routine tasks, tea and coffee making (see also Humphreys & Forde, 1998). Note that the same 'subtasks', spanning an intermediate temporal duration, may be utilised as a component of different tasks; equally, the same composite action may be harnessed by multiple subtasks. Note that this hierarchy is not limited to three levels; it is likely that tasks of differing degrees of complexity and temporal duration may have more or fewer hierarchical levels.

Competitive queueing

Insights from work on associative chaining have led to an alternative suggestion known as *competitive queueing* (Houghton, 1990). Rather than the serial activation of action representations corresponding to the actions required for the sequence, at the beginning of a particular sequence the internal representation of each component action in that sequence is activated *in parallel*, to a degree that is proportional to its temporal distance from the beginning of the sequence. This activation 'gradient' across component action representations is imposed by higher level units, which correspond to a higher level of the action hierarchy as described above. As each composite action is executed, its corresponding internal representation is subsequently inhibited, allowing the next action representation to receive levels of activation sufficient for its selection. This model has been extremely influential and is strong in as much as it has been able to successfully model co-articulation effects in speech (Houghton, 1990), has received support from electrophysiological work in primates (Averbeck, Chafee, Crowe, & Georgopoulos, 2002), and, notably, is able to cope with repeated actions in a single sequence (Houghton, 1990), despite early parallel activation models struggling with this function (Jordan, 1986).

However, competitive queueing might be inflexible to some extent, and thus more applicable to sequences with a fixed temporal order (such as the articulation for which the original model was proposed) as, like chaining accounts, competitive queueing does not naturally take nuanced temporal contextual information into account beyond the level of the overall sequence. Also, as pointed out by Botvinick & Plaut (2004), the nature of the sequencing mechanism would struggle to perform higher order tasks with more than two hierarchical levels of action. As such, for complex tasks where flexibility is allowed or even required depending on environmental factors, as is likely in ADLs as described above, the pre-determined sequencing of the competitive queueing account might be insufficient for control. Rather, a more dynamically adaptive mechanism seems more likely to be able to cope with flexible temporal orders and interruptions. Additions to the competitive queueing account have incorporated dynamic updating of action plans (see Rhodes et al., 2004), but as acknowledged by the authors, the level of additional complexity required for this function is potentially a concern.

Schemas

Where explanations of sequencing focus specifically on routine action, the vast majority of interpretations are provided in terms of action *schemas*. The term 'schema' was originally used to refer to the internal representation of sets of stereotypical features of categories of objects in the world, in order to aid perception. Later, the term was applied to motor actions, as a specification for the initiating conditions and production of an action (see Schmidt (1975) for an overview). The term remains in wide usage, though its general meaning has evolved to represent a more abstract idea. However, the concept is rarely operationalised, and as such, it is not always entirely clear to what the term refers. However, common to many accounts is the suggestion that a schema is a cognitive structure representing a set of features required for the direction and performance of a sequence, subsequence, or action.

Importantly, and consistent with descriptions of a hierarchy of action, schemas are proposed to exist at several levels of description, with higher level or 'intention' schemas activating lower level or 'child' schemas, possibly with intermediate levels. For example, a 'motor response schema' (Schmidt, 1975) describes a fairly context non-specific representation of the necessary parameters for the execution of a single action, and the conditions governing its execution. Similarly, Norman (1981) describes schemas simply as 'sensori-motor knowledge structures', suggesting control at a low level of abstraction. According to Cooper & Shallice (2000), low level, discretely represented motor schemas effectively compose higher level schemas or 'skill units', which capture stereotypical features of higher level task units, such as subsequences. Indeed, Shallice (1982) discusses schema at this higher task level, giving examples of schemas which may control 'doing long division, making breakfast, or finding one's way home from work' (p199), whereas the breadth of information that schemas may include is captured by Eysenck & Keane (2005), who cite the representations of 'generic knowledge ... events, sequences of events, percepts, situations, relations, and even objects' (p276).

Many authors describe routinised behaviour as the sequential activation of discrete schemas, where schemas are responsible for triggering the selection of appropriate motor representations of actions. This view will be the focus of much of this thesis; for our present purposes, schemas are considered to be cognitive structures which represent the actions required for the ongoing task, as well as the relevant goals and subgoals, and other relevant task-specific information where applicable.

Contention scheduling

According to a particularly influential account of action based on schema dynamics (originally Norman & Shallice, 1986; later Norman & Shallice (2000)), action selection occurs as a result of the activation of the corresponding action schema beyond some threshold. Selection of action schemas is carried out by a 'contention scheduling' system (CSS), which operates largely outside conscious control, and relies on mutual inhibitory links preventing the simultaneous selection of incompatible and competing schemas. Sequencing occurs through the satisfaction of pre- and post-conditions for the activation of schemas; activation of action schemas arises from features of the external environment and from excitation originating in higher level 'source schemas', which may be considered high level skill units, or schemas at the highest level of the action hierarchy.

For well learned or routine tasks, the CSS operates in a largely automatised manner. However, for novel tasks or sequences where relevant high level schemas or scripts have not been formed, a second system, known as the 'supervisory attentional system' (SAS) biases the selection of low level action schemas, embodying 'attention to action', or deliberate conscious control over the actions being selected (Norman & Shallice, 2000).

The theory thus posits two types of behaviour control, routine and non-routine, which are analogous to the automatic and controlled processing modes, respectively (Shiffrin & Schneider, 1977). Whereas the CSS is responsible for selection in both processes, the SAS is involved only in non-routine selection, and only via the biasing of selection, rather than its direct control (Norman & Shallice, 2000; Shallice, 1982). The CSS thus embodies the schema selection mechanism, which operates autonomously for well-learned tasks, but which may be biased by the SAS in non-routine processing (Shallice, 1982). This theory is similar to competitive queueing accounts in several ways, particularly in its suggestion that the activation of a single high-level schema simultaneously sends activation to all lower level schemas corresponding to those actions required for the task (parallel activation). However, it has a certain level of added flexibility in that it takes account of environmental or 'bottom-up' influences on action selection, and does not predetermine the activation order of each individual action, and thus need not add additional functionality for the implementation of dynamic updating of action plans.

Accounting for errors

Several authors interpret action slips and ADS in terms of some type of disruption to action schemas and the dynamics of the CSS. Norman (1981), for example, suggests that various action slips in normal performance result from the faulty triggering or activation of schemas, or failure to activate at all.

Schwartz and colleagues' early ADS case studies (Schwartz et al., 1991) suggested that the main deficit in ADS was the 'weakening of top-down formulation of action plans' (p409), resulting in a vulnerability to the inappropriate triggering of task-incongruent action schemas by sensory stimuli. This was also said to lead to interference between multiple action plans resulting in a 'blending' effect, where incompatible components of distinct schemas might be simultaneously activated. Later, more comprehensive studies examining the performance of several patients with closed head injury gave rise to a contrasting suggestion that the pattern of errors in ADS was due to a non-specific reduction in cognitive processing resources (Buxbaum et al., 1998; Schwartz et al., 1998, 1999).

In a series of related case studies, Humphreys and colleagues (Forde & Humphreys, 2002; Forde, Humphreys, & Remoundou, 2004; Humphreys & Forde, 1998; Humphreys, Forde, & Francis, 2000) provided a contrasting perspective on the cognitive mechanisms underlying action disorganisation, after failing to confirm several hypotheses derived from the 'non-specific cognitive resources deficit' theory proposed by Schwartz and colleagues (Forde & Humphreys, 2002; Schwartz et al., 1998). Interestingly, the authors combined a competitive queueing account and the contention scheduling of action schemas in order to interpret error patterns (Humphreys & Forde, 1998). They suggest that a corrupted activation gradient (see section 1.2.2) on component action schemas is manifested as a disruption to temporal order knowledge, causing the faulty ordering of composite action schemas and the subsequent observed breakdown in performance (Humphreys & Forde, 1998; Humphreys et al., 2000). Further, a breakdown in rebound inhibition after action execution was specifically postulated for continuous perseveration errors (Forde & Humphreys, 2002; Forde et al., 2004; Humphreys & Forde, 1998; Humphreys et al., 2000).

1.2.3 Computational models of routine action

A particularly common routine sequence which has been focused on in the neuropsychological literature is that of coffee-making (e.g., Schwartz et al., 1991). In theoretical approaches to understanding the mechanisms underlying normal and disrupted performance, two notable computational models simulating a coffee-making sequence have been presented which propose to capture the predominant features of these functional mechanisms (Botvinick & Plaut, 2004; Cooper & Shallice, 2000). Importantly, these models effectively examine the proposals that errors result from a reduction in top-down control of action schema activation (Cooper & Shallice, 2000; Schwartz et al., 1991), or the reduction of more general processing resources (Botvinick & Plaut, 2004; Schwartz et al., 1998).

Interactive Activation Network model

Cooper and Shallice's (2000) *interactive activation network* (IAN) model closely followed Norman and Shallice's (1986, 2000) theory, and aimed specifically to provide a computational account of the CSS in particular. Moreover, the model examined the effects of the reduction of top-down activation of action schemas, as suggested by Schwartz et al. (1991) as the cause of ADS.

The primary functional component of the model was a network of action schemas. Schemas were represented in a localist fashion, implemented as single nodes, each with a single dynamic value representing its activation, and were interpreted as 'abstractions over goal-directed segments of action', and as 'methods' for achieving those goals. A schema was considered to be selected once its activation exceeded a predetermined threshold. Schemas within the network occupied one of several possible hierarchical levels. Selection of a high-level schema (e.g., 'prepare instant coffee') resulted in the flow of activation to lower level component schemas (e.g., 'add coffee from jar'). The lowest level schemas represented discrete actions (e.g., 'pick up'), the selection of which ultimately caused the execution of the corresponding action by effector systems. The assignment of target objects to particular actions occurred by means of interaction with an 'object network', which was again composed of multiple nodes, each corresponding to the internal representation of an individual object. The object with the most highly active representation in this network corresponded to the 'argument' for the selected action schema, defining the object that was the subject of the selected action.

The authors emphasised the important combination of top-down cognitive control and the automatic, bottom-up triggering of action schemas by environmental factors. Indeed, a key component of the IAN model was a parameter entitled *Internal:External*, the value of which determined the balance of top-down (from higher level and source schemas) and bottom-up (from the external environment) influences on schema selection. Variations in the value of this parameter tested the hypothesis of Schwartz et al. (1991) that errors result from an imbalance of top-down and bottom-up influences on behaviour. This manipulation indeed gave rise to errors in performance which have commonly been observed in studies of action slips and ADS. While rates of particular error types in patients were not replicated, and certain critical types of error, such as recurrent perseveration, were not observed, this model provided strong initial support for the contention scheduling approach to understanding the cognitive dynamics underlying routine behaviour, and for taking a schema-oriented view to understanding performance error. Indeed, later adaptations of the model accounted for error patterns in multiple disorders of action by selectively disrupting various parameters (Cooper, Schwartz, Yule, & Shallice, 2005).

Simple Recurrent Network model

In contrast to popular schema focused theories which explicitly rely on a hierarchy of cognitive processing structures, Rumelhart and colleagues (1986) re-conceptualised a schema as nothing that truly 'exists' as a representational structure in the brain, but an abstract concept that describes the knowledge and rules associated with a familiar situation. Accordingly, a contrasting model of routine action developed by Botvinick and Plaut (2004) took a qualitatively different approach to examining its mediation. This model rejected the necessity for schemas as discrete cognitive structures guiding behaviour. In particular, they suggested that the explicit hierarchical structure in the IAN model resulted in a degree of inflexibility that would cause difficulty in the performance of certain types of task; namely those with what they termed to be a 'quasi-hierarchical' structure, where there exist slight variations of a procedure towards the same basic goal. The authors argued that the localist representations of goals and schemas at high levels of the action hierarchy in the IAN model would struggle to capture important similarities in these variations, making their processing inefficient.

Alternatively, they advocated a recurrent connectionist approach to understanding the production of routine sequences, again focusing on a coffee-making task for comparison with the IAN model and previous observational studies, but also introducing a similar tea-making task. They proposed a *simple recurrent network* (SRN) model consisting of three layers of nodes; an input layer, an output layer, and a fully recurrent intermediate 'hidden' layer. The pattern of activation of the input layer represented the currently held and fixated objects; that of the output layer represented the actions selected by the model. The hidden layer received activation from the input layer in addition to its own recurrent influence, allowing it to maintain a dynamic record of its own activation history as well as the current environmental state. The authors emphasised this feature as a critical distinction from explicitly hierarchical models, allowing the SRN model to retain a representation of the overall task context within the hidden layer. Furthermore, whereas the IAN model embodied a hard-coded structure determining the organisation of the schema hierarchy, the SRN model utilised back-propagation in order to adopt an appropriate pattern of connectivity within the hidden layer to mediate the desired sequences.

Again, the model was able to replicate a wide range of data from observational studies on human error, in both healthy participants and patients. In this case, however, this resulted from the addition of noise to the activation values of the hidden layer nodes, consistent with the suggestion that ADS results from a general reduction in cognitive resources (Buxbaum et al., 1998; Schwartz et al., 1998). Consistent with the human errors data, the model showed a general decline in performance with increasing noise, and in particular, the authors stressed the replication of the 'omission rate effect', demonstrated by several authors (Morady & Humphreys, 2009; Schwartz et al., 1998). Moreover, the model was able to perform three versions of a coffee-making task, each of which required different amounts of sugar, as an example of the 'quasi-hierarchical' task structure described above. They claimed that this performance demonstrated greater efficiency and flexibility in the underlying representations than would be possible with a strict hierarchical structure. Despite these successes, however, certain errors that were commonly reported in the preceding neuropsychological studies were extremely rare (Cooper & Shallice, 2006a), suggesting that more specific deficits may be responsible, at least for certain patterns of error in ADS.

Common weaknesses

These models have taken strikingly different approaches, both theoretically and implementationally. Where one uses a strictly hierarchical, abstract schema based approach (Cooper & Shallice, 2000), the other implements a pointedly non-hierarchical recurrent neural network model (Botvinick & Plaut, 2004). While both of these models have provided tremendous insight into the potential functional underpinnings of routine sequence production, importantly accounting for certain error patterns observed in behavioural studies, notably neither is constrained from a biological perspective, thus the question of whether either approach might reasonably be implemented neurally remains.

While the original theory detailing the CSS and SAS (Norman & Shallice, 1986, 2000), suggested that the SAS might be predominantly located in the frontal lobes, particularly the left hemisphere (Shallice, 1982), the precise form that schemas may take, the mechanisms by which they are selected and deselected, and the means by which they are arranged hierarchically, remains unspecified. Likewise, although Botvinick & Plaut (2004) suggest the involvement of prefrontal cortex and basal ganglia in the neural implementation of the mechanisms they propose, without mapping specific functions to particular regions of the model it is difficult to determine whether such mechanisms might be plausibly embodied by these structures. The opaque functioning of the emergent representations of the hidden layer exacerbates this problem, as does the employment of back-propagation, which is also criticised as being biologically implausible (Grossberg, 1987).

A later implementation of the SRN model (Botvinick, 2007) addressed evidence suggesting a functional hierarchy in prefrontal cortex, with increasingly abstract information being represented at successively more anterior regions of prefrontal cortex (Courtney, 2004; Fuster, 2001; Koechlin, Ody, & Kouneiher, 2003), though the precise nature of this increasing abstraction is unclear (Badre, 2008). This work included multiple, hierarchically arranged hidden layers; results showed that higher layers came to encode increasingly greater amounts of contextual information and less regarding immediate stimulus-response mappings. However, this model might be more accurately described as neurally 'inspired' rather than 'constrained', as beyond the simple hierarchy implemented, no neuroanatomical principles are followed. The implausibility of back-propagation also continues to detract from the model's direct implications for understanding sequence processing in the brain.

1.2.4 Summary

Various accounts of routine sequencing have been proposed, often focusing on typical error patterns produced by healthy participants and ADS patients. Proposals of Schwartz and colleagues (Schwartz et al., 1991, 1998) have been directly or indirectly tested in computational models with significant success, but to date, such models have not taken account of neuroanatomical constraints. Additionally, while the competitive queueing account of ADS (Humphreys & Forde, 1998) provides an intuitively appealing explanation of sequence production and of many of the errors observed, which is consistent with those explaining action slips to an extent (Norman, 1981), it is yet to be tested in a computational model, and therefore remains subject to the more general questions regarding competitive queueing discussed above (section 1.2.2). Clearly, more research is required in order to shed light on some of these issues.

In this thesis, we present a theory of routine sequential action that obeys the neuroanatomical and functional constraints of the likely underlying neural substrate of the production of sequential behaviour. In doing so, we hope to unite cognitive accounts of behaviour with mechanistic and neurally focused ones, in order to further our understanding of the abstract cognitive mechanisms embodied by the neural substrate. Furthermore, by attempting to map the proposed cognitive mechanisms of existing accounts onto the likely neural hardware, it may be possible to reach a reconciliation of the competing models within a biologically plausible architecture.

1.3 Neural substrates of routine sequences

1.3.1 Action selection

The problem of action selection or 'choosing what to do next' refers to how an organism or agent resolves the competition between multiple behavioural options at a given time (Brom & Bryson, 2006; Redgrave, Prescott, & Gurney, 1999; Tyrrell, 1992). It is a problem that must be solved, in a more or less sophisticated manner, by all organisms. Options for action at any time are often incompatible, competing for access to the same cognitive or motor resources, and thus may not be performed simultaneously. A mechanism that gates access to those resources quickly, efficiently and flexibly is vital to adaptive behaviour. Such a mechanism must be able to take into account the external context of the task at hand, as well as internal motivational signals in order to efficiently select an appropriate action from several possible contenders. It must be able to allow continuous engagement with the task until completion, possibly in the face of interference from other demands for attentional or physical resources. However, it must also be able to disengage appropriately, possibly from an ongoing or incomplete action, if faced with a more urgent or important task. It must be amenable to modification, allowing change of strategy if necessary. It should also respond to the learning of frequently performed actions, so with practice, selection and performance become more efficient and less computationally expensive. In vertebrates, is is widely accepted that the basal ganglia are essential for the solution of this problem (Mink, 1996; Redgrave et al., 1999). This group of subcortical nuclei have been shown to display several features which suggest its suitability for this function, including their wide ranging afferent and efferent connectivity with other neural structures, topographical organisation, and internal dynamics resulting from their intrinsic functional architecture.

1.3.2 Basal ganglia thalamocortical loops

Macro- and micro-loop architecture

Here, we provide a brief summary of the architecture of the basal ganglia and its connections with other neural structures. For a more thorough overview the reader is directed to any of several classic reviews (Bolam, Hanley, Booth, & Bevan, 2000; Gerfen & Wilson, 1996; Middleton & Strick, 2000; Mink, 1996; Parent & Hazrati, 1995a; Smith, Bevan, Shink, & Bolam, 1998; Wickens, 1997).

The basal ganglia, taken together, receive massive glutamatergic afferents from almost all regions of cortex, with the exception of primary sensory cortex (Mink, 1996), as well as subcortical components of the limbic system including the hippocampus and amygdala (Groenewegen et al., 1999; Humphries & Prescott, 2010). The efferent projections of basal ganglia predominantly target distinct regions of thalamus (Bolam et al., 2000), and brainstem (McHaffie, Stanford, Stein, Coizet, & Redgrave, 2005). These regions of thalamus are, in turn, reciprocally connected to cortex, forming thalamocortical feedback ‘loops’, and also connecting basal ganglia in loops with cortex (Parent & Hazrati, 1995a), illustrated schematically in figure 1.2.

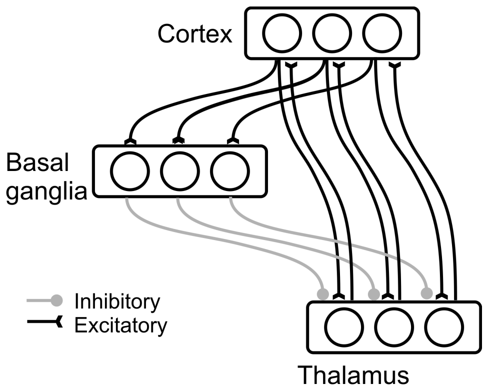

Figure 1.2: Illustration of the basic macro/micro loop architecture of the basal ganglia thalamocortical system. A single macro loop is composed of multiple, largely topographic, micro-loops. See text for details.

This connectivity obeys a topographical scheme whereby the organisation of afferents to basal ganglia is largely maintained through their internal structure and their projections to distinct regions of thalamus, which, in turn, send afferents to the same originating regions of cortex (Alexander, DeLong, & Strick, 1986). This organisation thus results in a series of functionally specific basal ganglia-thalamocortical (BGTC) topographical loops. This topography has been described at a very fine level depicting ‘micro’ loops, to the extent that the distinct representation of specific body parts and limb movements in distinct regions of basal ganglia has been documented (West et al., 1990), following the arrangement of sensorimotor cortices (Alexander & DeLong, 1985). However, BGTC loops are more often considered on a 'macro' scale according to the nature of the information that appears to be processed in wider territories of the overall system. Several schemes have been proposed to account most accurately for this regional distinction of basal ganglia. Alexander and colleagues (1986), for example, proposed five distinct territories for the processing of anterior cingulate, lateral orbitofrontal, dorsolateral prefrontal, oculomotor and motor cortical inputs to basal ganglia. A slightly more parsimonious view was explicated by Parent (1990), in which functional BGTC 'territories' may be broadly classed as limbic, associative and motor, with limbic territories originating in hippocampus, amygdala and orbitofrontal cortices, associative in dorso- and ventrolateral prefrontal cortices, and motor in premotor and primary motor cortices, each impinging primarily upon distinct regions of basal ganglia and thalamus. Broadly speaking, limbic cortices are located more anteriorly within the frontal lobe, with motor cortices occupying posterior regions and associative territories located centrally (Badre, 2008), whereas basal ganglia, particularly striatum, obey a more ventromedial-dorsolateral gradient (Voorn, Vanderschuren, Groenewegen, Robbins, & Pennartz, 2004). This latter, three territories account is better specified, and the view to which the remainder of this thesis will subscribe.

Intrinsic functional architecture of basal ganglia

The general structure and architecture of basal ganglia anatomy is preserved in both the primate and the rat, though certain differences exist which mainly reflect structurally distinct - though functionally homologous - nuclei between the species. Unless otherwise specified, the following discussion will focus on the primate anatomy, and is illustrated schematically in figure 1.3.

The basal ganglia consist of four main nuclei: the striatum, the subthalamic nucleus (STN), the globus pallidus (GP) which, in the primate is further divided in to internal (GPi) and external segments (GPe), and the substantia nigra pars reticulata (SNr). In rodents, the GP is a homogeneous structure, analogous to the primate GPe, whereas a distinct structure, the entopeduncular nucleus (EP) reflects the function of the primate GPi.

The striatum serves as the main input structure of the basal ganglia, and is subdivided into the nucleus accumbens, caudate nucleus and putamen, which serve as its primary limbic, associative and motor components, respectively (Parent, 1990). The striatum is the recipient of afferents from cortex and the limbic system. Corticostriatal projections are collaterals from cortical efferent neurons projecting primarily to brainstem and thalamus, and, particularly for motor territories, are believed to convey efference copies of domain-specific information, such as action representations (Lévesque, Charara, Gagnon, Parent, & Deschênes, 1996; Redgrave et al., 1999). Excitatory afferents to striatum impinge primarily upon the medium spiny neuron (MSN), which itself comprises around 75-95% of striatal neurons (Tepper, Koós, & Wilson, 2004). MSNs also account for the primary projection neuron of striatum, and, being tonically quiescent and GABAergic, impose an inhibitory influence upon their target when sufficiently excited (Crutcher & DeLong, 1984). As indicated by the nomenclature, their dendritic trees show an extensive covering of spines, and branch widely, allowing great convergence of cortical axons onto single MSNs (Gerfen & Wilson, 1996; Mink, 1996). Given this, it is widely suggested that striatum has an important integrative function (Horvitz, 2002).

MSNs in the striatum may be roughly divided into two types; those expressing substance P and the D1-type dopamine receptor, and those expressing enkephalin and the D2-type dopamine receptor. Notably, these neurons project primarily to distinct target structures, giving rise to two main pathways through the basal ganglia, traditionally referred to in a widely accepted scheme as the 'direct' and 'indirect' pathways (Albin, Young, & Penney, 1989). The direct pathway encompasses D1-receptor expressing MSNs of the striatum, which project directly to the primary output nuclei of the basal ganglia: the GPi/SNr (primates) or the homologous EP/SNr (rodents). The indirect pathway arises from D2-receptor expressing MSNs, projecting to the GPe. The GPe, in turn, impinges upon the STN, which finally projects to the output nuclei. The STN is unique among basal ganglia nuclei for two reasons; firstly as the only excitatory structure within the group, and secondly for the diffuse nature of its efferent projections. Whereas the striatum, GPe, GPi and SNr retain the topographic organisation of cortical inputs to basal ganglia (Alexander et al., 1986), STN contacts its own targets in a more distributed manner (Parent & Hazrati, 1995b).

The SNr and GPi are also GABAergic, but in contrast to striatum, are tonically active, thereby imposing constant inhibition upon their own target regions of thalamus and brainstem (Mink, 1996). As a result, the inhibitory striatum, when activated, transiently inhibits the output nuclei via the direct pathway, resulting in a net disinhibitory effect on thalamocortical targets (Chevalier & Deniau, 1990; DeLong, 1990), from which regions the excitation of striatum originated. In contrast, the indirect pathway amounts to a net inhibitory effect on targets of basal ganglia output, via the disinhibition of STN by GPe, and consequent excitation of SNr/GPi. Importantly, given the distributed nature of STN projections, this excitation is widespread and diffuse, in contrast to the focused disinhibition arising from the direct pathway, resulting in an overall 'off-centre on-surround' effect on its targets, also described as an 'accelerator-brake' model (Graybiel, 2000).

This model, while popular, fails to take into account certain significant projections within basal ganglia, most notably, that STN is also an important input nucleus of basal ganglia (Parent & Hazrati, 1995b). Updated models incorporated this projection, dubbing the resulting route to output nuclei as the 'hyper-direct' pathway, which has been suggested to serve an important generalised inhibitory function (Nambu, Tokuno, & Takada, 2002). Even this updated model, however, fails to incorporate the now well documented reciprocal projection from STN to GPe, resulting in a complex focused-inhibitory/diffuse-excitatory feedback loop between the two structures (Smith et al., 1998). Additionally, inhibitory topographical projections have been documented from GPe to basal ganglia output structures. An alternative view of basal ganglia function incorporates this additional intrinsic basal ganglia connectivity (Gurney, Prescott, & Redgrave, 2001a, 2001b; Mink & Thach, 1993), and critically accounts explicitly for the attenuating effects of dopamine on the activity of D2-receptor expressing MSNs in striatum (Gerfen et al., 1990). While effectively still positing an off-centre on-surround selection mechanism, this alternative view posits a more subtle action selection system than the simple direct-indirect pathway model.

Importantly, the effects of dopamine on D2-receptors in striatum result in a relatively re-

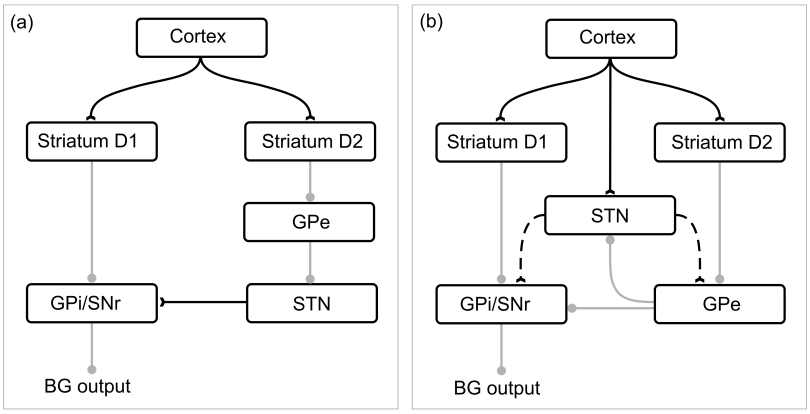

Figure 1.3: (a) Architecture emphasised by the original direct and indirect pathways model of selection in the basal ganglia (Albin, Young & Penney, 1989). Disinhibition of basal ganglia targets is achieved through activation of the direct pathway through striatum (D1) and GPi/SNr. Activation of the indirect pathway via GPe and STN produces inhibition of target nuclei, resulting in an 'off-centre on-surround' functionality. (b) Selection and control pathways model (Gurney, Prescott & Redgrave, 2001), emphasising the diffuse nature of STN projections and incorporating additional connectivity. Selection is performed via projections through striatum (D1) and GPi/SNr, whereas capacity scaling is achieved through STN and its connections with GPe and GPi/SNr. Excitatory projections are illustrated by black lines; inhibitory ones in grey. Dashed lines account for the emphasis of diffuse STN projections in the selection/control pathway model.

duced level of inhibition of MSN targets in GPe compared with SNr/GPi. This affects basal ganglia function in two ways. Firstly, via the inhibition of STN, the distributed excitation of output nuclei is attenuated to a degree which is relative to the overall amount of external excitation of basal ganglia. This 'scales' the output of the basal ganglia relative to the total input. Indeed, this function was termed 'capacity scaling' by the original authors (Gurney et al., 2001a), and allows the maintenance of an appropriate balance of excitation and inhibition of basal ganglia output nuclei, regardless of the total input. Further, via topographic projections from GPe to GPi/SNr, the attenuated inhibition of GPe acts synergistically on output nuclei with that from D1-receptor expressing regions of striatum, further enhancing the contrast of the off-centre, on-surround selection mechanism. This led to the identification of 'selection' and 'control' pathways through basal ganglia, reflecting this updated interpretation of basal ganglia functionality (Gurney et al., 2001a).

Modelling work

Compelling evidence for the role of basal ganglia in action selection has been the performance of a vast number of computational models based on the neuroanatomy of the basal ganglia and their connectivity with frontal cortex. While models of basal ganglia cover far more than action selection, also focusing heavily on action learning (Beiser, Hua, & Houk, 1997; Cohen & Frank, 2009; Gillies & Arbuthnott, 2000; Joel, Niv, & Ruppin, 2002), selection models are numerous and extensive (Berns & Sejnowski, 1998; Dominey & Arbib, 1992; Dominey, Arbib, & Joseph, 1995; Dominey, 1995; Frank, 2005; Fukai, 1999). Moreover, anatomically constrained models of basal ganglia have been shown to allow the selective gating of information into working memory (Beiser & Houk, 1998; Frank, Loughry, & O'Reilly, 2001; Frank, Scheres, & Sherman, 2007; Hazy, Frank, & O'Reilly, 2007; O'Reilly & Frank, 2006), indicating the likely generalisation of the selection function to multiple domains. Action selection has also been shown in models of *sub*cortical loops through basal ganglia (Houk et al., 2007).

1.3.3 The GPR model

Many of the above models include lateral inhibitory mechanisms in striatum, which effectively implement winner-take-all (WTA) selection mechanisms in basal ganglia; indeed, several of these models include only a 'direct' pathway, calling into question the relevance of the abundant connectivity through STN, GPe and D2-expressing regions of striatum for selection (e.g., Beiser & Houk, 1998; Dominey & Arbib, 1992; Dominey, Arbib & Joseph, 1995). While evidence exists for such inhibitory mechanisms in striatum (Gerfen & Wilson, 1996), the selection/control pathway view (Gurney et al., 2001a) suggests that the architecture of the basal ganglia as a whole should facilitate, if not dominate, action selection.

A rate-coded neural network model of the basal ganglia, dubbed the 'GPR' model, was developed which tested the selection capabilities of the selection/control pathway account (Gurney et al., 2001a, 2001b; Humphries & Gurney, 2002). The architecture of the GPR model is illustrated in figure 1.3(b). Following the fine-grained topography or micro-loop architecture of BGTC loops discussed above, the GPR is arranged in topographical 'channels', where each channel represents a distinct action; the model as a whole may be considered to reflect the macro BGTC loop occupying motor territories of the brain.

Each channel receives a distinct scalar input value, known as 'salience'. This value is considered to represent the strength of the current urgency for the corresponding action to be executed. Intrinsic competition between channels, mediated by the balance of activation in selection and control pathways, results in the selective disinhibition of a single channel in the thalamocortical loop. This, in turn, allows the thalamocortical loop to integrate activation beyond the level of its original salience, up to some selection threshold, at which point selection of that action is deemed to be achieved (Humphries & Gurney, 2002). Generally, selection of the action (channel) with the greatest input salience value occurs, as would be expected with a WTA model. However, the GPR was shown to have desirable selection properties over and above those of simple WTA mechanisms, including resistance to distractors, and greater contrast enhancement of selected and non-selected channels in basal ganglia output nuclei (Humphries & Gurney, 2002). Further developments of the GPR model have replicated experimentally observed oscillatory phenomena in a spiking version of the model (Humphries, Stewart, & Gurney, 2006), successfully implemented embodied action selection in a *khepera* robot model (Prescott, Montes González, Gurney, Humphries, & Redgrave, 2006), and enhanced performance in an existing model of the Stroop task (Stafford & Gurney, 2007).

Single actions

The GPR effectively examines the selection of single actions in isolation, and while it has been shown that the GPR model is capable of more complex sequential behaviours, this is only possible via the deliberate engineering of the salience values over time by the experimenter (Humphries & Gurney, 2002; Prescott et al., 2006). Given the success of the GPR model in the domain of action selection, it is of interest to apply the insights gained from its simulation to more complex behaviours which may be mediated more autonomously, rather than by this deliberate engineering of input signals. Within the context of the GPR model, understanding sequential action selection may then be re-conceived of as understanding the dynamics of the 'salience' signal, or, more generally, the sources and dynamics of influences on the processing of the motor BGTC loop.

1.3.4 BGTC loops as a single hierarchy

The well documented topography of the BGTC loop system has led to the widely accepted view that these loops, particularly on the 'macro' or territory scale, are almost entirely segregated, such that processing in one loop does not directly influence processing in its neighbour (Alexander et al., 1986; Alexander & Crutcher, 1990; Hoover & Strick, 1993; Middleton & Strick, 2000; Mink, 1996; Parent & Hazrati, 1995a; Selemon & Goldman-Rakic, 1985, 1991; Strick, Dum, & Picard, 1995). This view implies that motor territories are not subject to direct influence from processing in adjacent associative loops, calling into question the means by which cognitive processing may influence the selection of motor action. However, several strands of research are increasingly demonstrating a more complex relationship between loops, in which a degree of connectivity appears to allow integration in their processing. While this is not a new idea *per se* (Chevalier & Deniau, 1990; Joel & Weiner, 1994; Percheron & Filion, 1991), it has become refined in recent years with emphasis being given to understanding the precise role of interactions between loops and the means by which they communicate (Pennartz et al., 2009; Yin & Knowlton, 2006).

Corticocortical connectivity

It is well accepted, for example, that a high degree of direct communication exists between those cortical areas involved in BGTC loops in the limbic, associative and motor territories. For example, in primates, primary motor cortex (M1) is known to receive projections from supplementary motor regions, including SMA-proper and pre-SMA (Felleman & Van Essen, 1991; Geyer, Matelli, Luppino, & Zilles, 2000), and caudal premotor areas (Muakkassa & Strick, 1979; Takada et al., 2004). Evidence suggests this organisation holds in humans (Xiong, Parsons, Gao, & Fox, 1999). Premotor cortices themselves receive afferents from dorso- and ventro-lateral prefrontal cortex (dlPFC, vlPFC) (Barbas & Pandya, 1987; Takada et al., 2004) which, in turn, are influenced by orbital and medial prefrontal areas (Cavada, Compañy, Tejedor, Cruz-Rizzolo, & Reinoso-Suárez, 2000; Tanji & Hoshi, 2008).

It is of note that a divide appears to exist between rostral and caudal premotor cortices

(Geyer et al., 2000), including in humans (Matsumoto et al., 2007), with rostral sections receiving strong projections from PFC but not projecting significantly to motor regions, and caudal areas receiving little direct influence from PFC but reliably influencing more caudally located SMA and M1. However, connectivity between these regions implies that communication along this route is possible (Barbas & Pandya, 1987; Luppino, Rozzi, Calzavara, & Matelli, 2003; Takada et al., 2004). Indeed, it has been suggested that rostral premotor regions serve as a 'gateway' between cognitive and motor cortices (Hanakawa, 2011; Nakayama, Yamagata, Tanji, & Hoshi, 2008).

Despite this clear hierarchical pathway through cortices, this is rarely cited as a viable mechanism by which BGTC loops may communicate directly. This may simply be due to the focus on BGTC loops in processing - and more specifically, learning - within the basal ganglia themselves, rather than the peripheral cortical components of the system (Haruno & Kawato, 2006). Further, whereas layer V of cortex gives rise to the majority of corticostriatal and corticothalamic axons (Lévesque et al., 1996), corticocortical connections arise primarily from layer III (Haber & Calzavara, 2009; Muakkassa & Strick, 1979; Rockel, Hiorns, & Powell, 1980), indicating a possible distinction in the information being passed between cortical regions, and that being propagated to basal ganglia. The degree of this distinction is unclear, however, given the observation that thalamocortical projections conveying information from basal ganglia are received by multiple cortical layers, including layer III (McFarland & Haber, 2002). It thus seems reasonable to suppose that corticocortical projections allow at least some integration of information between loops that has hitherto largely been neglected.

Thalamocortical loop

Reciprocal and non-reciprocal bidirectional connections between thalamus and cortex have also been shown in the frontal lobes, consisting of corticothalamic 'hotspots', or regions of thalamus receiving focused projections from multiple regions of cortex, and divergent projections from a single thalamic region across multiple cortical regions (Haber & Mcfarland, 2001; Haber, 2003; Haber & Calzavara, 2009; McFarland & Haber, 2002). For instance, the ventral anterior (VA) nucleus of thalamus, its primary associative region, receives a non-reciprocal input from medial PFC, which has been more strongly associated with limbic function than the typically 'assocative' lateral PFC, with which VA is reciprocally connected. Likewise, the ventral lateral (VL) nucleus of thalamus, its motor territory, receives a non-reciprocal influence from rostral premotor regions, associated with associative function, in addition to its reciprocal connections with caudal premotor and motor regions (Haber & Calzavara, 2009). These sites may act as important integrative points for information processed by distinct BGTC loops. Again, these connections appear to propagate information in a rostro-caudal direction, consistent with the idea of a hierarchical organisation of the BGTC system.

Corticostriatal projections

Evidence also suggests the existence of diffuse corticostriatal projections across distinct BGTC loops and sub-loops, which may be facilitated by the vast axonal arborization of corticostriatal neurons and the resulting high level of convergence of cortical neurons on single MSNs (Zheng & Wilson, 2002). In primates, for example, corticostriatal projections from prefrontal area 8 in monkeys have been shown to impinge on almost the entire length of the caudate nucleus (Goldman & Nauta, 1977) which also receives significant projections from premotor cortices (Calzavara, Mailly, & Haber, 2007). More specifically, evidence for overlapping corticostriatal projections in primates has been found not only for multiple motor cortices such as SMA and M1 (Inase, Sakai, & Tanji, 1996; Nambu, Kaneda, Tokuno, & Takada, 2002), and premotor cortices and SMA/Pre-SMA (Tachibana, Nambu, Hatanaka, Miyachi, & Takada, 2004; Takada, Tokuno, Nambu, & Inase, 1998), but more significantly from apparently distinct territories including associative and premotor (Calzavara et al., 2007), and limbic and associative (Haber, Kim, Mailly, & Calzavara, 2006). Evidence from distinct tracing studies further suggests that projections from premotor cortices may terminate in the same regions of caudate as those from prefrontal areas, particularly the striatal cell bridges (Calzavara et al., 2007; Tachibana et al., 2004). Evidence for similar pattern of corticostriatal convergence has also been shown in humans using a novel analysis of magnetic resonance imaging data (Draganski et al., 2008), and also in rats (Reep, Cheatwood,

& Corwin, 2003), where overlapping projections from associative/premotor cortex and regions analogous to the primate orbitofrontal cortex (Dalley, Cardinal, & Robbins, 2004) were observed. This corticostriatal overlap again suggests a particularly integrative role of premotor regions, with an ‘open associative loop’, originating in PFC and terminating in premotor regions which in turn send efferents to motor striatum via a ‘closed motor loop’ (Joel & Weiner, 1994).

Striatonigral and striatopallidal projections

Both striatonigral and striatopallidal projections may be important in integrating information between loops. SNr, for example, has been cited as a potentially important source of integration within basal ganglia, with overlapping projections from all regions of striatum (Joel & Weiner, 1994), allowing limbic and motor information to reach associative PFC - the cortical target of SNr output via thalamus - whereas medial GPi may mediate the outflow of infomation from associative to motor regions via premotor cortex (Joel & Weiner, 1994). Finally, several authors also describe striato-nigral-striatal ‘spirals’ which form a hierarchical pathway for the flow of information from limbic, through associative, to motor regions of basal ganglia via dopamine releasing areas of the midbrain (ventral tegmental area (VTA) and substantia nigra pars compacta (SNc)) in both rats and primates (Haber, Fudge, & McFarland, 2000; Haber, 2003; Ikemoto, 2007; Joel & Weiner, 2000), which is now well accepted and inspiring contemporary theoretical perspectives on the function of BGTC loops (Pennartz et al., 2009; Yin & Knowlton, 2006).

STN connectivity

Efferent projections of STN have long been accepted to be diffuse, rather than focal in nature (Mink & Thach, 1993; Parent & Hazrati, 1995b), but its role in integrating information between loops is unclear. It is possible that STN is important for integrating over neighbouring micro loops, rather than loops on a larger scale, as would be suggested by the selection/control pathway view of basal ganglia function. However, some evidence does suggest explicit influences across macro-loops via STN, in both afferent and efferent connections (Joel & Weiner, 1997). Specifically, that associative regions of GPe influence motor regions of STN, and associative STN projects to motor *and limbic* regions of GPi/SNr, creating two 'open indirect pathways'. This suggests two mechanisms by which associative information is propagated across loops at the level of STN (Joel & Weiner, 1997). Interestingly, this contrasts with other mechanisms which seem to be consistent across levels of the hierarchy, mediating a limbic → associative → motor flow of information. The propagation of associative information alone here may highlight the importance of associative information in co-ordinating the activity of both higher and lower loops.

Summary

Despite popular opinion to the contrary, a high degree of connectivity appears to exist between BGTC loops, both at the macro- and micro-scale. While evidence suggests connectivity between loops allows the flow of information in both directions to some extent, the majority indicates a flow of information from limbic, through associative, to motor territories, forming a hierarchy of BGTC loops. While this connectivity is far from exhaustive, and the 'closed-loop' scheme traditionally described generally holds, it seems clear that sufficient integration is possible such that motor output may be directly influenced by processing in associative and limbic regions *within* the overall BGTC system.

1.3.5 BGTC loops in sequential action

In addition to neuroanatomical evidence suggesting the role of higher BGTC loops in the modulation of actions selection processes in lower loops, a significant degree of functional evidence also points to the involvement of multiple regions of the BGTC loop hierarchy in sequential action, particularly the associative loop, encompassing PFC. Many excellent reviews exist documenting the role of PFC and related functional areas in cognitive processing, often with a focus on sequencing (Courtney, 2004; Curtis & D'Esposito, 2003; D'Esposito, 2007; Fuster, 2001; E. Miller, 2000; E. Miller & Cohen, 2001; Tanji & Hoshi, 2008); therefore what follows is merely a brief summary of the most relevant and reliable findings.

While functionality in PFC is localisable to some extent, when taken together, prefrontal regions are sensitive to and appear to maintain and manipulate contextual information required for the production of action sequences, where, after Lashley (1951) we consider 'context' to consist of any information required for the correct execution of the present action, and correct transition to the next. A common feature of sequential tasks, for example, is the requirement for working memory, given the need to keep in mind relevant plans, task rules, current progress, and other contextual information for the duration of the sequence in order to guide action, often over the course of some delay or distracting event. Since the early 1970s (Fuster & Alexander, 1971), it has become well established that PFC neurons show sustained activation in the absence of ongoing sensory stimuli, and this is frequently purported to reflect working memory processes (Curtis & D'Esposito, 2003). Importantly, sustained activation in PFC reflects task-relevant information; distracting or irrelevant information is rarely maintained in this manner. Indeed, an influential theory suggests it is precisely this maintained activity which encodes those rules, plans and strategies for the guidance of voluntary sequential behaviour (E. Miller & Cohen, 2001). There is a significant degree of evidence to suggest that this high level information is maintained in PFC, with studies showing PFC activity related to ultimate goals (Tanji & Hoshi, 2008), action plans modulated by task requirements (Hoshi, Shima, & Tanji, 1998), sequence level action representations (Ostlund, Winterbauer, & Balleine, 2009), as well as the encoding of current sequence identity (Averbeck, Crowe, Chafee, & Georgopoulos, 2003; Averbeck, Sohn, & Lee, 2006) and current sequence category (Shima, Isoda, Mushiake, & Tanji, 2007). Moreover, individual PFC neurons involved in active maintenance have been shown to be selective for, or modulated by, stimulus features which are relevant for guiding action, such as object identity, shape, colour and location, as well as more abstract information such as expected reward (see Miller & Cohen, 2001 and Tanji & Hoshi, 2008 for reviews).

Perhaps most relevant is the repeated observation of activity related to the temporal sequencing of movements; for instance, monkey oculomotor tasks have shown PFC neurons to be selective for the current stage of a task, effectively 'tracking' progress of the sequence (Hasegawa, Blitz, & Goldberg, 2004). Similarly, other studies showed neurons selective for the order of presentation of visual stimuli (Funahashi, Inoue, & Kubota, 1993, 1997). Another showed sets of neurons encoding complementary information for sequential action, including temporal relationships of presented stimuli *and* progress tracking, highlighting the involvement of PFC in both serial encoding and action production (Barone & Joseph, 1989). Furthermore, simple sequential tasks are known to be affected by frontal lobe damage in humans (Petrides & Milner, 1982). In addition to PFC, associative regions of basal ganglia have been implicated in planning for sequencing (Houk & Wise, 1995), and, like PFC, basal ganglia neurons show sensitivity to task context and are critical for working memory (Hikosaka, Takikawa, & Kawagoe, 2000; Tanji, 2001).

It is important to note that other cortical areas in the frontal lobes have been heavily implicated in the control of sequential behaviour, most notably, the SMA and pre-SMA, just rostral to primary motor areas (Tanji, 2001). Such regions, along with premotor cortices (Hanakawa, 2011; Nakayama et al., 2008) may be components of distinct, intermediate BGTC loops that deal with neither purely associative nor motor information, but some combination of both. In the interests of parsimony, and in order to retain a focus on a clear associative and motor distinction, we do not explicitly address the role of SMA or pre-SMA in this thesis; rather, we effectively subsume all 'associative' processing into PFC and related subcortical areas, and all motor processing into a single motor BGTC loop. This is a highly simplified scheme however, and we acknowledge that SMA and pre-SMA are likely to be extremely important in the control of sequential action, possibly mediating important communication between more strictly associative and motor regions.

PFC in routine tasks

While it is almost unanimously accepted that PFC and related territories of basal ganglia are involved in novel sequence processing, the involvement of the area in routine tasks is less so, given the interchangeability of 'routine' and 'habit' in much of the literature. Historically, a habit has been defined as any behaviour that is evoked merely by the presence of a particular stimulus, and in the absence of a related goal state (Dickinson, 1985). Previous work on well learned or routine tasks has consistently referred to such performance as 'habit' (Reason, 1984), 'automatic' (Schwartz et al., 1995), and that 'conscious attentional control is not necessary' (Norman & Shallice, 2000). In a particularly strong emphasis of this point, Botvinick & Plaut (2004) had no requirement for a corresponding goal for carrying out the tea- and coffee-making tasks in their SRN model, implying the interchangeability of 'routine' and 'habit'. Given evidence that associative regions of the BGTC hierarchy play little, if any, role in habitual behaviour (Graybiel, 2008; Redgrave et al., 2010), this implies that routine action proceeds without any significant contribution of these regions.

It seems likely, however, that routine behaviour is not truly habitual in this manner. Firstly, the mere sight of a kettle rarely incites one to begin making a cup of tea, as is suggested by the sensori-motor response definition of habit. Indeed, even where routine action is discussed as automatised, usually a corresponding goal is assumed (Aarts & Dijksterhuis, 2000; Cooper & Shallice, 2000; Reason, 1979; Wood & Neal, 2007). Secondly, such an extended, complex and flexible sequence would exceptionally rarely be performed to completion without a concurrent and corresponding goal state. Indeed, Fuster (2001) writes that,

> 'even after repetition and automation of their performance, sequences retain a degree of representation in lateral PFC. Whereas the automatic aspects of motor behaviour may have been relegated to lower structures, the more abstract and schematic representations of sequential action, as well as the general rules and contingencies of motor tasks, appear to remain represented in prefrontal networks'. (p322)

Other authors also explicitly suggest that cognitive structures required for routine sequential action rely on prefrontal regions (Zalla, Pradat-Diehl, & Sirigu, 2003), and that the contextual information required for such action, possibly within schemas, is represented prefrontally (Badre, 2008; Tanji & Hoshi, 2008) as long term knowledge or stable representations (Courtney, 2004). Additionally, the patterns of errors observed in frontal lobe patients while performing routine tasks (Humphreys & Forde, 1998; Schwartz et al., 1998) imply a role for these regions in such tasks, while monkey models of routine action have directly pointed to the involvement of PFC (Ryou & Wilson, 2004). This is further supported by evidence that action 'slips' in such sequential performance are more common when the performer is distracted (Reason, 1979, 1984), suggesting that some higher cognitive resource is required for their consistent and accurate performance. It seems likely, then, that routine tasks like ADLs are sufficiently complex that performance in their entirety is beyond the scope of habitual mechanisms, however well learned they may be. For these reasons, we do not regard these behaviours as truly habitual. This accounts for the likelihood that schema or schema-like representations exist to support their performance in a general setting, for which suggestion evidence exists (Fuster, 2001; Humphreys & Forde, 1998), without requiring strong assumptions about the redundancy of goals or nature of external stimuli as inevitable triggering conditions.

Together, this evidence suggests that higher regions, particularly associative territories, of the BGTC hierarchy are important for the direction of routine sequences. The observed interconnectedness within this hierarchical system may be critical for the robust and flexible learning and execution of goal-directed action sequences, in a manner which may be more efficient than the preservation of a strict segregation of loops. By allowing the direct biasing of processing in one loop by another, this scheme avoids having to recruit alternative structures in order to integrate various types of information, and is thus likely to be faster and less computationally expensive than a segregated architecture.

Existing models implementing multiple BGTC loops

Several computational models have utilised a multiple loop BGTC architecture - some explicitly implementing cross-territory connectivity - in order to model complex functionality within BGTC loops that would not be possible without communication or direct arbitration between distinct loops. For example, certain models implementing reinforcement learning of actions in the basal ganglia, have embodied different learning processes within this paradigm known as the ‘actor’ and the ‘critic’ (see Sutton & Barto, 1998) in more dorsally and ventrally located BGTC loops, respectively (Khamassi, Girard, Berthoz, & Guillot, 2004), with the more heavily neurally constrained models specifying the particular cross-territory connectivity responsible for their interaction, such as corticostriatal (Şengör, Karabacak, & Steinmetz, 2008). In these models, dorsal regions are responsible for the selection of actions, while higher ventral regions of the hierarchy learn to modulate this selection according to particular task demands. This loop-wise scheme represents a reinterpretation of more traditional accounts which posit both processes in distinct functional compartments of striatum, but within the same macro-loop (Joel et al., 2002).

Girard et al. (2005) propose a two-loop model based on the GPR which, interestingly, uses upstream interloop connectivity from dorsal (motor) STN to SNr of a ventral loop (Joel & Weiner, 1997). The ventral loop is concerned with 'appetitive' (searching) actions; the dorsal is equated strongly with the original GPR and mediates 'consummatory' actions. The model integrates navigational actions selected by the ventral loop with non-locomotor actions in the dorsal loop. Selection of a consummatory action in the dorsal loop inhibits the selection of a further navigational strategy, allowing the agent to exploit its current environment. Otherwise, the ventral loop selects one of three navigational strategies to explore its environment. This upstream inhibition mechanism, deployed when the lower loop is engaged, prevents interference by higher regions of the hierarchy. Haruno and Kawato (Haruno & Kawato, 2006; Kawato & Samejima, 2007) present a 'heterarchical' learning model of instrumental learning, utilising striatonigral 'spirals' following Haber and colleagues (Haber et al., 2000; Haber, 2003), as connectivity between prefrontal and motor loops. Again modelling reinforcement learning, errors in reward prediction are represented in a coarse manner in the PFC loop, which guides learning of finer grained predictions in the motor loop. The authors suggest that the information provided to the motor loop by the PFC loop may be representative of subgoals, teaching signals or a coarse approximation of the same prediction, suggesting that the model is open to interpretation to some extent. fMRI experiments, however, supported several of the model's predictions. Nakahara et al., (2001) present a model of 'visual' and 'motor' BGTC loops, located more ventrally and more dorsally. Each loop learns a visuomotor sequence concurrently, but using visual and motor co-ordinates, respectively. Communication between the two loops is mediated by the pre-SMA, which acts as a 'co-ordinator', resolving inconsistencies between the output of the two loops via its projections to striatum of both loops. Different speeds of learning result from the different co-ordinates used for selection, with the result that selection is primarily guided by the PFC-loop early in learning and by the motor loop later in learning, successfully modelling the transition from cognitively guided to automatic performance. A further, particularly influential model also posits competition and arbitration between distinct behavioural 'controllers', representing goal-directed and habitual learning and control in associative and motor territories of the BGTC system (Daw et al., 2005). While this model does not specifically address the role of direct communication between distinct loops, an interconnected loop scheme is an appealing one for understanding the neural implementation of the two controllers. Finally, Chambers and colleagues (Chambers & Gurney, 2008) utilise cross projections between associative and motor regions in order to implement rule-based 'inaction' selection which, critically, may be *un*learned according to changing task demands.

1.4 Summary and proposal for programme of research

From the above discussion, we can conclude that all 'sequential behaviour' cannot be accounted for by a single process, and evidence suggests that different types of sequential behaviour may rely on distinct structural and functional mechanisms for their successful execution. Within the domain of routine action, existing theories frequently centre around cognitively focused explanations, often concerning the dynamics of hierarchically organised action schemas. Such approaches have been utilised to explain distinctive patterns of human error, both in healthy participants and sufferers of frontal lobe injury. Though two particular proposals have been either directly or indirectly investigated in computational models, modelling work is yet to incorporate neuroanatomical constraints, thus the applicability of the conclusions of this research is limited.

Despite this, a great degree of neuroanatomical, behavioural and physiological evidence points to the importance of the BGTC hierarchy in mediating control of routine sequences, particularly in light of increasing evidence that distinct territories of this system are interconnected in a hierarchical fashion. The GPR model of sophisticated action selection and its descendents, based on the architecture of this system, are well specified and produce impressive results, but tend to focus on the selection of single actions, and have rarely been applied in the associative domain. Multiple BGTC loop models incorporating action selection are also receiving attention. However, while each of these models provides a unique and important contribution to understanding the utility of multiple, specialised processes for the control of action, to our knowledge none has yet been applied to understanding the dynamics of abstract cognitive structures for routine action frequently discussed in the psychological literature. In this thesis, we attempt to unite these two approaches by proposing a computational model of routine sequence production based on the interconnected BGTC loop architecture. More specifically, we exploit the functionality of the existing GPR model of action selection in a model of associative and motor BGTC loops, in a neuroanatomical interpretation of the theoretically powerful results from earlier models of routine actions.

1.4.1 Research questions

In the forthcoming work, there are a number of specific issues arising from the above discussion that we wish to consider:

1. How are cognitive structures such as schemas, containing representations of context, organised within the BGTC loop architecture, and how does the architecture facilitate their dynamics and maintenance in working memory?

2. How are these representations of context translated into action, and how does the interconnected BGTC architecture support this process?

3. What are the required mechanisms for the sequencing of cognitive structures, and how might they be plausibly implemented in the BGTC loop architecture?

4. What information must be represented by schemas or their equivalent in order to allow 2 and 3?

5. How might the structure, organisation and dynamics of these cognitive structures account for observed patterns of errors in healthy participants and sufferers of ADS?

In the following chapter we begin by addressing question 1, by proposing a novel interpretation of the GPR model as the associative BGTC loop. Chapters 3 and 4 expand upon this, incorporating motor territories of the BGTC system in order to develop a model for the performance of routine action sequences, focusing specifically on questions 2-4. Finally, chapter 5 introduces disruption to the model in order to examine our last question regarding error behaviour in normal and impaired populations.

Chapter 2

Selection and maintenance of distributed representations

2.1 Introduction and aims

In the general introduction, we discussed evidence indicating the important role of associative regions of the BGTC hierarchy in processing contextual information for routine sequential tasks. Here, we begin our experimental investigation by addressing the question of how the associative BGTC loop architecture facilitates the selection and maintenance of cognitive representations of this information. After the discussion in the previous chapter, such representations may reasonably be considered schemas; however, given the neuroscientific emphasis in this chapter we refer to these cognitive structures as 'representations' to allow clearer comparison with previous work. In addressing this question, we consider insights from existing modelling work examining associative regions of cortex and basal ganglia, in addition to specific computational constraints. We also draw on psychological theory and neurophysiological evidence to develop a model of the associative BGTC loop for the processing of cognitive representations. We use as a basis for our modelling work the GPR model of action selection in the basal ganglia (Gurney et al., 2001a, 2001b), or more specifically, a later development which included the addition of a thalamocortical loop (Humphries & Gurney, 2002). For clarity, we refer to this extended model as the *TC-GPR*.

2.2 A structured approach

In order to address selection of cognitive representations for routine tasks in the general case, we refrain from considering their necessary semantic content for any particular task. This allows us to examine the necessary principles of the underlying architecture to mediate selection of such representations for any task, rather than being constrained by the demands of a single task. To structure the ensuing investigation, a set of functional criteria was devised in order to help guide the development of a model of the associative BGTC loop for routine sequences. These criteria were designed to reflect the computational requirements of a flexible, context sensitive, sequence production mechanism, particularly where that mechanism may plausibly be implemented in the neural hardware.

1. **Form, or 'select' a unique representation of the current temporal task context.**
   As outlined in the introduction, we consider the temporal 'context' of a task to consist of the information necessary in order to drive correct selection of the current action, and a correct transition to the next, at any one time. This criterion reflects the suggestion that PFC is responsible for maintaining a representation of features of the current task context such as goals and task rules (E. Miller & Cohen, 2001; Tanji & Hoshi, 2008); for routine tasks, possibly in the form of some schema (Badre, 2008; Courtney, 2004; Fuster, 2001). It is important to note that, as the current context changes as the task progresses, this representation of context must be dynamic. In order for this scheme to be effective in influencing the appropriate actions and transitions between them, it is required that each contextual representation is unique, such that influences on their targets, including appropriate action representations lower in the BGTC hierarchy, are unambiguous.

2. **Represent this information in an efficient manner.**
   This follows from the previous criterion and refers to the computational impracticality and implausibility of a 'localist' coding scheme, whereby all possible task contexts are represented with entirely orthogonal *coding elements*, such as neurons or unique groups of neurons (Averbeck et al., 2002; Botvinick & Plaut, 2004; Rolls & Treves, 1990). As discussed in more detail below (section 2.3.1), it is likely that complex,

multi-dimensional information, which is likely to comprise schema-like representations of context (e.g., Shallice, 1982), is represented in a more efficient fashion, with the re-use of individual coding elements across multiple representations.

3. **Autonomously maintain selected representations in the absence of the external stimuli that initiated their selection.**
   Working memory, as discussed earlier, is a well established function of PFC, and sustained activity of ensembles of PFC neurons encoding stimulus and task features is hypothesised to underlie this functionality (Curtis & D'Esposito, 2003; E. Miller & Cohen, 2001). It is thus important that any plausible model of the associative BGTC loop is able to replicate this function, by the autonomous maintenance of activity in any selected representation of context.

2.3 Implementation of criteria

In implementing the above functional requirements in a biologically plausible model, there are several constraints from neuroanatomy and neurophysiology which must be taken into account. Consistent with functional criterion 2, neurophysiological recordings from prefrontal cortical neurons strongly suggest that representations of complex stimuli and sequential task progress are distributed in nature (Tanji & Hoshi, 2008), and recent modelling work has begun to examine the processes governing the emergence of PFC representations (Reynolds & O'Reilly, 2009; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). Beyond this, however, to the best of our knowledge no analysis of the cortex, its architectural relationship with other neural structures, or the nature of the computations they perform, has yet provided any great insight into the precise structure of the information contained within these representations, or whether there are particular limitations on its organisation (O'Reilly, Herd, & Pauli, 2010). Here, we consider neuroanatomical evidence and computational arguments leading to the suggestion that these representations take a particular structural form.

2.3.1 Efficiency & GPR functionality

Functional criterion 2 requires the efficient representation of unique representations of context in the associative BGTC loop. Its implementation demands an understanding of the functional architecture of the system. In the *motor* territories of BGTC loops and models grounded therein, the proposed relationship between cortex and basal ganglia is a relatively well understood one. To a great extent, there is evidence to suggest that motor cortices are arranged topographically, probably according to action specifications within particular limbs (Georgopoulos, Kalaska, Caminiti, & Massey, 1982; Graziano, Aflalo, & Cooke, 2005). The topographical anatomical relationship between cortex and basal ganglia, while hugely convergent, implies that this organisation is maintained throughout basal ganglia (Parent & Hazrati, 1995a). This in turn suggests a rough preservation of a topographical projection pattern throughout the entire loop. While, within this topography, representations are likely to take a complex form, at the systems level, the organisation effectively results in a localist coding scheme throughout the motor loop; this is reflected in the channel-wise organisation of the GPR model of selection in basal ganglia (Gurney et al., 2001a, 2001b).

In contrast, the significant degree of evidence that suggests PFC displays a more distributed coding scheme is consistent with distributed representation theory (Hinton, McClelland, & Rumelhart, 1986), which emphasises the efficiency problems associated with encoding each slight variation of a particular feature with a unique neuron or orthogonal group of neurons. This problem is usually discussed with reference to the representation of objects in the inferotemporal cortex and colloquially referred to as the 'grandmother cell' problem (Gross, 2002), though its principles extend to any region of the brain encoding complex stimuli or abstract concepts. The precise degree to which representations in PFC are distributed, however, are unclear. An increasing amount of evidence has now been gathered to indicate at least a certain degree of sparseness (Bowers, 2009; Gross, 2002; Quiroga, Reddy, Kreiman, Koch, & Fried, 2005), suggesting a semi-localist or feature-based scheme may be common. If, then, we accept the evidence that PFC representations are unlikely to be structured in a truly localist fashion, it is unlikely that this strong 'one-to-one' relationship with basal ganglia is maintained in the associative loop.

While a distributed scheme in cortex is plausible, it is functionally incompatible with the well-supported notion of basal ganglia as a topographic, competitive selection mechanism. Imagine, for example, a series of distributed representations which may be individually activated in PFC. If basal ganglia is to 'select' any particular representation at the level of cortex, it must be able to distinguish competing representations at the level of its own organisation, the 'channel'. The competitive selection mechanism hypothesised in Redgrave et al. (1999), and formalised in Gurney et al. (2001a, 2001b) therefore dictates that an individual channel in basal ganglia would be required for each possible distributed representation, in order to emulate the closed 'micro-loop' architecture on which the mechanism relies. Essentially, the mechanism by which basal ganglia performs selection effectively requires that it adopts a localist code, whereas evidence suggests that PFC employs at least a partially distributed one. There are, however, two computational problems with this scenario:

1. Assuming recurrence in PFC, as commonly implemented in computational models utilising distributed representations in PFC (Beiser & Houk, 1998; Dominey et al., 1995; Rougier et al., 2005), any overlap between different representations - resulting from any individual neuron's involvement in multiple representations - may cause a spread of activation through the cortex, degrading the quality of the 'selected' representation. Conversely, total orthogonality of representations effectively results in a localist code and thus fails our efficiency criterion.

2. A localist replication in basal ganglia of each distributed cortical representation effectively negates the computational benefits to efficiency afforded by a distributed scheme in cortex; this raises the question of what the distributed scheme achieves.

As an alternative to a fully distributed code, a feature-based scheme throughout the BGTC loop is also potentially problematic. In this case, it is again reasonable to suggest that each distributed representation consists of a set of individually localist feature representations. Here we might imagine that a separate channel in basal ganglia is dedicated to encoding each possible feature. This would require basal ganglia to concurrently select all relevant features related to the current task context. However, this 'multiple selection' is precisely one of the issues that the very architecture of the basal ganglia as a selection mechanism is hypothesised to *prevent* in order to avoid problematic behavioural manifestations of it, such as 'dithering' (Gurney et al., 2001b; Humphries & Gurney, 2002); such a scheme would thus be unable to encode incompatibility between competing features, and unable to implement the clean selection properties displayed by the GPR model and its descendents.

2.3.2 Mechanisms supporting sustained activity

As mentioned above, randomly connected recurrent networks have been implemented as PFC in previous models of basal ganglia loops (Beiser & Houk, 1998; Dominey et al., 1995; Rougier et al., 2005), in order to form unique, distributed representations of task context. Connectivity in such recurrent networks may be hard-coded or may be modified by learning algorithms to store and retrieve particular representations. The 'storage' of a particular representation indicates the ability of the network to self-sustain activation of that representation in the absence of external input. If network weights are modified according to the Hopfield prescription (Hopfield, 1982), recurrent networks are generally accepted to store around 0.15N stable patterns, where N is the number of elements comprising the network, though recently, Hopfield nets have been shown to have a surprisingly large storage capacity of up to N (Wu, Hu, Wu, Zhou, & Du, 2012). In the present model, however, there are particular constraints on the manner of the connectivity between PFC, basal ganglia and thalamus, and on the intrinsic connectivity of basal ganglia. This additional structure may impact on the storage capacity of the system. In developing the model then, it is important to take into account the nature of the connectivity between regions of the BGTC loop, and implement a means of sustaining activation of selected representations without compromising the potential storage capacity of the system.

2.3.3 Role of PFC in selection

Given the wealth of evidence accrued on basal ganglia function, it seems clear that the micro-loop structure of the BGTC system facilitates the selective disinhibition of the representation with the greatest input 'salience', whether this be at the level of actions or cognitions. However, this raises the question of whether selection must be *wholly* mediated by BG, or whether cortex itself possesses selection mechanisms to expedite or otherwise enhance this process, as a great deal of previous research would suggest (e.g., Rowe et al., 2000). As such, the more complex processes that are likely to underlie the selection of non-localist representations may rely on complementary selection mechanisms within basal ganglia and PFC itself.

2.3.4 Deselection

It is also important to address the important issue of 'deselection', or the ability of the loop to deactivate a currently maintained representation in the presence of increased support for a competitor. This must be balanced with the need for sustained activation. Put another way, the mechanisms allowing sustained activity should be sufficiently flexible that an intrinsically maintained representation should be easily deactivated in the presence of sufficient activation of a competing representation. Such a mechanism seems trivial, but its exclusion from previous work (Beiser & Houk, 1998; Frank et al., 2001) highlights the need to address its mediation in an explicit manner. This deactivation may rely on several mechanisms, for example, recurrent inhibition within PFC itself. However, the basal ganglia functionality illustrated by the GPR model should be preserved such that the inhibition from basal ganglia output nuclei is a component of this process.

2.3.5 Information preservation

Finally, we consider the issue of information preservation, and whether the information contained within the cortical representation of task context must be entirely preserved throughout the entire associative loop, or whether it is sufficient for basal ganglia to utilise only key features of this representation in order to perform the task. The degree of converging projections from cortex (see Mink, 1996) would suggest that some kind of compression or dimensionality reduction is likely, and some evidence for this idea also comes from modelling work (Bar-Gad, Morris, & Bergman, 2003).

2.4 Novel architectural solution

2.4.1 Proposition for a new architecture

Clearly, in its current guise the TC-GPR model is unable to meet the functional demands of a model of the associative loop. As such, we propose a modified version of the model based on a novel architectural concept, exploiting the utility of a recurrent network as PFC in order to implement efficiency and storage of stable, feature-based representations in concert with basal ganglia.

We suggest that basal ganglia may perform selection in a coarse manner at this level of the BGTC hierarchy. By this, we mean that each channel in basal ganglia may select a *subset* of the total representations available to PFC. Recurrent connectivity within cortex itself then affords a complementary selection role, disambiguating those representations forming the selected subset. A combination of intrinsic PFC recurrence and excitatory feedback between cortex and disinhibited thalamic channels allows maintenance of the selected representation until sufficient competition is encountered that drives inhibition of thalamus to deactivate the representation. While this scheme results in a partial loss of information in basal ganglia, the architecture of the loop as a whole is such that stability of representations in cortex - and subsequently the richer information therein - is preserved.

This embodies a ‘divide and conquer’ approach to the problems of selecting, maintaining and deactivating non-localist cortical representations utilising a localist architecture in basal ganglia, whereby basal ganglia and cortex both perform crucial and complementary roles in the selection and maintenance processes. We further propose a partially distributed or feature-based coding scheme to be utilised in cortex; this affords a more predictable system of overlap between representations, whereby those representations sharing a particular feature will share the same neural substrate to represent that feature, thus making this a particularly pragmatic implementational choice for modelling. Beyond this, however, it is likely that this semi-localist structure is in fact more plausible than a fully distributed scheme (Bowers, 2009; Gross, 2002; Quiroga et al., 2005).

2.5 General model description

As discussed above, we base our model on the existing basal ganglia thalamocortical loop model introduced by Humphries & Gurney (2002), itself a development of the channel-wise GPR model of the basal ganglia (Gurney et al., 2001a, 2001b). While we are interested in the interplay between loops occupying different functional territories of the BGTC system for mediating action sequences, in this chapter we focus on modelling the associative loop in isolation in order to directly address the issues laid out in sections 2.1 and 2.3. Nor do we address the mechanisms underlying the appropriate *sequencing* of these representations here; rather we wish to fully investigate the functional capabilities of the novel architecture we propose, in terms of selecting and maintaining individual cognitive representations, such as may be required for sequential action selection in an extended model.

2.5.1 Basic model architecture

The organisation of the modelled basal ganglia and thalamus is illustrated in figure 2.1(a). This figure emphasises the typical channel-wise or one-to-one connectivity between the majority of the nuclei modelled therein, as discussed in detail in Gurney et al. (2001a, 2001b) and Humphries & Gurney (2002). For the reasons explained above, we wish to avoid a continuation of this localist representational scheme in cortex. This in turn excludes the preservation of a channel-wise organisation of this region of our model. Rather, we now implement PFC as a recurrent network with a greater number of elements than the existing modelled nuclei (figure 2.1(b)). In contrast to previous GPR models, we now consider the selection of a pattern of activity across several elements in PFC, rather than the selection of a single element or channel. Consequently, we are now able to efficiently implement contextual representations as conjunctions of *features*. Critically, each feature may be a component of multiple representations, resulting in overlap between representations sharing particular features, with clear conceptual benefits regarding the similarity of representations which encode similar contexts, such as generalisation (Hinton et al., 1986) and graceful degradation (Rolls & Treves, 1990).

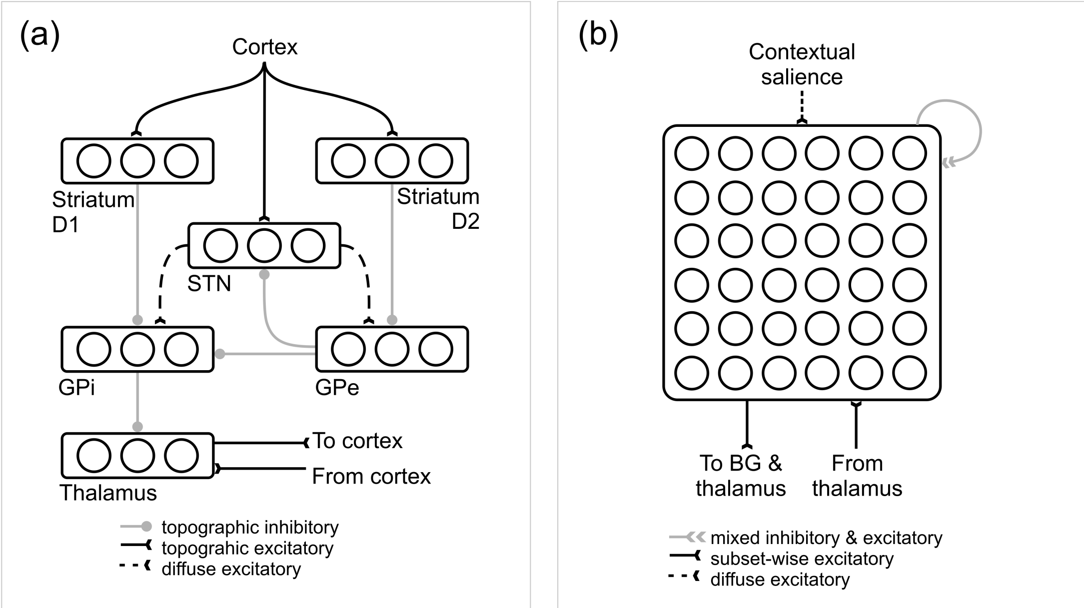

Figure 2.1: (a) Schematic diagram of the TC-GPR architecture, emphasising its topographic organisation. We propose an expansion of cortical dimensions to allow distributed representations. (b) Proposed organisation of the present model. Multiple representations project to and receive afferents from each channel in basal ganglia (see text for detailed discussion).

2.5.2 Content and structure of representations

According to our discussion above, the temporal task context should be represented by a *unique* pattern of activity in associative regions of cortex such as PFC, such that unambiguous information may be communicated to those brain regions involved in effecting motor action for contextually appropriate performance. As mentioned in the introduction, we consider the temporal context to consist of any information which is required to ensure the correct selection of actions and their appropriate temporal ordering. Thus, it is likely that the content of representations of context will vary between tasks. As a general rule however, these representations are likely to include information pertaining to multiple different features of the current internal and external environments, such as the status of relevant objects in the environment and the ultimate goal towards which the agent is working.

While, in this chapter, we do not examine the production of sequential tasks in full, in order to facilitate our implementation of the novel architecture of the loop it is useful to conceive of several example distributed representations, comprising multiple conceptual features.

Though refraining from anchoring these hypothetical representations semantically so that we may consider the general case, they may be considered to represent distinct *stages* of one or more sequential tasks, where a ‘stage’ of a task is completed by the selection of an action by the motor loop, thus initiating the next stage. In order to constitute a valid and plausible illustration, these hypothetical representations should be partially overlapping, in order to reflect involvement of the same feature in multiple contexts (e.g., a common goal). However, our uniqueness constraint dictates that, while overlapping, these representations must also be at least partly orthogonal to one another. By adopting a feature-based scheme for their design and adhering to the 6x6 element network architecture in figure 2.1(b), we created a set of six simple representations with which to analyse the functional capabilities of the proposed architecture. These are illustrated in figure 2.2; each consists of a conjunction of two features and overlaps with two other representations.

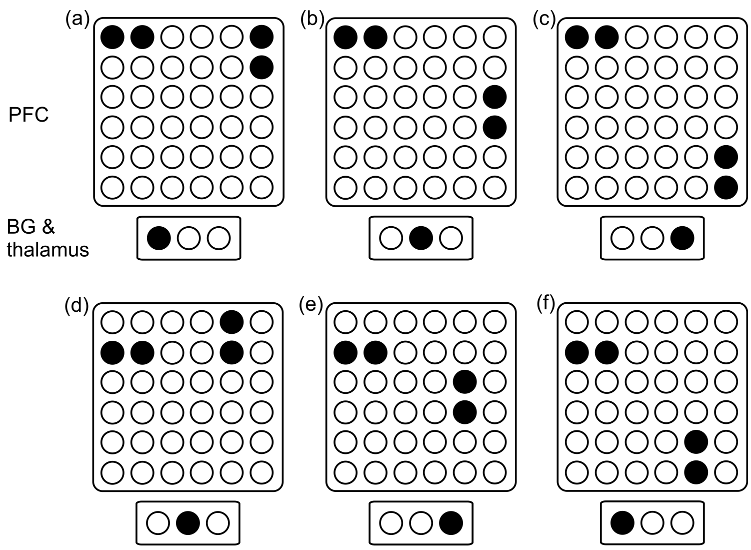

Figure 2.2: Illustration of six example feature-based representations as expressed by the recurrent network illustrated in figure 2.1(b), and the corresponding channels in basal ganglia and thalamus that support them. Each PFC representation comprises two features; each feature is represented by two adjacent nodes. Note overlapping features in a, b, & c, and d, e, & f. Each basal ganglia channel supports multiple representations (see text for details) Filled nodes = active; empty nodes = inactive.

Where we are examining the mediation of six distinct representations, we use only three channels in basal ganglia and thalamus to do so. It is important to note this disparity; in the novel architectural scheme we present here, rather than selecting a particular representation, each channel in basal ganglia is now responsible for disinhibiting a *subset of two* of the six possible representations. Thus, two distinct representations are simultaneously disinhibited by activation of a particular channel in basal ganglia. This results in excitation to multiple representations via a single disinhibited thalamic channel, and therefore has the potential to result in the simultaneous activation of two competing representations. However, intrinsic recurrence in PFC acts as a supplementary selection mechanism, resolving any remaining ambiguity and suppressing activation of the 'losing' representations.

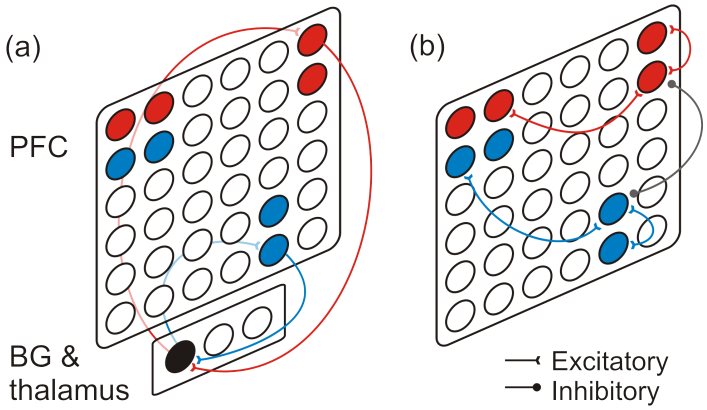

Figure 2.3: Schematic illustration of the connectivity underlying the subset-based selection scheme in the current architecture. (a) Two distinct PFC representations (illustrated in red and blue, respectively) send excitatory projections to a single channel in basal ganglia (illustrated in black), and receive influence from the corresponding single channel in thalamus. Note that only select connections are illustrated for clarity: all coloured PFC nodes project to and receive projections from this channel in this manner. (b) Within PFC, excitatory and inhibitory recurrent connections, within and between distinct representations respectively, support the selection of a single representation from among the subset selected by the single channel in basal ganglia. Again, illustrated connectivity is representative rather than exhaustive.

The connectivity between PFC, basal ganglia and thalamus allowing this selection scheme is illustrated in figure 2.3. Here, two distinct PFC representations are illustrated in red and blue (these correspond to representations (a) and (f) in figure 2.2, respectively). These distinct representations are both mediated by a single channel in basal ganglia and thalamus. As such, the eight nodes comprising these two representations send excitatory efferents to, and receive excitatory afferents from a single corresponding channel in thalamus. This results in positive feedback between this single channel in thalamus, and the eight PFC nodes comprising both of these representations (figure 2.3a). However, as mentioned, multiple representations are prevented from becoming simultaneously active due to the presence of recurrent inhibition between the competing representations within PFC. Individual nodes comprising a single PFC representation are, conversely, mutually excitatory (figure 2.3b). Note that, for simplicity in this example, the representations disinhibited by a single channel in basal ganglia are non-overlapping. This is not, however, a demand of the functional architecture. This idea is explored in chapter 4, in which more complex and more heavily overlapping representations are utilised for the mediation of a particular task.

2.6 Specific model description

We now go on to provide a specific description of the nuclei modelled and their connectivity. Unless notation indicates otherwise, projections between nuclei are topographic or 'channel-wise'. All notation is summarised in appendix A.

2.6.1 Basic neuron equation

As in Humphries & Gurney (2002), we implement a systems level neural network model, utilising leaky-integrator model neurons to represent neuron populations, where their output simulates the *average firing rate* of that population. Each leaky integrator has a dynamic activation value, a, defined by,

$$\dot{a} = \frac{-1}{\tau_m}(a - u), \tag{2.1}$$

where u is a weighted sum of inputs to that neuron, τ_m is the membrane time constant of that neuron governing the rate of decay, and $\dot{a} \equiv da/dt$. By extension, the equilibrium value $\tilde{a}$ of that neuron is given by,

$$\tilde{a} = u. \tag{2.2}$$

The output y of each neuron is calculated by the following piecewise linear function of its activation a:

$$y = \begin{cases} 0 & \text{for} \quad a < \epsilon \\ m(a - \epsilon) & \text{for} \quad \epsilon \le a \le 1/m + \epsilon \\ 1 & \text{for} \quad a > 1/m + \epsilon \end{cases} \tag{2.3}$$

where m is the gradient of the output function and ϵ the output threshold.

2.6.2 Prefrontal cortex

PFC is implemented as a 6x6 recurrent network (figure 2.1(b)) as a suitable mechanism to fulfill the criteria set out in section 2.1. Each element or node in PFC is represented by a single leaky-integrator neuron, described by equations 2.1-2.3, and each node received three contributions to its total input: an external salience signal, an excitatory projection from thalamus, and a weighted sum of output from all other PFC nodes.

PFC subsets: connectivity and notation

As discussed above, the functional relationship between PFC and the remaining nuclei in the loop may be regarded in terms of 'subsets' of multiple, stable cortical representations which are disinhibited by a single basal ganglia channel. This scheme relied upon a particular pattern of projections between PFC and the channel-wise regions of the model, whereby each channel received afferents from - and in turn projected to - all PFC nodes active for *any* of the representations within the subset selected by that channel.

Let X_{PFC} be the total set of PFC nodes, where X_{PFC} has nodes x_i, with $i = 1...N$. The subset of PFC nodes associated with basal ganglia channel i is denoted X_i. This is referred to by the index set J_i, where $J_i = \{k \in \mathbb{Z} : x_k \in X_i\}$, and where x_k is the k^{th} PFC node.

Thalamic contribution

Ventral anterior nucleus of thalamus (VA), widely regarded as the associative territory of thalamus (Haber & Calzavara, 2009), is modelled in the standard channel-wise form, and each channel i sends excitatory efferents to those PFC nodes in the corresponding subset X_i. If the output from VA channel i is y_i^V, the contribution from VA to PFC node x_j can be formalised by

$$u_j^{Vx} = W_{ij}^V y_i^V, \tag{2.4}$$

where W_{ij}^V is the synaptic strength from thalamus channel i to PFC node j.

Intrinsic contribution

Each PFC node projected to all others via a synaptic weight matrix, W^ρ. This matrix was hard-coded to support the representations illustrated in figure 2.2; full details are given in appendix B. The contribution from PFC to the inputs of a particular PFC node j may be written as

$$u_j^\rho = \sum_{i=1, i \neq j}^{N} W_{ij}^\rho y_i^x, \tag{2.5}$$

where N is the total number of nodes in PFC, y_i^x is the output of PFC node i and W_{ij}^ρ the synaptic strength from PFC node i to node j.

External salience

The external input to the model is described in terms of a set of six scalar ‘saliences’ corresponding to the six representations. The external input to the model may be conceived of as pre-processed multidimensional information reflecting the wide range of afferents to PFC originating in, for example, orbitofrontal, temporal and parietal cortices carrying limbic and secondary sensory information (see Miller & Cohen, 2001, and Tanji & Hoshi, 2008, for thorough overviews). Consistent with previous GPR models (Gurney et al., 2001a, 2001b) the strength of the salience to each representation reflects the relative ‘urgency’ for the selection of that representation (Redgrave et al., 1999). Each salience signal ζ_i is supplied to those nodes comprising the corresponding representation R_i. Where a single node is involved in two competing representations, the salience input to that node is the sum of the two competing saliences. The external input to PFC node x_j may then be described as

$$u_j^\zeta = \sum_{i=1}^{N^R} \zeta_i, \tag{2.6}$$

for $x_j \in R_i$, and where N^R is the total number of representations.

Summary

The total input to each PFC node may therefore be summarised by,

$$u^x = u^{Vx} + u^\rho + u^\zeta. \tag{2.7}$$

2.6.3 Basal ganglia and thalamus

We now describe in brief the architecture of the modelled basal ganglia and thalamus. These, as mentioned above, are faithful to the architecture described in Humphries & Gurney (2002), to which we direct the reader for further detail. Figure 2.1(a) presents a schematic diagram of the basic architecture.

PFC output: subsets

Each input nucleus of basal ganglia receives afferents from a single PFC subset X_i, as described above in section 2.6.2. The output Y^{X_i} from PFC subset X_i is quantified as the sum of the outputs y^x of all PFC nodes J_i comprising subset X_i, as follows:

$$Y^{X_i} = \sum_{k \in J_i} y^x_k. \tag{2.8}$$

Caudate

Caudate is accepted as being the primary associative region of striatum in the primate (Parent, 1990). In the TC-GPR model, striatum receives input from cortex and an external salience signal. However, in the interests of clarity given the new architectural distinctions between PFC and striatum, caudate here receives input only from PFC. If W^X is the synaptic strength from each PFC neuron, these inputs to caudate channel i from PFC may then be written as $W^X Y^{X_i}$. The facilitatory and antagonistic effects of tonic dopamine at the synapses of D1-receptor and D2-receptor expressing MSNs of striatum, respectively, are accounted for by a multiplicative factor λ_A, reflecting the tonic level of dopamine in associative striatum (Gurney et al., 2001a, 2001b). Inputs to D1-receptor expressing MSNs (the selection pathway, S) are enhanced by the factor $(1 + \lambda_A)$, and those to D2-receptor expressing MSNs (the control pathway, C) attenuated by $(1 - \lambda_A)$. Total inputs to each channel of caudate may thus be described by

$$u^S = W^X Y^X (1 + \lambda_A), \tag{2.9}$$

$$u^C = W^X Y^X (1 - \lambda_A). \tag{2.10}$$

Ventromedial STN

Ventromedial (vm) STN is considered to be the associative component of STN (Hartman-von Monakow, Akert, & Künzle, 1978; Parent & Hazrati, 1995b). In the current model, caudate and vmSTN receive identical PFC afferents. As a component of the control pathway, as described in the general introduction, vmSTN additionally receives an inhibitory influence from associative regions of GPe. If W^{XD} and W^{GD} are the synaptic strengths from PFC and associative regions of GPe, respectively, then

$$u^D = W^{XD} Y^X - W^{GD} y^G, \tag{2.11}$$

describes the input to vmSTN where y^G is the output of associative GPe.

Dorsomedial GPe

Generally, dorsomedial (dm) regions are accepted to comprise associative GPe (Joel & Weiner, 1997; Parent & Hazrati, 1995a). In the current model, dmGPe receives excitatory afferents from vmSTN and an inhibitory projection from D2-receptor expressing regions of caudate. Excitation from vmSTN is diffuse rather than topographic, thus we consider its contribution in terms of the sum of the output of its channels,

$$Y^D = \sum_{i=1}^{n^A} y_i^D, \tag{2.12}$$

where n^A is the total number of channels in associative basal ganglia. If W^{DG} indicates the synaptic strength of the vmSTN $\rightarrow$ dmGPe pathway, and W^{CG} that of the D2 $\rightarrow$ dmGPe pathway, then the total input to dmGPe is given by

$$u^G = W^{DG} Y^D - W^{CG} y^C. \tag{2.13}$$

Dorsomedial GPi

Associative regions of basal ganglia output nuclei are also generally considered to lie dorsomedially (Joel & Weiner, 1997; Sidibé, Bevan, Bolam, & Smith, 1997). Consistent with the TC-GPR model, we collapse both primary output regions - GPi and SNr - into a single functional region, here termed dmGPi. As a primary component of the selection pathway, dmGPi receives topographic inhibitory influences from D1-receptor expressing MSNs, but also receives activation via the control pathway as low level inhibition from GPe and a diffuse excitatory influence from vmSTN. If W^{SH} is the synaptic strength of the D1 $\rightarrow$ GPi pathway, W^{GH} that of the dmGPe $\rightarrow$ dmGPi pathway, and W^{DH} that of the vmSTN $\rightarrow$ dmGPi pathway, then inputs to dmGPi may be described by

$$u^H = W^{DH} Y^D - W^{SH} y^S - W^{GH} y^G, \tag{2.14}$$

where y^S and y^G are the outputs from caudate (D1 regions) and dmGPe respectively.

VA thalamus

Ventral anterior nucleus of thalamus (VA) receives afferents from PFC in the same manner as vmSTN and caudate, such that those PFC nodes composing subset X_i and projecting to channel i in caudate and vmSTN also project to the same channel i in VA. VA also receives topographic inhibitory projections from dmGPi. Let W^{XV} be the synaptic strength from PFC, and W^{HV} the synaptic strength from dmGPi. Inputs to VA may then be written as

$$u^V = W^{XV} Y^X - W^{HV} y^H, \tag{2.15}$$

where y^H is the output of dmGPi.

2.6.4 Parameters

The synaptic weight matrix W^{ρ} is detailed in appendix B. Synaptic weights W^X and W^{XV} were set to 0.25; this smaller strength compared with that in Humphries & Gurney (2002) compensated for the contribution from multiple PFC nodes. W^{XD} was set to 0.4. W^V_{ij} was set to 0.8 for PFC node $x_j \in X_i$, otherwise W^V_{ij} was set to zero. Synaptic strengths between all other basal ganglia nuclei were as described in Humphries & Gurney (2002), and all time constants and output parameters were also consistent with earlier work (but see appendix B for full details).

Selection of a PFC representation was deemed to have occurred when the output y^x of all nodes composing that representation reached the *selection threshold* $\theta = 0.9$. This threshold is inspired by evidence that action initiation is directly related to the level of activation in motor regions of cortex, manifesting as a threshold for action which is constant over tasks (Hanes & Schall, 1996). Though we are currently examining the selection of *cognitive* representations rather than action representations, this was deemed a suitable approximation by which to determine selection in the current systems level investigation. Noise was not included in the current simulations.

Simulations were implemented in MatLab and utilised a forward Euler update solution. All relevant simulation parameters are detailed in appendix B. Simulations were designed to establish the selection, maintenance and deselection abilities of the new associative TC-GPR architecture, and are described in detail in the following section.

2.7 Results

2.7.1 Selection between channels

Initially, the model was tested on its ability to resolve competition between two distinct PFC representations which were mediated by *different* channels in basal ganglia; this simulation therefore examined the complementary selection mechanisms in basal ganglia and cortex.

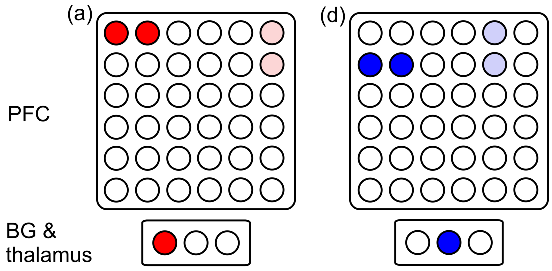

Figure 2.4: Replication of representations (a) and (d) and their corresponding basal ganglia channels. PFC representations consist of two features; PFC nodes representing feature 1 in both cases are filled in bold colour, and are shared across three representations (see figure 2.2). Nodes representing feature 2 are filled in pale colour. Colours correspond to the results graphs in figure 2.5

At stimulus onset, representations (a) and (d) (see figure 2.4) were supplied with external salience, taking values of $\zeta_a = 0.26$ and $\zeta_d = 0.25$, respectively. This external support was supplied for the duration of the simulation; this simulation therefore examines only selection, rather than maintenance. These values were chosen in order to examine the model's ability to distinguish saliences similar in magnitude, and its ability to integrate small input values to reach the greater selection threshold of $\theta = 0.9$. Outputs from the two distinct 'features' of both representations from this simulation can be seen in figure 2.5 (top and bottom left). Feature 1 overlaps with other representations; feature 2 is unique to the representation in both cases (see figure 2.2). Outputs of thalamus and GPi are also shown to illustrate channel-wise activity in the basal ganglia loop.

Low level activity is observed in both representations at stimulus onset. Between $t = 500ms$ and $t = 1000ms$, intrinsic excitatory connectivity in PFC causes increases in activity in both representations, and a corresponding *increase* in GPi activity due to fast transmission via the hyperdirect pathway through PFC → STN → GPi, as discussed in chapter 1. PFC recurrence begins to accentuate small salience differences between the two representations from approximately $t = 600ms$. This subsequently results in a selective *decrease* in GPi activity in channel 1, which mediates the most active PFC representation (a). In turn, this reduces inhibition on the corresponding channel in thalamus at roughly $t = 750ms$. This disinhibition allows activity in the corresponding thalamocortical loop to be rapidly integrated to reach the selection threshold $\theta = 0.9$, successfully demonstrating selection of the correct

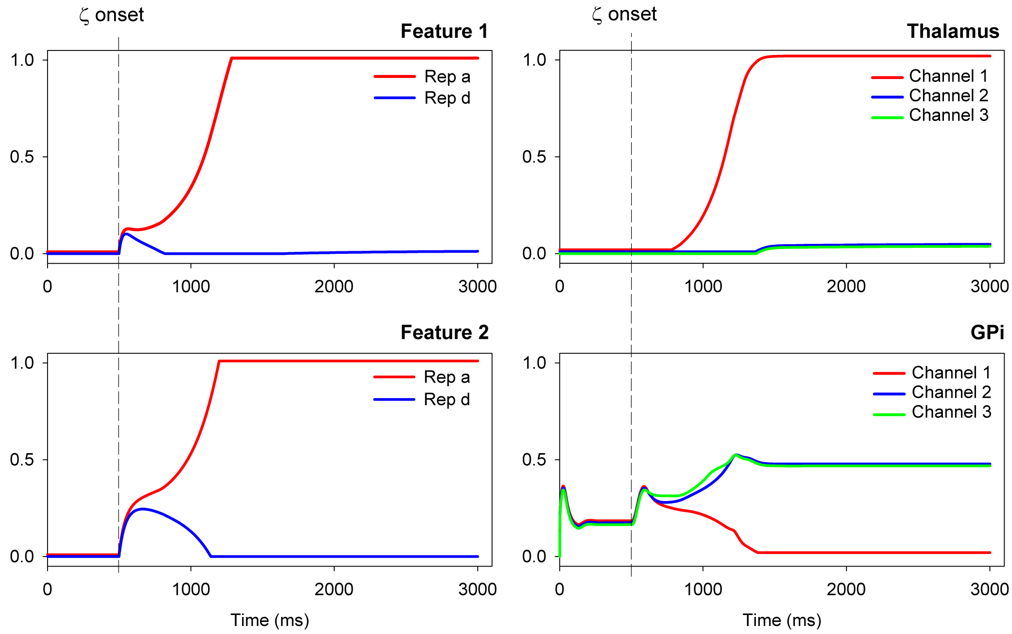

Figure 2.5: Output of key nuclei for simulation 2.7.1. External salience was supplied to competing representations (a) & (d) at $t = 500ms$. Left graphs show the output for features 1 and 2 of these representations, as illustrated in figure 2.4. Complementary selection processes in PFC and basal ganglia result in successful selection of the most strongly supported representation; activation of both features for representation (a) exceeds the selection threshold $\theta = 0.9$; activation of competing representation (d) remains well below threshold (see text for details).

representation.

Note that the time course of activity is slightly different for features 1 and 2 of the PFC representations. This is due to the overlap of feature 1 with other representations in both cases (see figure 2.2). Because of this overlap, feature 1 nodes receive support from all basal ganglia channels. As a result, stronger intrinsic PFC inhibition is required between these nodes than others. As this intrinsic inhibition takes effect more rapidly than the competition within basal ganglia, the initial separation of activation values is faster for competing feature 1 nodes than for feature 2. However, note that once competition begins to be resolved in basal ganglia, this supports selection in PFC and feature 2 ultimately reaches selection slightly faster than feature 1.

As representations (a) and (d) are supported by separate basal ganglia channels, in this scenario, basal ganglia takes an active role in the selection of the appropriate representation. Recurrence in PFC primarily assists by enhancing the effect of initial differences in ζ_a and ζ_d, such that basal ganglia is able to further resolve the competition and disinhibit the correct thalamic channel. PFC recurrence is also necessary for preventing the spread of activation to nodes which are not part of the winning representation, but which *are* supported by the disinhibited thalamic channel (in this case, the nodes comprising representation (f) receive excitation from basal ganglia channel 1, but are prevented from becoming active by direct inhibition from representation (a); see figure 2.2). It is important to note that such small differences in external salience values may be insufficient for successful selection *without* intrinsic PFC recurrence.

2.7.2 Selection within channels

This simulation was also concerned with the model's ability to select appropriately between competing representations; in this scenario, the competing representations are both supported by the *same* channel through basal ganglia and thalamus. As such, this simulation preferentially examines the ability of the intrinsic PFC connectivity to resolve this competition, where both representations receive excitation from the disinhibited thalamic channel. As before, representation (a) was supplied with salience of $\zeta_a = 0.26$; in this case however, representation (f) was supplied with salience $\zeta_f = 0.25$, where both (a) and (f) are supported by channel 1.

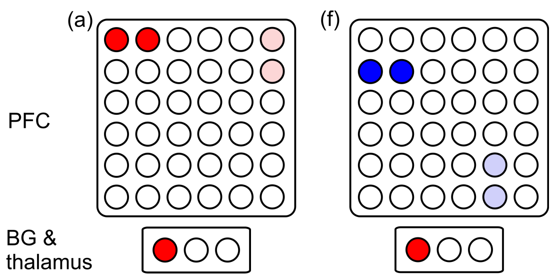

Figure 2.6: Replication of representations (a) and (f) and their corresponding basal ganglia channels. Again, PFC nodes representing feature 1 in both cases are filled in bold colour, those for feature 2 are filled in pale colour. Note that both representations are supported by the same basal ganglia channel, illustrated in red in both instances for consistency with the results graphs in figure 2.7

Results of this simulation are detailed in figure 2.7, and indicate the ability of the model to utilise intrinsic inhibition within PFC in order to suppress a simultaneously supported competing representation. Comparison with figure 2.5 indicates a similar overall activation trajectory in each of the modelled nuclei. However, by overlaying the activity of the winning PFC representation from simulation 2.7.1 (grey traces in figure 2.7), it is clear that selection of the winning representation takes place more rapidly here. This is also true of basal ganglia activity, and is expected due to reduced competition within basal ganglia itself. As this further implies, selection between the competing representations within PFC is quicker to be resolved.

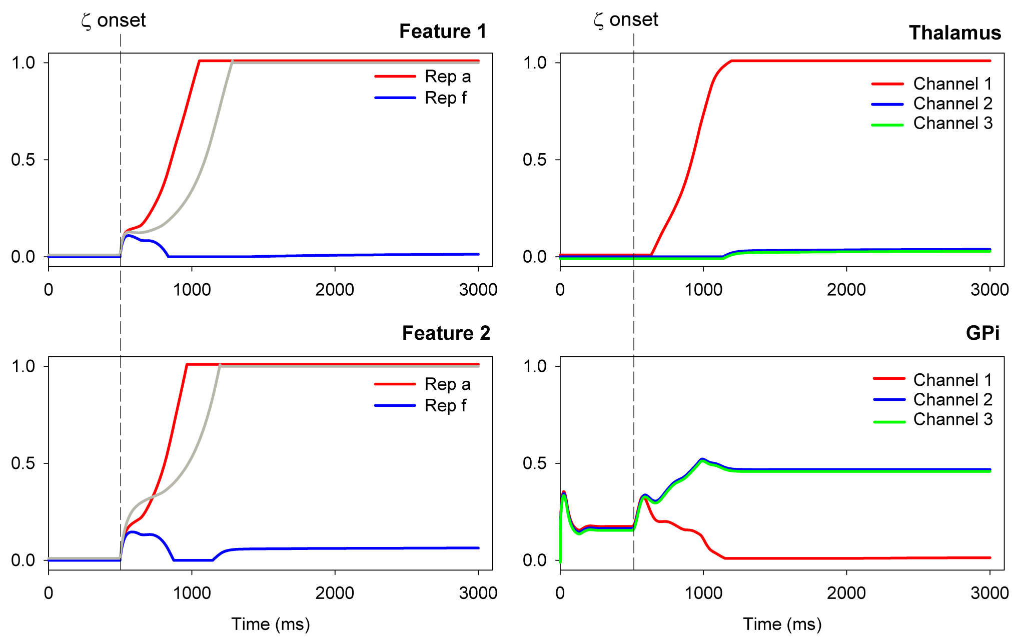

Figure 2.7: Output of key nuclei for simulation 2.7.2. Here, both representations are mediated by basal ganglia channel 1. Rapid selection of this channel is observed in GPi (bottom right) and thalamus (top right); intrinsic PFC recurrence rapidly determines selection of both features of the most strongly supported representation (a), while activity in the competing representation (f) remains below threshold $\theta = 0.9$. Grey traces in the left plots of PFC output show the activation of the winning PFC representation from simulation 2.7.1 for comparison.

2.7.3 Maintenance

This trial examines the model's ability to maintain above-threshold activity in selected cortical nodes after the external supporting salience to that representation is removed. This reflects the requirement for ongoing activation of a cognitive representation after a phasic initiating stimulus. At stimulus onset, representation (a) was supplied with external salience of magnitude $\zeta_a = 0.26$. For reasons of simplicity, in the current simulation no competing representations were simultaneously activated. Activity in PFC representation (a) was allowed to reach the selection threshold. At stimulus offset, $t = 1500ms$, external salience was reset such that $\zeta_a = 0$.

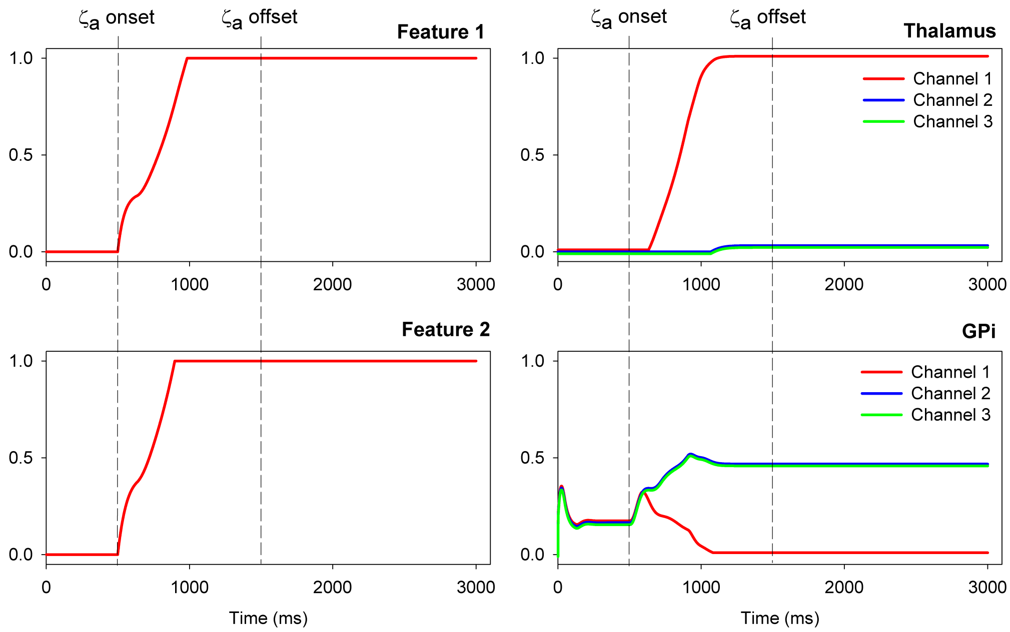

Figure 2.8: Output of key nuclei for simulation 2.7.3. Selection of the supported representation (a) is observed after salience onset at $t = 500ms$. Once selection is achieved in cortex, external support to representation (a) is removed at $t = 1500ms$. Continued disinhibition of the corresponding thalamus channel 1 allows activation of the selected representation (a) to be maintained after the removal of external support.

Figure 2.8 illustrates sustained activation of the selected representation upon stimulus offset at $t = 1500ms$. No reduction in PFC output is observed, indicating the ability of the model to maintain a 'working memory' representation of a prior stimulus, whereby high activation in cortex causes continued suppression of SNr output and subsequent maintained disinhibition of thalamus, allowing the thalamocortical loop to retain a high level of activation.

2.7.4 Deselecting a maintained representation

Having demonstrated sustained activity, we now examine the model's capabilities to deselect or *switch* from an active sustained representation to a new pattern. Again, representation (a) was supplied with a salience value of $\zeta_a = 0.26$ until the selection threshold $\theta = 0.9$ was reached; external salience was subsequently reset to $\zeta_a = 0$ at $t = 1500ms$. At $t = 2000ms$, external salience was introduced to a competing representation, (d). The competing representations are illustrated in figure 2.4. Note that representations (a) and (d) are supported by distinct channels in basal ganglia and thalamus, requiring a 'switch' in the selected channel in basal ganglia in addition to PFC.

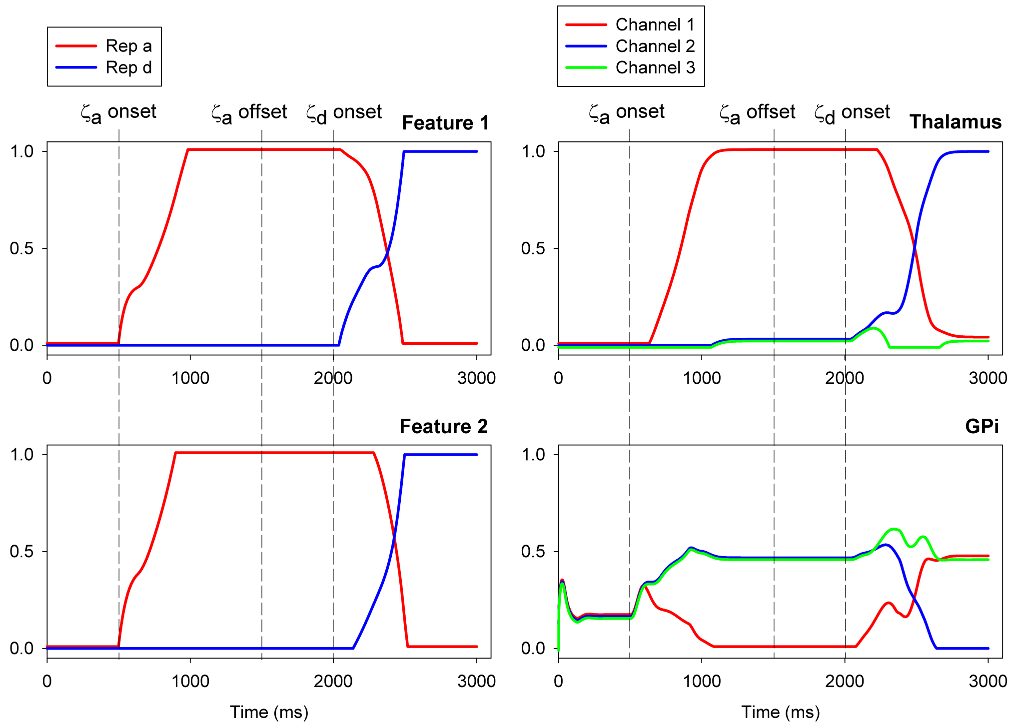

Figure 2.9: Output of key nuclei for simulation 2.7.4. Maintenance of a selected representation (a) is observed as in simulation 2.7.3. After the removal of salience to representation (a) at $t = 1500ms$, external support is applied to representation (d) at $t = 2000ms$. This representation is supported by a distinct channel in basal ganglia. Switches in all nuclei are subsequently observed, demonstrating the model's ability to deselect a maintained representation on the basis of sufficient competition.

Figure 2.9 clearly demonstrates a clean and rapid switch in all modelled nuclei. Initially, with the introduction of additional activity in cortex, the overall amount of activation in the model is increased, resulting in peaks of activity in GPi as salience is propagated through the fast hyperdirect pathway. This increase - particularly in the previously inhibited GPi channel 1 - rapidly inhibits thalamocortical loop activity for the corresponding channel. Subsequent suppression of GPi channel 2 activity allows activation in the thalamocortical loop involving channel 2 to integrate the newly supported representation to the selection threshold. Note, however, that a significantly greater salience value of $\zeta_d = 0.4$ was required in order to cause switching from a currently selected competing representation; lower values of ζ_d were insufficient to result in switching behaviour (not demonstrated here). This is consistent with Humphries & Gurney (2002), in which comparatively higher salience values were required to cause switching behaviour.

2.7.5 Deselecting a supported representation

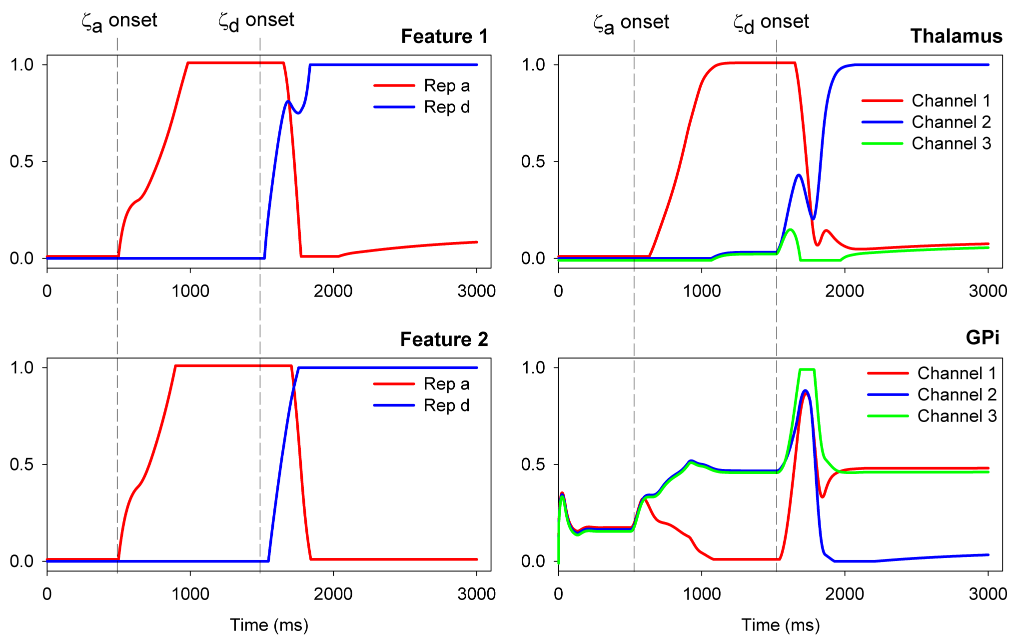

Figure 2.10: Output of key nuclei for simulation 2.7.5. Here, switching behaviour is observed from a currently selected representation (a) to the competing representation (d), after the onset of salience to representation (d) at $t = 1500ms$. In contrast to simulation 2.7.4 however, representation (a) continues to receive external support for the duration of the simulation. Despite this continued support of the initially selected representation, switching is still possible given sufficient external support to the competing representation.

It is reasonable to infer from the results of simulation 2.7.4 that the model should also be able to switch from an externally supported selected representation; however, it is likely to require significantly stronger external salience to do so. We ran simulation 2.7.4 again, omitting the offset of ζ_a, and increasing the salience to the competing representation to $\zeta_d = 0.6$. Figure 2.10 displays the results of this simulation. Again, we see a successful switch, though with exaggerated activity peaks in GPi due to higher overall levels of activity at the onset of ζ_d, and slight interference in overlapping feature nodes and thalamus during the switch. This interference is likely due to an initial rise in thalamic activity in the competing channel 2 being temporarily and incompletely suppressed by a brief peak in GPi activity. However, as competition is resolved, GPi channel 2 activity is inhibited and selection of the new representation is achieved. Again, however, note the greater salience value of $\zeta_d = 0.6$ that was required for this switch. This further increase in external support was necessary to counter the increased competition resulting from the ongoing support to the representation (a).

2.7.6 Progressing through a task

We now examine the model's ability to switch between representations which share a particular feature. If we imagine that this shared feature represents a common goal, we might imagine that this switching behaviour reflects the dynamics underlying the progression through subsequent stages of a task. Importantly, task features such as 'goal' have a longer relevant temporal duration than others, which might relate only to a single action. It is thus important that any sequencing mechanism is able to maintain activation of those features which endure for multiple stages of a task.

Figure 2.11 replicates representations (a) and (b) from figure 2.2, which share a common feature 1, illustrated in bold colour. Note that though representations (a) and (b) overlap in PFC, they are mediated by separate channels in basal ganglia. At stimulus onset, salience $\zeta_a = 0.26$ was presented to representation (a). At $t = 1500ms$ this salience was removed, and $\zeta_{b'} = 0.4$ was presented only to the *non-overlapping nodes* of representation (b). This simulation therefore tested the model's ability to sustain activation in feature 1 during switching in basal ganglia and of feature 2. Indeed, the model was able to perform an appropriate switch in both basal ganglia and PFC nodes, as demonstrated by the graphs

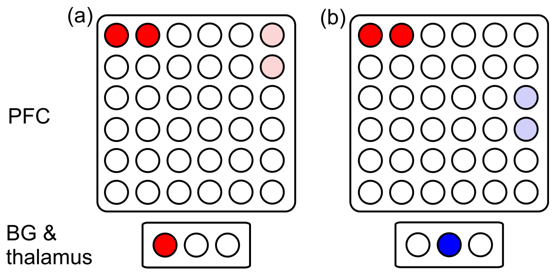

Figure 2.11: Replication of representations (a) and (b) and their corresponding basal ganglia channels. Again, PFC nodes representing feature 1 in both cases are filled in bold colour, those for feature 2 are filled in pale colour. Note that though the representations share the same feature 1, illustrated in red in both instances for consistency with the results graphs in figure 2.12, these representations are supported by distinct channels in basal ganglia.

displayed in figure 2.12. This indicates the model is able to effectively move through stages of a task without deactivation and subsequent reactivation of enduring features.

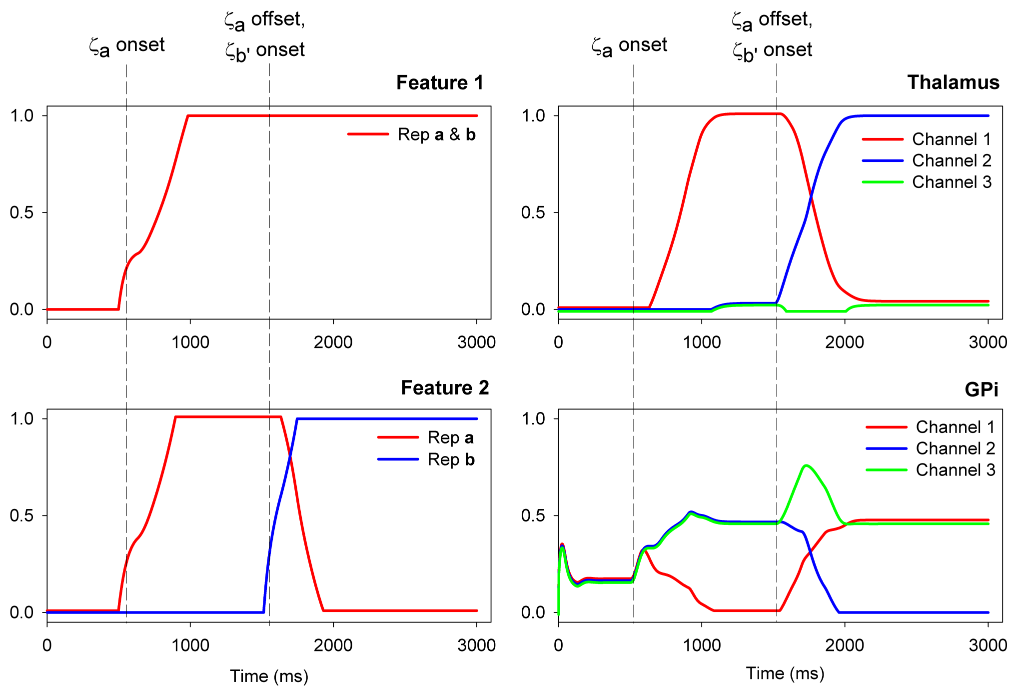

Figure 2.12: Output of key nuclei for simulation 2.7.6. Representation (a) is selected after salience onset at $t = 500ms$. At $t = 1500ms$, salience to (a) is removed and simultaneously applied to representation (b). Note that representations (a) & (b) share the same feature 1. The model is able to sustain activity in the shared feature while switches occur in basal ganglia and feature 2. This pattern of activity might be expected from the progression through two stages of a task which share a common goal, the representation of which stays constant through the switch.

2.7.7 Resistance to interference

We also explicitly examine the model's behaviour under conditions of interference, and its ability to maintain a selected representation under transient, moderate competition from a second representation. To remain consistent with simulations of the same nature in Humphries & Gurney (2002), we applied salience $\zeta_a = 0.26$ to representation (a) and allowed activity in this representation to reach the selection threshold $\theta = 0.9$. We then transiently applied external salience of the same magnitude $\zeta_d = 0.26$ to the competing representation (d) for $1000ms$.

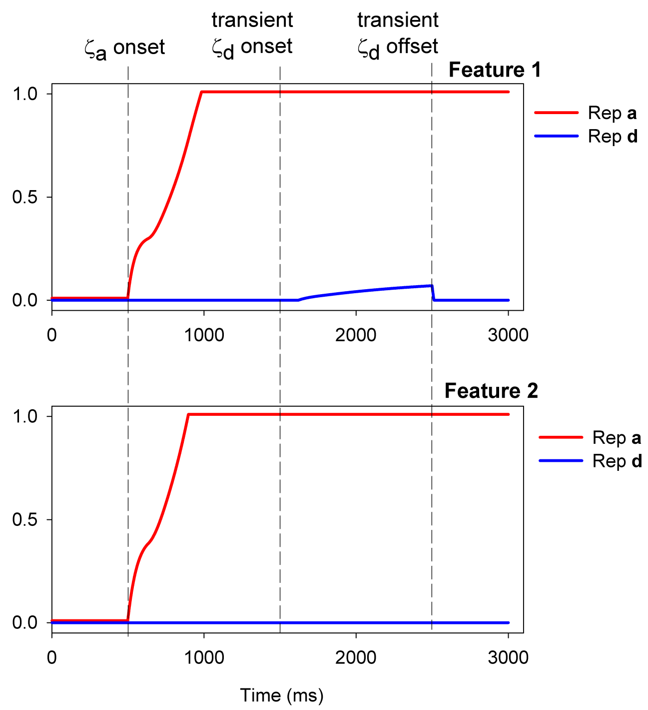

Figure 2.13: Output of key PFC nodes for simulation 2.7.7. Activity of PFC nodes for the selected representation (a) shows resistance to transient interference from a competing representation (d) for moderate levels of salience. While low level activity is observed in the competing representation, this does not approach the selection threshold $\theta = 0.9$, and no disturbance to the activity of the selected representation is observed.

Output of feature nodes for representations (a) and (d) are shown in figure 2.13. At the onset of salience to the competing representation at $t = 1500ms$, small amounts of activity are observed for the competing representation feature 1 nodes. It is likely that this activation is observed for feature 1 and not feature 2 due to lower overall recurrent inhibition on feature

1 nodes resulting from their involvement in multiple representations, in addition to the extra support from thalamus received by feature 1 nodes, also as a result of their participation in multiple representations supported by multiple basal ganglia channels. However, this activity does not interfere with the currently selected representation, which is maintained throughout the duration of the application of salience to the competing representation. This clearly shows the ability of the novel architecture to resist interfering salience to competing representations.

2.7.8 Lesion studies

(i) Lesioning basal ganglia output

This simulation aimed to confirm the necessity of basal ganglia involvement in selection and maintenance of representations. Recurrent networks, such as we have implemented as PFC, have been shown to have far greater capacity than required for this model (Wu et al., 2012), raising the question of the necessity of basal ganglia for selection with such a network. However, the interaction of PFC with basal ganglia allows greater *economy* of intrinsic recurrence; only those representations which are supported by the same channel in basal ganglia must inhibit one another strongly. Inhibition from basal ganglia therefore effectively reduces the space in which PFC must apply inhibition in order to generate stability. Here, we wanted to confirm that this more efficient scheme in fact depended on basal ganglia.

We lesioned the projection from GPi to thalamus, resulting in a simple excitatory feedback system between PFC representations and thalamus. We then re-ran simulation 2.7.1, where two representations were presented with external support simultaneously. As can be observed in figure 2.14(i), the model was unable to resolve the competition between these representations, and we see 'dual selection' of both representations (a) and (d). In addition, we also observe spreading activation from the two externally supported representations (a) and (d) to others via excitation from additional channels in thalamus. This indicates that the selection properties of basal ganglia are required for desirable performance with economic recurrence in PFC. It is also possible that this interaction of a recurrent network and a central selection mechanism results in greater overall storage capacity than a recurrent network alone; this idea is discussed in detail in section 2.8.

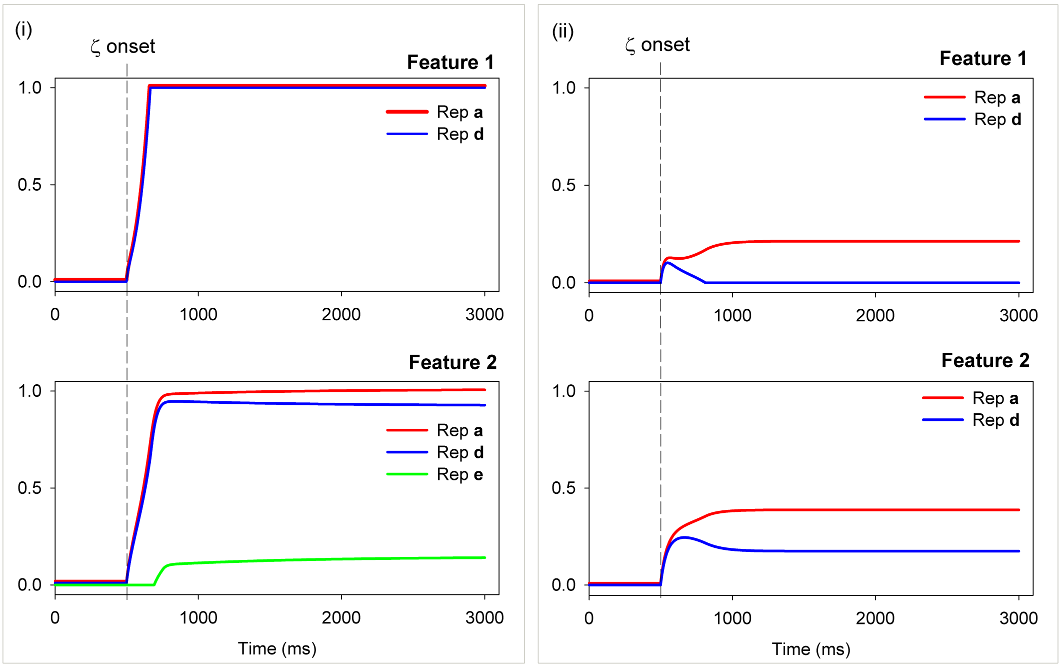

Figure 2.14: Output of key PFC nodes for simulation 2.7.8(i) & (ii). (i) In the absence of inhibition from GPi, the model is unable to resolve competition between representations (a) & (d). Uninhibited excitation in thalamus allows rapid integration of both representations to the threshold $\theta = 0.9$, resulting in multiple selection. Also notice the spread of activation via excitatory recurrence into a third representation (e). (ii) Without recurrent excitation from thalamus, no representation reaches the selection threshold $\theta = 0.9$. While some separation of the supported representations (a) & (d) is observed, resulting from intrinsic PFC recurrence, neither representation can be said to be 'selected', indicating the critical importance of thalamic activity for selection, and, by extension, of basal ganglia.

(ii) Lesioning the thalamocortical projection

To further explore this idea, we also examined the performance of the recurrent network in isolation from thalamus. Again, we re-ran simulation 2.7.1 after lesioning the thalamocortical projection. Results (figure 2.14(ii)) demonstrate the insufficiency of the PFC alone to mediate the task. Whilst a greater delineation is observed between the two representations than that seen in simulation 2.7.8(i), and no spread of activation to other representations occurs, neither representation is integrated to a level of activity required for selection. This emphasises the critical role of the disinhibition of thalamus - and subsequent recurrent excitation in the thalamocortical loop - for selection.

It is difficult to show the requirement for the BGTC loop as a whole in the role of maintenance in isolation from its role in selection; as study 2.7.8(i) shows, lesioning basal ganglia output causes selection capabilities to break down and interference between distinct representations. Though not demonstrated here, this also occurs when external salience is applied only to a single PFC representation. We cannot assess the ability of a lesioned model to maintain activity in a single representation without its prior selection. However, simulation 2.7.8(ii) demonstrates the important role of the thalamocortical projection in the integration of activity to the selection threshold in a single representation. As such, we contend that both selection and maintenance functionality rely on the integrity of the loop as a whole in the current model.

2.8 Discussion

2.8.1 Summary of findings

The results presented above demonstrate the suitability of the novel architectural scheme we present for the mediation of selection, maintenance and timely deselection of multiple cortical representations, where those representations outnumber the total channels in basal ganglia. We have shown that the model is able to resolve competition between multiple representations, both within and between subsets, and is able to switch between expressed representations upon the introduction of a sufficiently strong competing input salience, whilst maintaining expression of a single representation in the presence of interference. Furthermore, by lesioning connectivity in the model we have demonstrated the advantages of implementing complementary selection mechanisms in basal ganglia and cortex for the selection and maintenance of cortical representations in the current neurally inspired architecture.

2.8.2 Effects of functional constraints

The present model was constructed in order to ensure adherence to a set of pre-determined functional criteria. A number of constraints from anatomy and neurophysiology were also considered, leading to a novel structural and functional organisation of the associative loop, whereby basal ganglia disinhibited *subsets* of cortical representations in order to solve a complex incompatibility regarding the coding schemes employed in basal ganglia and cortex, which we now discuss in detail.

Form unique representation of the temporal task context.

While this constraint may seem self-evident, it is important to note explicitly due to its direct relevance to the heavy demands on the underlying substrate. More specifically, representations of context are likely to incorporate many individual items of information; an adaptive solution to this problem requires some means of combining them in a flexible manner to comprise a unique overall representation. In the present study, we have structured the illustrative representations around the basis of a conjunction of currently relevant features. While we have not examined the production of sequences *per se*, in chapters 4 and 5 we go on to show that this organisation allows the expression of sufficient information to uniquely represent individual stages of multiple tasks. Importantly, although each representation shared components of its pattern of activity with at least two other distinct representations, the overall pattern was unique for each. While semantics were not specifically referred to here, we suggest that these representations are able to reflect a specific temporal task context, and thus effectively provide a cognitive specification for a particular action to be performed.

Efficiently represent task context

This criterion imposed heavy constraints on the structure of the model, and led to an intriguing analysis of the relationship between basal ganglia and PFC. Ultimately, this particular consideration guided the proposition that a degree of information loss is possible in basal ganglia, and that this may be reflected in the degree of convergence in corticostriatal axons. This in turn required additional selection mechanisms in PFC, complementary to those documented in basal ganglia, in order to disambiguate representations beyond the level of basal ganglia sensitivity. However, the contribution of basal ganglia to the selection process allowed a greater economy of connectivity within PFC than would otherwise be required for stability. Most notably perhaps, this scheme allowed fewer channels in basal ganglia than the total number of representations, whilst maintaining the crucial involvement of basal ganglia in the selection process, and its associated benefits, such as resisting interference.

While we have investigated the mediation of only six representations in a network of 36 nodes, the feature-based representation scheme we have adopted indicates, in principle, the ability of the model to support more unique patterns than would be possible with a fully localist scheme, further increasing efficiency. It is notable here that each representation has a unique component as well as a shared feature; this effectively disambiguates one representation from another on the basis of a single feature. It is however possible - even likely - that more realistic representations have no individual unique features and the overall pattern of activity is the only means of distinguishing one representation from another. Though we have focused on simple and partly orthogonal representations, within the current architecture - excepting specific patterns of weights - we would expect the system as configured to display the same selection abilities with no individual distinct features. Indeed, this is explored in detail in chapter 4.

Autonomously maintain selected representations

Again, this criterion required particular architectural features; given the other considerations on model function - in particular, the requirement for deselection - the excitatory thalamocortical loop is proposed as the primary mechanism underlying maintenance, in concert with intrinsic connectivity in PFC. Our lesion studies explicitly showed the requirement for positive feedback from thalamus for this, further implicating the continued involvement of disinhibition from basal ganglia during maintenance. This suggests complementary roles for basal ganglia and PFC in both selection and maintenance. Though strong maintenance was observed, this scheme was shown to allow timely deselection of a currently maintained representation, thus indicating the flexibility of working memory traces in the current model.

2.8.3 Significance of the novel architecture

Underlying mechanisms

The novel architecture that resulted from the constraints discussed above determined a unique underlying functionality, allowing the model to fulfil the proposed criteria. Specifically, the selection behaviour of the associative loop may be understood in terms of attractor dynamics. We suggest the new organisation results in a 'two-tiered' attractor space in the BGTC loop. Via the disinhibition of subsets of cortical representations, basal ganglia causes a movement of the activation of PFC into a coarse level attractor space; we term this a *subset-attractor space*. Within this subset attractor space, there are multiple finer-level fixed point attractors, each corresponding to an individual representation within the subset; we refer to these as *state-attractors*. These ideas are visualised in figure 2.15. The precise pattern of external input to PFC determines which of these state-attractors the pattern of activity in PFC ultimately settles into. Although the present study only examines the function of the model with input to PFC, this implies that by directly modulating activity in caudate via additional influences, the selection process can be modified in terms of which subset-attractor space is ultimately reached, but not the state-attractor.

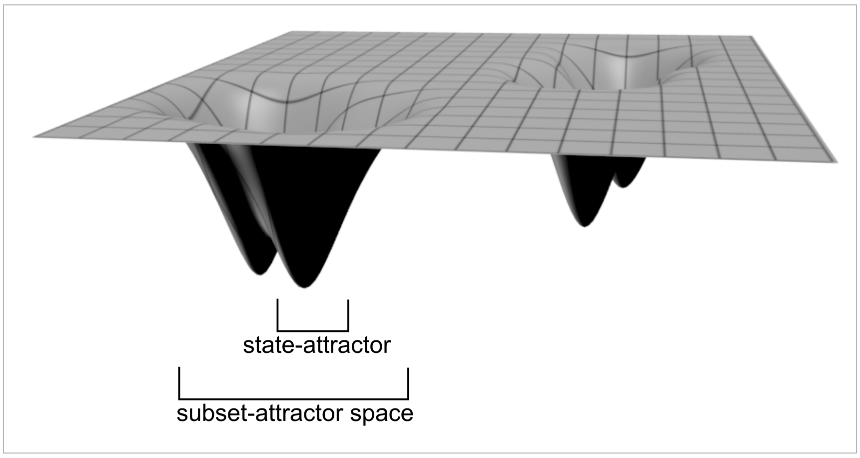

Figure 2.15: Simplified schematic diagram of the hypothesised attractor space of the associative loop system, indicating its possible structure in terms of *subset-attractor spaces* and *state-attractors*. See text for details.

This perspective helps explain the difference in latency for selection when competing representations lie within the same or different subsets, or coarse level attractors (see comparison in figure 2.7). The amount of energy required to move between state-attractors within a subset-attractor space may be less than that required to move *between* subset-attractor spaces. It follows that a greater latency would be expected for resolving competition between competing attractors that are further away in state space, than those that are close by, given the same magnitude of external support.

Storage capacity

Where we have equated 'efficiency' with the ability to represent information in a distributed manner, the efficiency of the model is also affected by its ability to *store* patterns. In a typical Hopfield network, depending on the precise algorithms used to generate stability, the theoretical capacity limit tends to vary between $0.15N$ (Hopfield, 1982) and N (Wu et al., 2012), where N is the number of elements in the network.

Earlier, we pointed out that the structural constraints imposed by neuroanatomically plausible patterns of connectivity in our model might impact on the storage capacity of the system compared with, for example, a Hopfield net of the same size (54 elements). However, the number of individual representations that may be supported by the current model is likely to depend on a number of factors, including the number of features involved in the representation, the number of active and inactive nodes in each pattern, the degree of overlap between different representations, the number of channels in basal ganglia and the number of representations supported by a single channel. In the current demonstration, we have only concentrated on the mediation of six separate representations (though not demonstrated here, all patterns were observed to be stable), which is below the more conservative theoretical capacity limit of the system at $0.15N = 8.1$. However, given the number of 'unused' nodes in PFC (see figure 2.2), we contend that it is likely that a greater number of representations may be supported by the current model. As such, the novel interpretation of the BGTC architecture does not appear to have impacted negatively on the theoretical storage capacity of the system compared with a Hopfield net of equivalent size. Indeed, in chapter 4 we go on to show that the system is capable of sustaining a greater number of patterns than suggested by this more conservative theoretical limit.

2.8.4 Comparison with existing models

Several models of PFC interactions with basal ganglia have been proposed with specific reference to working memory and the encoding of task context. To our knowledge however, no existing models of the basal ganglia at this level of the BGTC hierarchy specifically address the questions we examine here; namely, the selection, maintenance *and* deselection of distributed cortical representations with reference to additional constraints based on the plausibility of particular anatomical relationships between PFC and basal ganglia. However, it is useful to examine the mechanisms utilised by previous models of this region of the BGTC hierarchy in order to examine in context the focus we have taken here, and the combined insights from this and previous work.

Mechanisms for working memory in basal ganglia models

The importance of basal ganglia for adaptive working memory is widely accepted, with support from experimental studies of various flavours (Baier et al., 2010; Lewis, Dove, Robbins, Barker, & Owen, 2004; McNab & Klingberg, 2008; Voytek & Knight, 2010), and most models of basal ganglia at associative levels of the BGTC hierarchy address the issue directly (O'Reilly, 2006). However, the precise role of basal ganglia in working memory is debated, with various accounts emphasising the gating of information into working memory (Beiser & Houk, 1998; Brown, Bullock, & Grossberg, 2004; Gruber, Dayan, Gutkin, & Solla, 2006; O'Reilly & Frank, 2006; Vitay & Hamker, 2010), versus maintenance itself (Dominey & Arbib, 1992; Monchi & Taylor, 1999; Schroll, Vitay, & Hamker, 2012; Taylor & Taylor, 2000).

A significant number of different mechanisms have been explored for the purpose of maintaining information in PFC over time (Curtis & Lee, 2010), many of which do not rely on basal ganglia involvement for maintenance. Those models that have used distributed representations in cortex have tended to rely on cortico-centric mechanisms for maintenance, such as bi-stability in cortical cells (Frank et al., 2001; O'Reilly & Frank, 2006) and local recurrence (Amos, 2000; Gruber et al., 2006; Ponzi, 2008). Conversely, those models which primarily utilise the excitatory thalamocortical loop to maintain information, as we have implemented here, tend, conversely, to rely on a fully localist architecture (Dominey & Arbib, 1992; Monchi & Taylor, 1999; Schroll et al., 2012; Taylor & Taylor, 2000). As we have argued above, this is an unlikely means by which cortex represents information at this level of the hierarchy. We are unusual, then, in proposing an architecture which utilises thalamocortical feedback for maintenance in concert with distributed representations, and thus cannot directly compare the performance of our model with existing work. However, we look in detail at two particularly influential narratives with implications for the understanding of the present model.

Dominey and colleagues (Dominey et al., 1995; Dominey, 1995) utilise distributed PFC representations of context for a saccade task requiring active maintenance of sequential information. Superficially, these models appear to be quite similar in structure to the one we present here. However, the authors use a unique combination of mechanisms to maintain state information, including recurrent excitation with thalamus as we have focused on, as well as motor efference copy from output structures (in this case, superior colliculus) and a damped self input from a second PFC layer incorporating several different time constants. While these models are unquestionably powerful and have yielded many useful insights, whether true self-maintenance is achieved is unclear; several of their simulations involve tonic visual input, which may arguably provide critical ongoing support to maintain the current state. Performance of the model under conditions of phasic stimuli was still impressive at 85% correct performance (though performance suffered compared to tonic stimuli conditions). However, it is difficult to determine whether the model would be capable of the same degree of maintenance in the absence of implausibly long time constants in PFC, or alternatively whether the model could still perform well with longer delays between the presentation of stimuli. In short, information in PFC is maintained for a critical duration in order to allow response selection, but perhaps not indefinitely. It seems likely that, left without new visual or efference copy input for more than a few seconds, the representation would begin to decay. Moreover, a minimal representation of basal ganglia is employed, and additional selection takes place in superior colliculus itself, suggesting that, in this model, basal ganglia is predominantly a locus for associative learning rather than performing crucial roles in selection or maintenance *per se*.

Frank, O'Reilly and colleagues (Frank et al., 2001; Hazy et al., 2007; O'Reilly & Frank, 2006) have produced what is arguably the most substantial modelling contribution to understanding working memory functions of the basal ganglia to date. Again, these models appear superficially similar to the model we present in this chapter, particularly with their focus on feature-based cortical representations. They, however, emphasise selection - or more specifically ‘gating’ - functions of the basal ganglia, positing maintenance solely in bi-stability of cortical neurons, rather than in the interactions between cortex and basal ganglia. Indeed, they actively discount recurrent excitation in thalamocortical loops as a mechanism for maintenance of distributed representations, suggesting that it is not possible to maintain multiple representations in cortex with this mechanism if adhering to the neuroanatomical principle of corticothalamic convergence (Frank et al., 2001). Here however, we have shown this not to be the case. Following from this, they also claim that thalamocortical interaction for maintenance of working memory traces is impractical because it requires the continued disinhibition of thalamus for the duration of the maintenance period. Conversely, we argue that this continued disinhibition is necessary, due to the requirement of basal ganglia-based inhibition for *de*selection. Notably, this is a feature which few, if any, models of cortical-basal ganglia interactions explicitly address, though some neurophysiological evidence exists to suggest the ongoing involvement of basal ganglia during maintenance (Cromwell & Schultz, 2003). Without this inhibition-driven deselection, the inactivation of a selected representation (or a component thereof) remains an open question (Beiser & Houk, 1998; Frank et al., 2001), as does the general function of tonic inhibition from basal ganglia, particularly when maintained activity is observed in the presence of ‘NoGo’ activity in striatum (Hazy et al., 2007). We argue that the gating role postulated for basal ganglia in these models is further limited, where their framework implies that there is no direct competition between channels in basal ganglia. In these models, PFC consists of a series of ‘stripes’, each of which represents a particular category of information, for example the current task. Stripes consist of a series of nodes, each node representing a particular instance of that category. The particular pattern of activity over all stripes denotes the overall task context. Each node in basal ganglia is responsible for gating information into a single stripe. Theoretically, all PFC stripes may be updated simultaneously, providing no role for the well-evidenced competitive selection function intrinsic to basal ganglia. Additionally, their architecture may struggle with simultaneous competing inputs to the same stripe: either basal ganglia updates the stripe, in which case both stimuli would be stored in working memory, or it does not, in which case neither would.

In contrast to these models, the present study provides mechanisms for true competition based selection - rather than gating - in basal ganglia, and true maintenance of selected representations. Thus, we do not suffer from the potential for ambiguous updating nor slow decay of cortical representations. This unique combination of functionalities in a model of the BGTC loop architecture stems directly from the novel organisation we proposed at the outset.

Short versus long term memory

It is important to note that a significant contribution to the maintenance of representations in our model is provided by hard wired intrinsic recurrence in PFC. While the sustained activation of the representations may be considered working memory, this weight-based storage of representations is likely to reflect longer term memory, and thus the activity in our model may be understood as the retrieval of long term memory traces into working memory. In contrast, many of the studies of working memory examined above consider tasks in which information is maintained over time, but is neither retrieved from nor later stored in long term memory, which would necessitate alterations in the patterns of intrinsic connectivity. Other models rely on a dynamic approach to the maintenance of information, where the nature of the stored representation changes over time (Botvinick & Plaut, 2006a; Rougier & O'Reilly, 2002). To an extent this is also true of the Dominey et al models (Dominey et al., 1995; Dominey, 1995), where dynamic representations record accumulating task history. A key question for future research then is when active maintenance of *information* differs from active maintenance of *a single pattern of activity*. Indeed, models are now exploring the possible mechanisms underlying the formation of PFC representations for both activation-based and weight-based working memory (Botvinick & Plaut, 2006a; Reynolds & O'Reilly, 2009; Rougier & O'Reilly, 2002; Rougier et al., 2005), though these models have yet to be implemented within a biologically plausible framework with basal ganglia.

Information compression

Interestingly, of those models of PFC and basal ganglia which stress the convergence of corticostriatal axons, redundancy of cortical representations and dimensionality reduction is either emphasised as the very reason convergence is observed (Bar-Gad et al., 2003), or stressed as a negative consequence of implementing thalamocortical mechanisms for working memory, and to be avoided (Frank et al., 2001). Here, we are able to reconcile the two approaches, by allowing recurrent thalamocortical mechanisms for working memory, whilst preserving information in cortex which is minimally redundant, despite a coarse level of sensitivity in basal ganglia leading to a loss of information in basal ganglia itself. The subset-based selection scheme we have implemented, in concert with recurrence in PFC, allows a flexible approach to the problem of information preservation in the presence of corticostriatal convergence.

2.8.5 Predictions

The model presented here results in a number of predictions. Firstly, reactivated working memory traces of stored long term memory should be identifiable by distinguishable patterns of activity which change little over time, either during a trial or between distinct events. Secondly, our model suggests that lesioning any of the pallido-thalamic projection, the thalamocortical projection, or intrinsic connectivity within PFC should result in specific patterns of degradation in the quality of working memory representations; basal ganglia output lesions would be expected to result in a difficulty selecting or switching between distinct representations, and interference would be expected. Lesions within cortex or the thalamocortical loop should result in difficulties maintaining any selected representation; depending on the severity of the lesion, extremely salient stimuli may also be required to result in selection at all. A degree of support for such lesion effects may be seen in Voytek & Knight (2010), but specific deficits were unclear and a great amount of work in this area is required before confident conclusions may be drawn.

2.8.6 Limitations

While the present study provides a comprehensive systems-level account of the possible mechanisms underlying selection and maintenance of distributed representations in PFC, the abstractness of the model inevitably omits known nuances of the neuroanatomy and neurophysiology which are clearly critical for much of this functionality. For instance, intrinsic bi-stability has been demonstrated in prefrontal neurons and is likely to play an important role in sustained activation. As mentioned above, we do not account for task-specific working memory representations which do not rely on long term weight-based storage, and those weights we do rely on are hard-coded, providing no account of the possible learning mechanisms at hand. More in-depth models utilising spiking networks and more accurate accounts of neuromodulators such as dopamine, and the incorporation of plasticity, may begin to address some of these points and establish the ability of the proposed functional architecture to generalise to different levels of abstraction.

2.8.7 Summary

The present chapter has established a general architecture based on the neuroanatomy of the BGTC loop for selecting and maintaining distributed representations in PFC. These representations are most easily understood as cognitive representations of context, and are critical for the execution of sequential tasks involving working memory. In the next chapter, we propose a larger scale architecture involving interactions between the current model and a comparable TC-GPR based ‘motor loop’. This aims to build on the current model and establish a general architecture for performing goal-directed sequences, addressing the questions of translating a cognitive representation to an appropriate motor command, and generating appropriate and timely transitions between stages of a sequential task.

Chapter 3

An architecture for sequential performance

3.1 Introduction and aims

In the previous chapter, we used abstract PFC representations which were free of semantics to examine the ability of the GPR model to mediate the selection and maintenance of multidimensional information. Up until this point, in order to allow focus on these novel problems, modelling has been concerned solely with the associative loop and the processing of individual cognitive representations in isolation. We have not yet begun to examine;

1. how these cognitive representations are translated into the activation of motor commands;
2. how transitions between representations are mediated in order to proceed through a goal-directed sequence, and;
3. what information must be included in these representations in order to allow 1 and 2.

This chapter addresses these questions, using arguments based on the known neuroanatomy, the computational requirements of context sensitive sequencing, and existing theories of action selection. We bring together the results of these discussions by developing a general theoretical architecture for a model of multiple BGTC loops for mediating goal directed action sequences. In doing so, we identify a number of features of the proposed architecture which suggest that particular meaningful categories of information should be represented in PFC for the direction of flexible routine sequences.

3.2 Translating cognitive representations to motor actions

3.2.1 Cognitive and sensory influences

For goal-directed tasks, there are two main classes of information that the motor centres of the brain must integrate in order to perform flexible action selection; information about the world extracted primarily from sensory data ('bottom-up'), and the abstract cognitively represented intention to act ('top-down'). More generally, this top-down information may be conceived of as a component of a greater representation of temporal task context, which, as outlined earlier, we consider to comprise the necessary information for driving action selection and sequencing. Here, we examine the precise role of top-down information in facilitating action selection, and its relationship with sensory influences. Most importantly, we consider whether contextual information is directly responsible for the *selection* of actions, or the *modulation* of selection processes.

It is important to recognise that associative and motor regions of the brain may effectively represent the same information in distinct, cognitive- or motor-centred form, respectively. Thus, an action representation at the motor level may have an equivalent representation at the cognitive level. Each representation is likely to have nuanced differences; for example, motor representations of an action may include specific execution-related parameters which may not be available at the cognitive level. As an illustration, evidence suggests that online control of movements may not be cognitively guided (Glover, 2002). Thus, understanding the means by which cognitive information influences action selection may be seen as addressing the problem of translating action information held in cognitive 'co-ordinates' into the appropriate motor commands.

3.2.2 The GPR model salience

The original GPR model (Gurney et al., 2001a, 2001b) encompasses the motor territories of basal ganglia and performs signal selection on the basis of an externally generated salience projecting to striatum and STN. The extended TC-GPR model (Humphries & Gurney, 2002) retains this input scheme, and additionally sends a copy of the salience signal to cortex. While these models are focused on demonstrating the fundamental selection properties of the basal ganglia rather than the nature of the signals being processed *per se*, the salience signal is described as 'pre-processed sensory data', originating in somatosensory cortex (figure 3.1). Here, we further examine the possible functional nature of this projection so that we may understand its significance for action selection and its relationship with contextual influences.

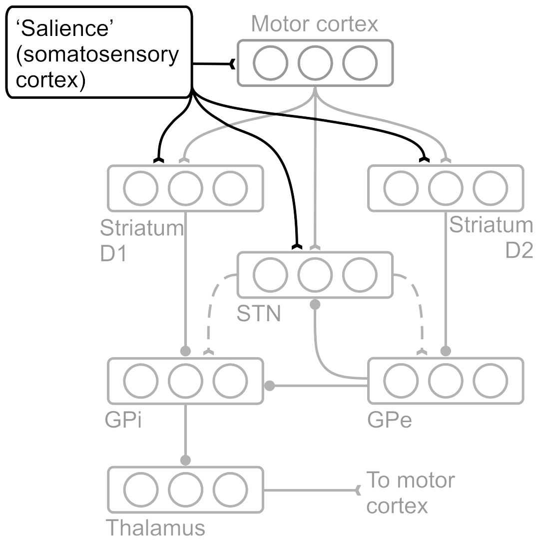

Figure 3.1: Schematic illustration of the salience input to the TC-GPR model of the motor loop in Humphries & Gurney (2002). This signal reflects processed sensory data, which is likely to specify candidate actions (see text for details).

3.2.3 Incorporating sensory influences

That the salience signal is proposed to originate in somatosensory cortex strongly implies its reflection of the bottom-up or sensory related information received by the motor loop. As the motor loop is concerned with action selection in motor space, we contend that a highly important component of this information is action specific, and relates to suitable actions given the current state of the immediate environment, which we term *candidate actions*. Further, as most actions are executed not in isolation, but directed towards particular objects in the environment, it is likely that these candidate action representations are, to an extent, object-oriented. This relates closely to the notion of action *affordances* (Gibson, 1979), or features of the immediate environment which tend to invite the performance of particular actions (see also Cisek, 2007).

Action affordances are by their very nature strongly tied to the immediate physical environment and the objects therein, relating primarily to the physical attributes of perceived objects and the particular manipulations they promote, or indeed, afford (Gibson, 1979). In order for any region encoding affordance information to accurately specify appropriate candidate actions, it is necessary that this region has access to information regarding the immediate availability of those objects. As such, it is likely that affordances are derived from information regarding visual perception and object recognition (Phillips, Humphreys, Noppeney, & Price, 2002; Riddoch, Humphreys, & Edwards, 2000). Figure 3.2 illustrates the proposed functional nature of the original salience input signal to the TC-GPR model, as action affordance information arising from the perception of objects.

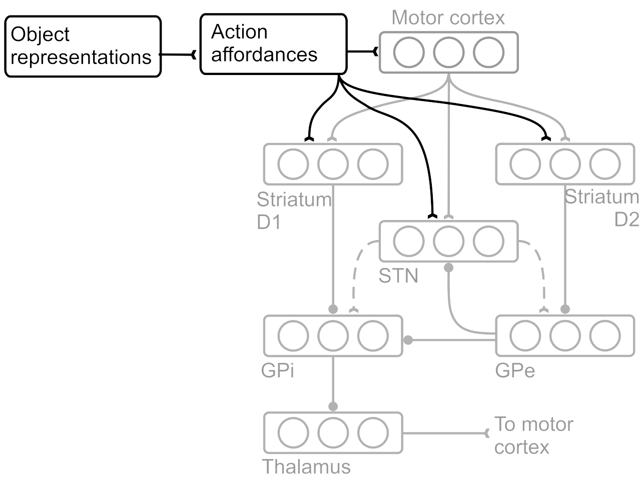

Figure 3.2: Specifying the functional nature of the sensory related external inputs to the TC-GPR. This is likely to define to object-related action affordances. Neuroanatomical support for such a pathway is discussed below.

Modelling sensory influences

We wish to retain the original TC-GPR salience projection to the motor loop in terms of its functionality, while defining its nature as 'action affordances', as discussed above. This sensory related projection to the motor loop therefore consists not of a simple description of the environment or some arbitrary input vector, but a specification of suitable candidate actions given the current state of the environment and the objects immediately available. We implement a distinct neural region in the model for this role, which is topographically organised in the same manner as the motor loop projecting in a one-to-one fashion to the channels of the motor loop.

Additionally, it is pertinent to include a region reflecting the internal representation of the currently fixated object, similar to the 'object network' employed by Cooper & Shallice (2000). This region consists of a series of nodes, each node representing a particular object involved in the task description. The activation of an object node indicates fixation of the corresponding object, and should result in the subsequent activation of any action affordances related to that object, thus specifying the object-related candidate actions for execution. Critically, multiple candidate actions may be specified by the fixation of any single object, with the result that multiple action affordances may be activated at any one time.

Neuroanatomical analogues

We suggest that regions functionally similar to our action affordance and object representation regions exist in the brain, and that similar (if far more complex) projection patterns exist between them, providing a biological, as well as a computational, imperative for the pathways we propose here.

Functions analogous to our postulated affordance specification have been proposed in parietal cortex; specifically, evidence suggests that several closely related parietal regions are loci of action related information derived from sensory sources, particularly vision (Cisek, 2007; Fagg & Arbib, 1998). Such regions are strongly involved in different types of action specification, such as saccade specification parameters in the lateral intraparietal area (LIP) (Gnadt & Mays, 1995), and reach specification parameters in the medial intraparietal area (MIP) (Batista & Andersen, 2001). Most relevant to our current model, the anterior intraparietal area (AIP) has been implicated in the specification of object-related grasping movements (Fogassi & Luppino, 2005), and may thus be a potential locus of visual-motor transformations for object manipulations (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995). Moreover, projections from AIP have been shown to impinge on motor related territories of the BGTC hierarchy, such as premotor cortex (Borra et al., 2008) and putamen (Cavada & Goldman-Rakic, 1991).

Origin of affordances

In terms of the sensory information that gives rise to object-related action affordances, this pathway should receive visual information; specifically that originating from the ventral visual stream, or the 'what' pathway (Goodale & Milner, 1992). Indeed, the terminus of the ventral visual stream and the primary locus of object recognition is well accepted as being inferotemporal cortex (IT), which in turn sends efferents to AIP (Borra et al., 2008). This pathway has been utilised in previous models as establishing grasp related parameters in response to the representation of objects (Fagg & Arbib, 1998), and evidence from neurologically impaired individuals suggests that internal representations of objects may directly activate action representations (Riddoch, Humphreys, & Price, 1989). The IT $\rightarrow$ AIP projection thus seems a likely analogue for the object-directed affordance specification pathway we have proposed here.

It is important to note that AIP also receives afferents from the dorsal visual stream (Borra et al., 2008), hypothesised to provide information for the specification of movement parameters based on the visual properties of objects regardless of their semantic associations. This pathway has been shown to be important for affordance specification, particularly for non-objects where semantic information is not available, suggesting the existence of two 'routes to action' (Phillips et al., 2002; Rumiati & Humphreys, 1998). It is likely, then, that AIP is involved in the integration of semantic and structural information for the specification

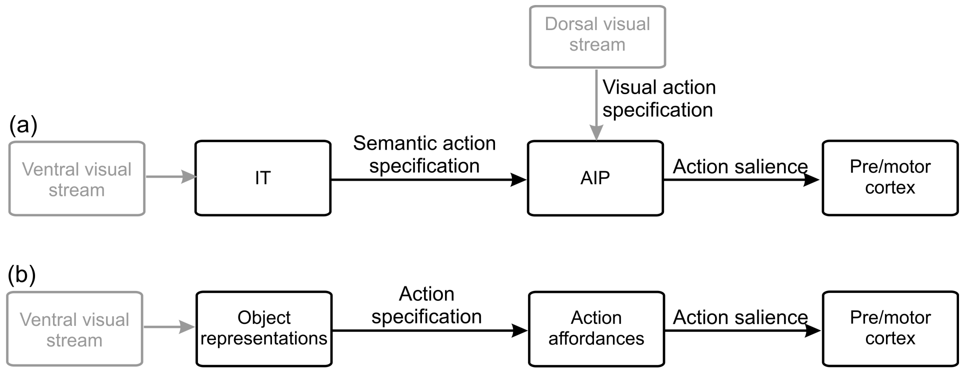

Figure 3.3: (a) Likely origins of semantic and visual specification of action affordances, based on neuroanatomical and neuropsychological evidence, discussed in the text. (b) Simplification for the present study emphasises semantic action specification based on object recognition.

of appropriate grasp movements (figure 3.3(a)). Here however, we have only incorporated regions corresponding to the former neural pathway through IT $\rightarrow$ AIP, encompassing semantic object information (figure 3.3(b)). This emphasises the role of object recognition on affordance specification and, as we expand upon in chapter 4, removes the requirement for detailed consideration of specific external parameters including spatial information in the full model.

3.2.4 Incorporating contextual influences

Having specified the likely source of the salience signal to the original TC-GPR model, this signal can be summarised as the general sensory related contribution to action selection, pertaining to the current environment and the objects therein. As discussed above, in addition to this sensory information it is vital that the motor loop also receives top-down influences on action selection, to allow action related components of internally represented temporal task context to guide action selection.

Though this cognitive influence is undoubtedly important for adaptive behaviour, it is likely that, under normal circumstances, sensory influences pose more of a direct constraint on action selection, whereas contextual information may simply bias the selection process. This is consistent with the schema focused SAS/CSS model (Norman & Shallice, 1986), as well as an examination of BGTC circuit interactions with cerebellum (Houk & Wise, 1995), and reflects the idea that sensory related information ultimately defines which actions may be performed in the current environment, particularly where those actions are object-directed. In real terms, one does not generally attempt to pick up a spoon if there is no spoon present to be picked up. However, if one's current goal does not involve picking up a spoon, there is no external constraint that will stop this action from being performed. Thus, while sensory-related or affordance contributions to the processing of the motor loop impose a direct constraint on selection by the specification of candidate actions, contextual influences support and modulate action selection by biasing the selection of a single action from the set of activated candidate actions. Similar ideas are also presented by Cisek (2007) in his 'affordance competition hypothesis':

> 'It is proposed here that the brain processes sensory information to specify, in parallel, several potential actions that are currently available. These potential actions compete against each other for further processing, while information is collected to bias this competition until a single response is selected' (p1585).

Importantly, the nature of the influence of basal ganglia on its thalamocortical targets means that such a modulatory effect may be implemented in the motor BGTC loop via direct influence to basal ganglia. As the influence of the basal ganglia on action selection is not directly excitatory but *disinhibitory*, additional excitation to motor basal ganglia, or more specifically, motor striatum, will not result in the activation of new action representations in motor cortex. Rather, such an influence can only modulate the thalamocortical activation of those actions which already have some degree of sub-threshold excitation in cortex provided by the action affordances. Figure 3.4 illustrates this theorised contextual influence to motor loop processing.

Modelling contextual influences

Based on the review of interconnected BGTC loops in chapter 1, top-down influences originate in the associative loop and consist of cognitive representations of the current task context. The mechanisms by which these representations are mediated were explored in chapter 2. The primary issue of interest here is the precise mechanism by which this information is propagated to the basal ganglia of the motor loop in order to exert its influence on

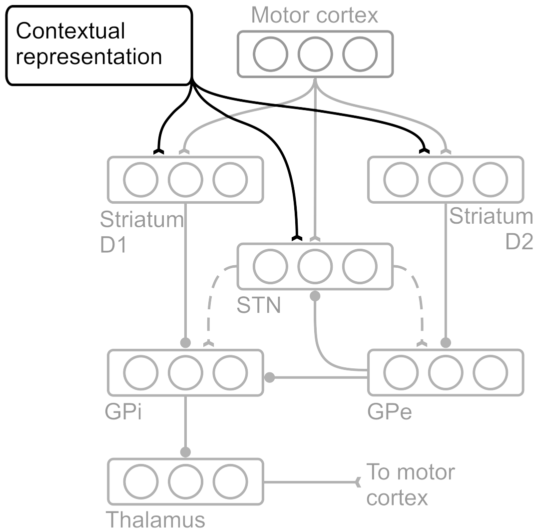

Figure 3.4: Inputs to the motor loop carrying contextual information are likely to bias, via basal ganglia, the processing of action representations which already show sub-threshold activation in the motor cortex, rather than providing direct excitation.

action selection.

As discussed in chapter 1, there are in fact several anatomical means by which regions in the associative BGTC loop may directly influence processing in motor territories. A primary focus in recent work has been striato-nigro-striatal 'spirals' (Haber et al., 2000; Joel & Weiner, 2000) and the convergence of thalamocortical projections (Haber & Mcfarland, 2001; Haber & Calzavara, 2009). However, there is also evidence to suggest partial overlap of corticostriatal projections from different functional territories. Such cross-territory corticostriatal projections have been utilised in recent computational modelling work (Doll, Jacobs, Sanfey, & Frank, 2009; Frank & Badre, 2012; Schroll et al., 2012), and have previously been proposed to convey cognitive influences to action related territories of the BGTC hierarchy (Calzavara et al., 2007; Draganski et al., 2008). Here, we adopt a corticostriatal projection from PFC to striatum of the motor loop to implement this influence. Crucially, via this pathway, the effect of contextual information on action selection is a modulatory one through the disinhibition of thalamocortical targets, rather than by direct excitation of these targets. This, as described above, allows the important constraining versus biasing influences on selection imposed by sensory and contextual information, respectively.

Figure 3.5 illustrates the general theoretical architecture resulting from the amalgamation of the proposed additions required for driving context sensitive action selection, derived from neuroanatomical and computational considerations. Note that any actions encoded by the motor loop - particularly if they relate to an object - may only be performed if a representation of that object is active and drives particular affordances for that object. In contrast to the direct excitatory effect of affordances on thalamocortical action representations, contextual influences are modulatory, enhancing or attenuating the effects of particular action affordances in motor basal ganglia, rather than directly driving motor commands in cortex.

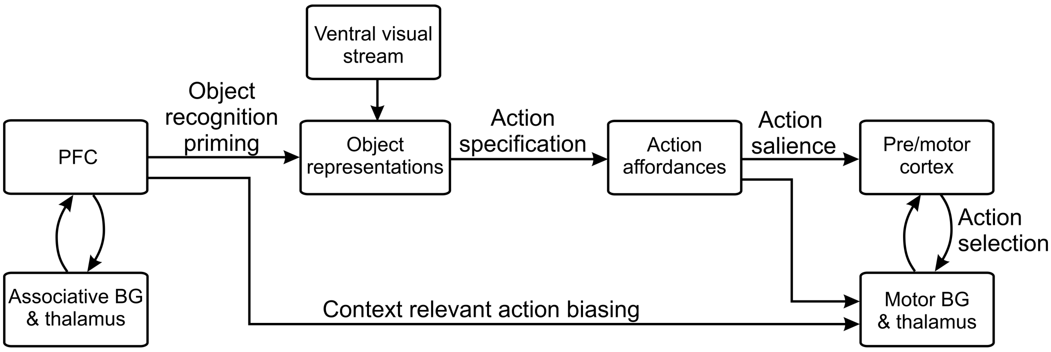

Figure 3.5: Diagram illustrating the theorised functional architecture for translating cognitive representations to motor output. See text for further details on the precise functions of pathways and postulated underlying anatomy.

Cognitive facilitation of object recognition

Note from figure 3.5 that we have included a contribution from PFC to the representations of objects; this embodies a top-down influence on object recognition (e.g., Bar, 2003). The influences of cognitive influences on visual search and perception are complex and multifaceted (Kastner & Ungerleider, 2000), and we do not claim to accurately represent them here. Though highly simplified, the projection we include here is supported by evidence from neuroanatomy showing a projection from PFC to IT (Webster, Bachevalier, & Ungerleider, 1994), and in the current model, the effects of this projection on the activity of the object representations are intended to reflect the *results* of a targeted visual search process; namely, fixation of the desired object.

3.3 Sequencing

To recap, in section 3.1 we set out three questions that we wished to address in this chapter. The general architecture presented above addresses the first of these questions (*how are cognitive representations translated into the activation of motor commands?*), but says nothing about the second, regarding the mechanisms that might mediate *transitions* between representations in order to allow the progression through a sequence. Specifically, when an action has been successfully completed, how does the system integrate this information with ongoing, higher level task-related information in order to drive a transition to an updated representation of the new task context? This is tightly bound to the third question we posed, regarding the necessary *content* of these cognitive representations. These two questions are addressed in this and the following section.

In order to retain the focus on the cognitive influences on action, we assume an external environment in which all relevant objects are immediately available for fixation. As a result, we may assume a static sensory influence to the object representations region (labelled 'ventral visual stream' in figure 3.5). Accordingly, any changes in activity in this region - reflecting a newly fixated object - will result from dynamic task-related priming influences from PFC. In turn, any changes to the object-related action affordances will ultimately result from these dynamic task-related influences on 'visual search'. Any dynamics in the motor BGTC loop in the current model are therefore the result of dynamics in the associative loop, via the influence of task-related priming on object perception and the direct corticostriatal projection to motor striatum. Conceptually, this may be summarised as a four-step process which we consider constitutes a single 'stage' of a task:

update cognitive representation → fixate new task-relevant object → specify candidate actions → select contextually appropriate action.

3.3.1 An external sequencing mechanism

As discussed extensively in chapter 2, the PFC representations are point attractors within the associative loop network, and, as such, are intrinsically stable and self-sustaining. Once the network comprising the associative loop settles into one of these attractors - thus expressing a particular PFC representation - some additional 'energy' must be supplied to the system in order to allow dynamics within the activation of the network. Given the intrinstic stability of the system, this energy must necessarily originate from an external source. As such, in order to generate transitions between distinct PFC representations - and, consequently, sequential behaviour - some external mechanism is required to supply this additional energy to the system.

Previous models of sequential behaviour have relied on intrinsic dynamics within recurrent networks to generate sequencing (e.g., Botvinick & Plaut, 2004). However, in these models, individual patterns of activation across the network, analogous to our PFC representations, do not represent stable attractors. Rather, connectivity within the network is important for producing the correct *trajectory* through state space, but not for stabilising individual representations. Rather, in such models, a new, discrete pattern of activation of the network is reached on every simulation timestep; no one pattern is intrinsically stable over multiple timesteps. A trade-off thus exists between the requirements for self-maintenance of activation patterns and intrinsic dynamics within a recurrent network. This may be considered an activation-based variant of the 'stability-plasticity dilemma' (Abraham & Robins, 2005), which relates to the difficulty in implementing synaptic plasticity in a network ('dynamics') and simultaneously retaining information within existing patterns of synaptic weights (stability). However, this is rarely discussed with direct reference to activation patterns in neural networks, due to the preferential utilisation of fixed point attractors to study retrieval of *single* memories in isolation (P. Miller, Brody, Romo, & Wang, 2003), and attractors with transient dynamics to study sequential processing (Botvinick & Plaut, 2004). Here, however, we require both stability and sequencing, whereby a sequence of individually stable PFC representations must be evoked. This trade-off between stability and dynamics has been addressed in the computational literature, and the requirement for an external sequencing mechanism has been noted (Rutishauser & Douglas, 2009).

Requirements for sequencing

The requirements and precise nature of a sequencing mechanism depend to an extent on task demands. For the purposes of composing flexible routine action sequences, any influence driving sequencing should take into account not only the current external environment or the single prior action but also the current temporal task context in order to avoid the problems associated with 'chaining' (Henson, 1996; Lashley, 1951), which were outlined in chapter 1. Temporal task context may incorporate various types of information depending on the particular task, but may include, for instance, the overall goal. Imagine, for example, two goal-directed tasks A and B, *make tea* and *make coffee*, the examples used earlier to illustrate the hierarchical organisation of action (figure 1.1). These tasks may share an initial subtask S_1, say *boil kettle*. However, their subsequent subtasks may differ, giving S_2a and S_2b, for example, *add teabag to cup* and *add grinds to cup*, respectively. Let us also imagine that after completion of the common subtask S_1, the external environment is identical in both cases, as is the immediate performance history. At this point then, it is entirely ambiguous from this information whether S_2a or S_2b should be performed next. It is only the internal representation of the overall task, A or B, that determines the correct course of action. In this instance then, while the representation of the overall task may have no effect on the performance of the *current* action or current subtask S_1, it acts as a critical contextual 'disambiguator', determining the correct *next* action to be performed.

Implementation

A mechanism for sequencing is thus required which is both external to the associative loop - such that the selected PFC representations therein remain stable - and which integrates information from the current external environment and the current temporal task context. In order to implement a solution which successfully achieves both these requirements, we propose a sequencing mechanism embodied by an additional region that we term *transition nodes*, similar to that utilised by Rutishauser & Douglas (2009) in their own examination of driving context sensitive sequencing of stable states.

In this solution, each stable PFC representation R_i that is available for selection by the associative loop has a corresponding transition node which, when active, provides excitation to all PFC nodes comprising representation R_i. Assuming representation R_i is sufficiently excited for a sufficient duration, R_i is subsequently selected, and any previously selected representation is deselected or inactivated through inhibitory and competitive selection mechanisms within the associative BGTC loop. Afferents to the transition nodes from PFC convey information pertaining to the current temporal task context. This tonic influence from PFC 'primes' - or activates to a sub-threshold level - any transition nodes which correspond to suitable subsequent PFC representations. Note that multiple transition nodes may be primed at any one time, in those circumstances where more than one action may feasibly follow the current one. When an action is selected by the motor loop, the external environment changes as a result of this action. Information reflecting this environmental change is propagated to the transition nodes as a phasic influence. This temporary additional activation is integrated in the transition nodes with the current tonic influence from PFC. The combination of influences from these two sources preferentially excites a single transition node, which in turn activates the PFC representation which is most compatible with both internal and external influences, thereby updating the PFC representation of task context according to both task history (from PFC) and current environmental state. Owing to the phasic nature of the external influence to the transition nodes, the activity therein is also transient, thus requiring stability within the associative loop in order to maintain the newly selected PFC representation. This addition to the model is illustrated schematically in figure 3.6.

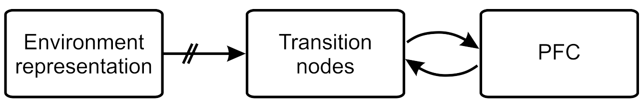

Figure 3.6: Schematic illustration of the postulated sequencing mechanism, 'transition nodes.' These receive a tonic influence from the currently expressed contextual representation in PFC and a phasic influence denoting a change in the external environment. This impinges upon PFC resulting in suitable context-sensitive sequencing behaviour.

3.4 Functional architecture of contextual representations

3.4.1 Components of representations

The general architecture presented above introduces three pathways originating in PFC, influencing the object representations, activity in motor basal ganglia, and priming transition nodes in order to trigger appropriate updating of PFC representations. These pathways are effectively specialised for carrying information related to the current object of interest, current intended action, and current contextual information relevant to the appropriate next action, respectively. Accordingly, the proposed architecture suggests that these particular categories of information must be represented in PFC, such that the relevant information may be extracted by the three pathways. As discussed in chapter 2, these categories of information may be represented in PFC in many different ways. However, we continue to consider a basic feature-based or 'semi-localist' structural form which, as we have shown, affords suitable selection and maintenance within the associative BGTC loop architecture.

Given the model architecture we have proposed, the necessary components of PFC representations for mediating the selection of component actions are the *object(s)* required for the desired action, as well as the intended *action* itself. These must be extracted individually by the two pathways we introduced in section 3.2, and integrated by the motor loop, via parietal affordance areas, in order to produce the correct action. It is important to note that these pathways - and the motor loop in general - need not receive any more detailed contextual information than this; these minimal features are sufficient for the selection of the correct current action. However, given that more detailed contextual information *is* important for appropriately concatenating single actions into sequences, any information which affects the temporal order of action execution must also be included in the overall PFC representation. This information does not need to impinge upon the pathways between the BGTC loops, but must be conveyed to the transition nodes introduced in section 3.3, which integrate this information with that from a dynamic environment to generate appropriate sequencing. Note that the intended action and object are important features of the overall context, but that they are unlikely to be sufficient to drive sequencing, except perhaps in very simple tasks in which only one temporal order is acceptable. In addition to these features, the transition nodes must also have access to any detailed contextual information which is required in order to drive the correct sequencing of the actions. As mentioned earlier, this is likely to vary from task to task, but examples might include the current *task*, perhaps the current *subtask* if a task is so divided, and a record of previous actions performed as part of the sequence.

3.4.2 Relationship with BG

Following from chapter 2, we propose a subset-based selection scheme, whereby the associative basal ganglia selects a subset of the total PFC representations. Nodes from multiple representations are supported by a single channel in basal ganglia, and recurrent inhibition in PFC refines and completes the selection process. Given the feature-based encoding scheme utilised in PFC, it is reasonable to suppose that subset selection takes place on the basis of a particular feature; that is, each 'channel' in associative basal ganglia encodes a particular value or instance of a particular feature category. Indeed, it is likely that associative regions of basal ganglia are particularly sensitive to some relatively high level task information or features (Kermadi & Joseph, 1995) that may facilitate the timely and efficient selection of PFC representations. This idea is explored in more detail with relation to specific tasks in the next chapter.

3.5 Summary

In the current chapter, we have argued from both a biological and computational standpoint for the inclusion of particular neural regions and pathways, and particular functional mechanisms which address three particular questions, initially set out in the general introduction and reiterated in section 3.1. Figure 3.7 illustrates the full theoretical architecture resulting from the additions we propose above.

In the following two chapters, we implement a computational model based on this theoretical architecture, and use it to address the specific and well researched tasks of tea- and coffee-making. In doing so, we are able to examine the model's ability to account for spe-

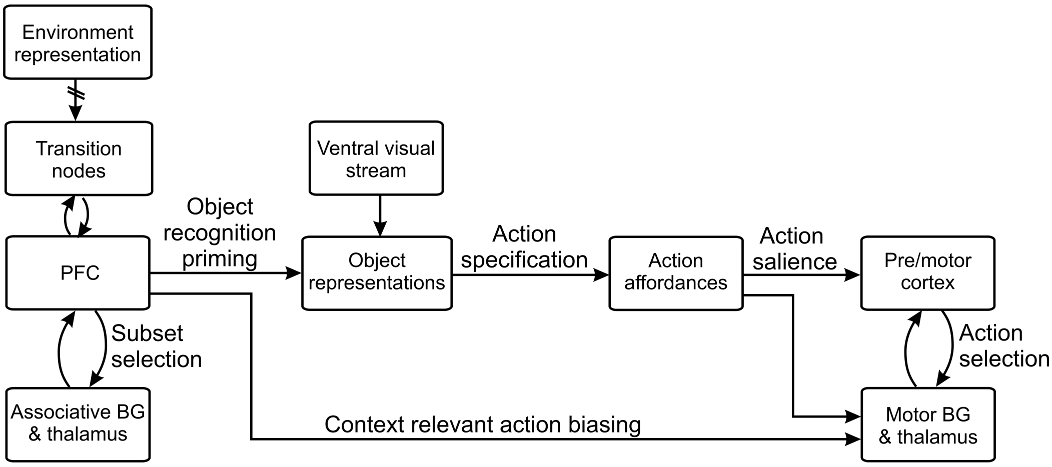

Figure 3.7: Diagram illustrating the full theoretical architecture resulting from the additions presented in the text, incorporating mechanisms for translating cognition to action on the basis of action affordances, context sensitive sequencing, and semantic sensitivity in associative basal ganglia.

cific behavioural data, as well as bringing a new, neuroanatomically focused perspective to an existing debate on the nature of the mediation of such sequences (Cooper & Shallice, 2000; Botvinick & Plaut, 2004).

Chapter 4

Sequential routine action selection in multiple BGTC loops

4.1 Introduction and aims

In the previous chapter, we established a general theoretical architecture for the mediation of multiple, goal directed sequential tasks, based on the known BGTC-loop hierarchy and a number of computational constraints. In this chapter, we assess the performance of two such sequential tasks in a computational model of this system, in order to demonstrate how this theoretically sound architecture mediates context sensitive sequential performance.

4.1.1 Previous modelling work

In examining the performance of the model on these tasks, we hope to build on an existing debate in the modelling literature, discussed briefly in chapter 1, on the organisation of the cognitive structures required for sequential performance (Botvinick & Plaut, 2004, 2006b, 2006c; Cooper & Shallice, 2000, 2006a, 2006b). This debate centred around two competing computational models of routine action that we introduced in the general introduction, and which proffered two distinct accounts of how this cognitive information is structured (Botvinick & Plaut, 2004; Cooper & Shallice, 2000). To recap, the first of these models, known as the *interactive activation network* (IAN) model, consisted of an abstract, hierarchical schema network, where schemas and goals were represented in a localist fashion at multiple levels of the network (Cooper & Shallice, 2000). The other utilised a *simple recurrent network* (SRN) model harnessing distributed, learned representations of context to guide action selection (Botvinick & Plaut, 2004). Both models focused on the task of coffee-making, as this type of task has routinely been used in order to assess the impact of frontal lobe injury on the performance of sequential tasks (Schwartz et al., 1991, 1995; Humphreys & Forde, 1998). The SRN model also performed a related tea-making task. Both models were consistently able to perform the assigned tasks in the absence of noise or disruption, but relied on apparently quite distinct mechanisms in order to do so.

In this chapter, we aim to show that by creating a biologically plausible model to mediate the performance of these sequential tea- and coffee-making tasks based on the known neuroanatomy of the BGTC-loop system, and computational constraints discussed in chapters 2 and 3, the resulting architecture naturally embodies key functional features of both of these models, and many of the issues debated by the two sets of authors approach a reconciliation.

4.1.2 Outline of the chapter

We begin by providing a detailed description of the tea- and coffee-making tasks to be performed by the model, followed by a precise specification of the contextual representations developed to support task performance. These take the same functional form of those introduced in chapters 2 and 3, but include additional complexity to account for the details of the tasks. We go on to examine the model's performance in a variety of simulations. We discuss the results of these simulations with regard to the underlying dynamics in cognitive regions of the model that support action selection, and their significance for understanding the organisation of cognitive information for mediating sequential action.

4.2 Task design

Beyond allowing useful direct comparison with the models of Cooper & Shallice (2000) and Botvinick & Plaut (2004), and bringing a new, neuroanatomically focused interpretation of the underlying functional mechanisms they discuss, our reasons for utilising the tea- and coffee-making tasks have an additional mechanistic basis. The tasks share many subgoals and actions, whilst satisfying distinct goals. A degree of similarity between the tasks is desirable, as it allows us to examine the ability of the model to complete distinct tasks which, for example, require the performance of the same action; this will require a dependence on internal representations of contextual information, placing particular demands on the model's ability to capture such information in a timely and efficient manner. Additionally, as detailed in previous work (Botvinick & Plaut, 2004), while the order in which particular subsequences are performed is often flexible, the order of the actions within those subsequences is rarely so. Employing the tea- and coffee-making tasks allows us to examine the ability of the model to mediate sequences with this type of flexibility, as several of the involved subsequences do not require execution in a particular temporal order.

The tea- and coffee-making tasks thus present a specific set of problems for the model. Importantly, many of these have caused difficulties for previous modelling work, such as reusing a single action within the same sequence, and utilising the same action in the context of different tasks (e.g., Rumelhart & Norman 1982; see also Botvinick & Plaut, 2004, and Dominey, 1995, for discussions). These tasks therefore allow us to directly compare the model's performance with previous modelling studies.

4.2.1 Details of the tasks

Hierarchical organisation

As described in the general introduction (see figure 1.1), it is well accepted that action sequences tend to comprise structural hierarchies, and evidence suggests that healthy participants conceive of sequential tasks in just such a manner (Humphreys & Forde, 1998). In order to reflect this general structure, we adopted a **task** → **subtask** → **action** organisational hierarchy, which is also consistent with the previous modelling work with which we are directly concerned (Cooper & Shallice, 2000; Botvinick & Plaut, 2004). As such, both our tea- and coffee-making tasks were composed of distinct subtasks, each of which was in turn completed by the performance of two or three distinct composite actions. The tea-making task consisted of three subtasks, and may be summarised as follows:

1. **add teabag to cup**: pick up teabag → put into cup;
2. **add water to cup**: pick up kettle → pour into cup;
3. **add sugar to cup**: pick up spoon → scoop sugar → pour into cup,

where subtasks are in bold, and the rightward arrow means 'is followed by' in a successful execution. The order in which the subtasks must be completed in the tea-making task was fixed as listed above. Conversely, the coffee making task consisted of four subtasks:

1. **add coffee to cup**: pick up spoon → scoop coffee → pour into cup;
2. **add water to cup**: pick up kettle → pour into cup;
3. **add milk to cup**: pick up milk → pour into cup;
4. **add sugar to cup**: pick up spoon → scoop sugar → pour into cup.

For the coffee task, however, the final two subtasks could performed in either order, such that,

add coffee to cup → add water to cup → add sugar to cup → add milk to cup,

was also a valid sequence with which to achieve the goal of coffee-making. It is important to note that the order in which the **add milk** and **add sugar** subtasks may be performed is *not* specified by the goal, which is simply **make coffee**. The hierarchical structure of the three valid sequences in terms of the task, subtasks, and actions is summarised in table 4.1.

Note that two of the four subtasks (**add water** and **add sugar**) were required for both tea- and coffee-making tasks, requiring the model to select the associated actions in multiple contexts. Importantly, these two subtasks must be preceded and followed by different subtasks in each case, so may not be regarded as embedded components of a greater single subroutine.

4.2.2 Representations

As discussed in the previous chapter, we conceive of a 'stage' of a task as being a four step process terminated by the selection of an action by the motor loop. For example, the

a	Make tea						
	add teabag		add water		add sugar		
	pick up teabag	put into cup	pick up kettle	pour into cup	pick up spoon	scoop sugar	pour into cup

b	Make coffee (i)									
	add coffee			add water		add sugar			add milk	
	pick up spoon	scoop grounds	pour into cup	pick up kettle	pour into cup	pick up spoon	scoop sugar	pour into cup	pick up milk	pour into cup

c	Make coffee (ii)									
	add coffee			add water		add milk		add sugar		
	pick up spoon	scoop grounds	pour into cup	pick up kettle	pour into cup	pick up milk	pour into cup	pick up spoon	scoop sugar	pour into cup

Table 4.1: Summary of valid sequences and their hierarchical organisation in terms of their composite subtasks and actions.

first stage of the tea-making task is initiated by the presentation to the model of the corresponding goal or instruction to **make tea**. This is followed by the selection of an appropriate cognitive representation of the current temporal task context, which initiates a saccade to the correct object and subsequent activation of appropriate action affordances. Finally, selection of an action by the motor loop terminates the current stage of the task and initiates the next. Consistent with the model of the associative loop outlined in chapter 2, each such stage of each task must be represented uniquely by PFC. This is so that the model may produce the correct action according to the current context, and so that actions may be performed in the correct temporal order, as discussed at length in chapter 3. Again, we hand-crafted the representations to be selected and maintained by the associative loop, allowing us to maintain full control of the information contained in the representations, in turn allowing us to better interpret the behaviour of the model in terms of the dynamics of the underlying cognitive representations. While this approach lacks the benefits of the 'discovery' of appropriate representations as shown by Botvinick & Plaut (2004), it has the potential to allow considerably deeper insights about the necessary components of cortical representations for the mediation of flexible sequential performance (see, for example, section 4.5.4).

As we have stressed in earlier chapters, and given the stereotyped nature of these representations, it is useful to conceive of these representations as schemas; more specifically, as high level schemas or skill units composed of lower level schemas which themselves are represented as individual features within the overall representation (Cooper & Shallice, 2000; see also section 1.2.2). This has important implications for understanding the behaviour of the model, as described in detail in section 4.5.1.

Included features

Chapter 3 introduced the idea that distinct categories of information or *features* should be represented in PFC in order to allow the appropriate selection of the current intended action, and to allow the correct temporal sequencing of actions. Together, the representation of each relevant feature constitutes the overall temporal task context. Within this representation of context, the current intended *action* and corresponding *object* should be represented explicitly in order to drive selection of the current action. However, a more detailed representation of task context is required to ensure the correct sequencing of actions. The specific details required for this function, as discussed in chapter 3, are likely to vary from task to task. In the present study however, three additional features are critical for correct sequencing of actions. Firstly, representations of the current *goal* and *subtask* are necessary. Given that the same actions may be performed in the context of different tasks, and as a component of distinct subtasks (see table 4.1), these features are necessary to disambiguate the forthcoming action.

For example, the action **pick up spoon** is required for both **add sugar** and **add coffee** subtasks. Without an internal representation of the current subtask, no information is available to determine the correct *next* action: either **scoop sugar** or **scoop grounds**. Equally, a representation of the current goal is necessary to determine whether the **add milk** subtask should be performed. Thus, these features are required not for the immediate selection of the appropriate action, but for appropriate *sequencing*.

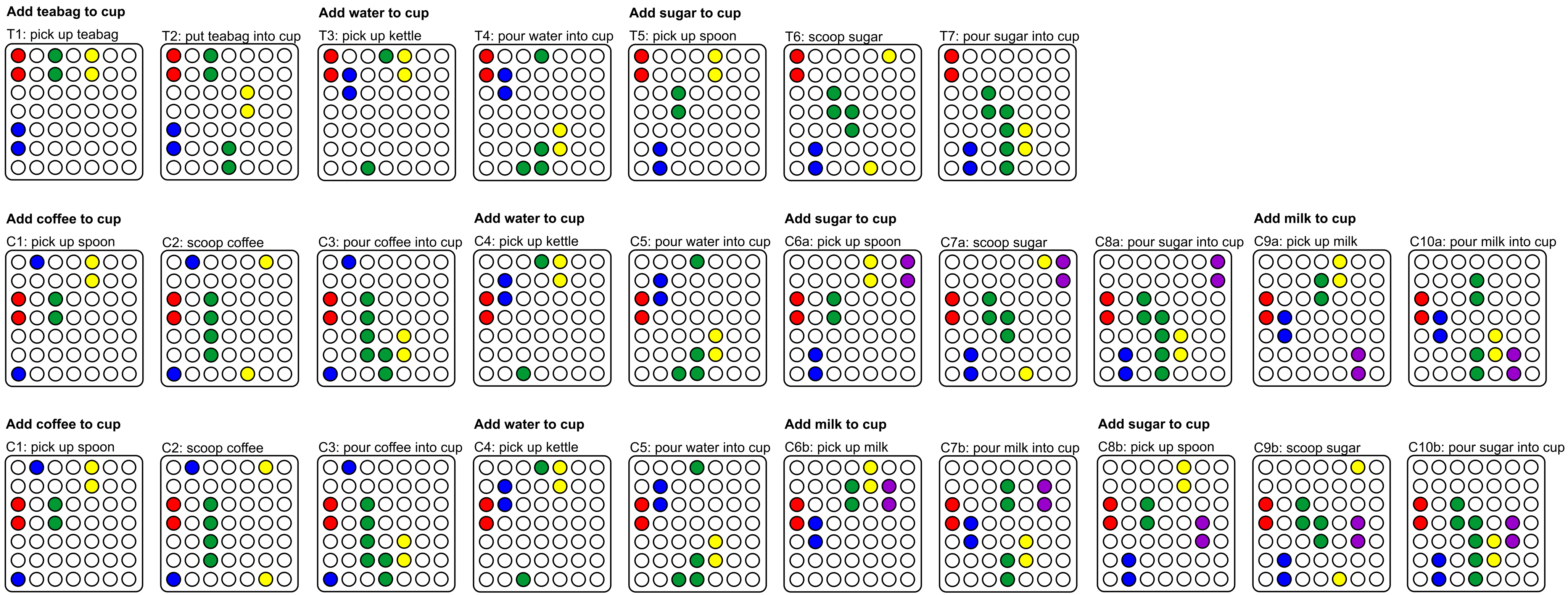

Figure 4.1: PFC representations for each stage of the tea-making task (T1-7) and both versions of the coffee-making task (C1-C10a & C1-C10b). Subtasks are denoted in bold text. Coloured nodes are those that are active for that representation of temporal task context. Nodes representing the current goal are illustrated in red; those representing the current subtask in blue; the objects required to complete the current stage of the task in green; and the corresponding actions in yellow. 'Rank order' nodes are illustrated in purple. Note that each feature is represented consistently across each pattern in which it appears. See text for further details.

Additionally, where subtasks may be performed in a flexible order (namely, the **add milk** and **add sugar** subtasks in the coffee-making task), we include a representation of the *rank order* of these subtasks. This indicates the order in which these subtasks are performed, and again is necessary for ensuring appropriate transitions between selected PFC representations. For example, if the **add sugar** subtask is completed prior to the **add milk** subtask, the model must keep in memory that the sugar has been added so that the subtask is not repeated. The representation of rank order makes this possible. Note, however, no such representation of rank order is necessary for the 'inflexible' subtasks of adding coffee, tea and water (and, for the tea-making task, sugar); as there is only one legal order in which they may be executed, the previous actions that have been performed are *implicit* in the subsequent representations. Hence, rank order is not represented where it is not necessary.

The inclusion of features representing these elements of context notably results in separate representations for each instantiation of an action or subtask even though the actions performed are identical; for example, three slightly different representations of the stages of the **add sugar** subtask are required, for encoding the differences between the tea-making and both versions of the coffee-making task. While this may seem to compromise efficiency, it is to be regarded as a strength, as it captures both the important similarities between different invocations, as well as contextual differences which, as discussed earlier, are key for ensuring the correct subsequent action. Importantly, these differences are represented minimally, with features which are consistent across different versions of a subtask being represented consistently each time. Moreover, this requires no additions or modifications to the basal ganglia mechanisms mediating the representations. This provides an effective resolution to Botvinick & Plaut's (2004) complaint that schemas are unable to capture these differences in an efficient manner; by encoding schemas at multiple levels of the task hierarchy as a single representation or high level schema, we are able to account for contextual similarities whilst maintaining a representation of the hierarchical organisation of the task.

The representations we used to encode each stage of each task are illustrated in figure 4.1. Note that each individual feature is represented consistently, whereby particular nodes encode particular features across all contextual representations. For instance, the top two leftmost nodes (illustrated in red in figure 4.1) encode the goal **make tea**; these remain tonically active throughout the course of the tea-making task, assuming no errors in representation occur. Notably, this scheme results in a greater degree of overlap between those representations which share a greater number of features. Such overlap is likely to underlie certain behavioural phenomena, such as action slips (Botvinick & Plaut, 2004). It is interesting to note that this gradient of representational similarity is an inevitable consequence of the feature-based coding scheme we have adopted here, but not necessarily of alternatives such as fully distributed representations.

4.2.3 Simplifications and omissions

Importantly, we have simplified the modelled tasks to a greater extent than has been done in previous modelling studies, predominantly by reducing the number of actions required for each of the tasks, and by omitting certain classes of action. The overall reduction in the number of actions does not affect the validity of the current study, as we are concerned primarily with the fundamental principles of sequencing and the underlying mechanisms that mediate them, which apply to sequences of arbitrary length. However, this reduction does allow us to retain a manageable framework; more extended sequences would necessarily require the inclusion of additional architecture with no real theoretical benefit. Additionally, the reduction in overall length does not result in a simplification of the inherent *structure* of the tasks, which, as emphasised, retain the **task** → **subtask** → **action** hierarchy explored in previous studies.

In addition to this overall reduction, we specifically omit two sub-classes of action from our current study: orienting actions, e.g. **fixate cup**, and release actions, e.g. **put down spoon**. There are multiple reason for these omissions. Firstly, this contributes to the reduction of the overall length of each task, allowing us to avoid unnecessary complexity, and focus on the *crux* and facilitatory actions within each subtask, where by 'crux', we mean the single composite action which achieves the goal of the current subtask (Schwartz et al., 1991). Secondly, given the relatively abstract level of description with which we are currently focused, there is no *qualitative* difference between orienting or release actions and the manipulative actions we include. Again, despite the minimal gain from including these actions, doing so would have required vast architectural additions to the model. However, the omission of these actions themselves requires us to include some other means of accounting for their effects.

Accounting for orienting objects

In the previous chapter, we introduced a region representing object-related action affordances. This region is based on parietal areas, most notably AIP, which in the brain receives afferents from the dorsal and ventral (via inferotemporal cortex; IT) visual streams (Borra et al., 2008). These afferents to AIP provide visual information including precise spatial parameters, and semantic information based on object recognition, related to the currently fixated object or region of the environment. This information is utilised by AIP to define suitable grasping actions given the current state of the environment. Representations of specified suitable actions are subsequently propagated to motor regions of the BGTC hierarchy, interpreted in the present study as action *salience* (figure 4.2(a)). Orienting actions are thus important for providing up-to-date visual and semantic information for the specification of object-targeted actions.

In the current model we have omitted orienting actions, but accounted for their effects on action specification as follows. As discussed in chapter 3, we currently include only a semantic component to action specification originating in the ‘object representations’ region, based on IT. This object representations region receives a top-down projection from PFC, which reflects a task-related priming influence on visual search and object recognition. Processing in the ventral visual stream is subsumed by a ‘gating’ influence on activity in the objection representations region which we term a *fixation gate*. This influence simply reflects the presence of an object in the environment, and its consequent availability for fixation. If this gate is ‘closed’, this indicates that an object is not currently present, and thus cannot be fixated (see figure 4.2(b)).

Via the projection from PFC, selection of a particular contextual representation in PFC causes downstream activation of the corresponding object representation. This process thus

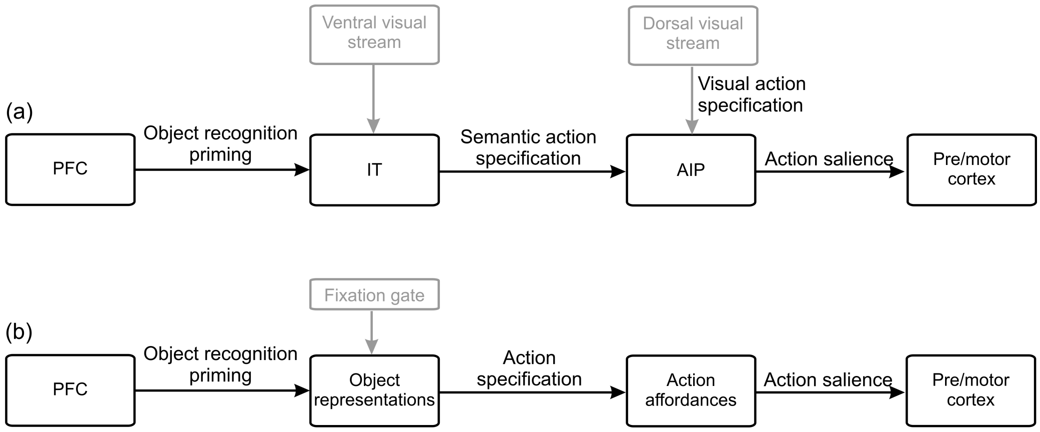

Figure 4.2: (a) Influences to the AIP originate in both ventral and dorsal visual streams carrying information related to semantic and visual components of the fixated object or region of space. (b) Simplifications implemented in the current model. Activity in the 'object representations' region indicates the currently fixated object. The fixated object defines action affordances according to semantic information, binding any action selected in motor cortex to that object. Influences from PFC reflect task-related priming influences on object recognition, whereas the 'fixation gate' subsumes ventral visual stream processes, preventing activation ('fixation') of an object which is not currently present in the environment. Note that dorsal ventral stream influences are not modelled in the present study. See text for more details.

subsumes task-directed saccades to the to-be-acted-upon object; affordances subsequently received by the motor loop relate *only* to the currently fixated object. This process embodies two assumptions. Firstly, that each action selected by the motor loop is preceded by a saccade to the to-be-acted-upon object. Secondly, that the motor loop is concerned *only with actions on the fixated object*. This constitutes a 'deictic' scheme, 'whereby the body's pointing movements bind objects in the world to cognitive programs' (Ballard, Hayhoe, Pook & Rao, 1997; p726), and similar schemes were utilised by both Cooper & Shallice (2000) and Botvinick & Plaut (2004). Note that these assumptions effectively result in a spatially nonspecific model, and additionally remove the necessity to incorporate representations of allocentric or egocentric space.

Accounting for release actions

Finally, release actions may be considered as secondary to the actions we do include, in that they must only be performed so that the next action may be initiated. They in themselves are not generally requisite actions for the completion of the task (though an exception to this is the **add teabag** subtask, in which **put down** is an integral part of the final composite action,

put into cup). Further, such actions are required at the end of every subtask. While this does not render them conceptually defunct *per se*, it adds little theoretical value given the existing final action common to all other subtasks, **pour into cup**. As such, we concluded the benefits of clarity brought by the exclusion of these release actions outweighed those of including them. Rather than modelling these actions separately, then, we included the assumption that any **pick up** action was preceded by a **put down** of the currently held object.

4.3 Full model architecture - general description

4.3.1 Amendments to associative loop

'Subtask selection'

The general form of the associative loop remains faithful to the architecture developed and explained in chapter 2. However, as the representations are no longer abstract but semantically anchored to particular tasks, the role of basal ganglia at the associative loop level must be clarified with respect to the current tasks.

In chapter 2, we introduced the concept of 'subset' based selection in associative basal ganglia, where each channel in basal ganglia supported multiple representations in PFC; recurrent inhibition within PFC was utilised to resolve competition between those representations concurrently supported by a single channel. We retain this general scheme here, whereby each channel in basal ganglia selects the subset of representations which share a common *subtask*. For instance, representations T3, T4, C4 and C5 (see figure 4.1) all correspond to actions within the **add water** subtask. As such, all four representations are supported by a single channel in basal ganglia (see also figure 4.3). Again, we rely on intrinsic recurrent inhibition within PFC to resolve any remaining ambiguity.

The reasons for this organisation are twofold. Firstly, given the **task** → **subtask** → **action** behavioural hierarchy, from a computational perspective encoding subtask at this level allows minimal switching between basal ganglia channels. Such switches are likely to be more costly than those only within PFC; this scheme minimises the number of channel-switches within a single task, thus reducing computational demands on the associative loop. Moreover, there is some evidence to suggest that at this level of the neuroanatomical hierarchy, basal ganglia deals with information at this level of description. For instance, Fujii & Graybiel (2003) found that activity in prefrontal neurons peaked at the beginning and end of learned sequences, suggesting a sensitivity to 'chunked' information in associative regions of the BGTC hierarchy. Additionally, Kermadi & Joseph (1995) showed many sequence-selective cells in monkey caudate in a saccade and reach task. These sequences were presented successively and in different orders on each trial, and as such might arguably be classed as 'sub-sequences', again reflecting a sensitivity to information at an intermediate level of description in the action hierarchy. Thus, encoding subtask at this level of basal ganglia is both computationally and theoretically appealing.

Here then, we include by necessity five distinct channels in basal ganglia, each of which represents one of the five subtasks utilised in the two tasks. Selection of a particular channel in basal ganglia disinhibits the corresponding subtask channel in thalamus, providing excitation to *all* nodes which are active for *any* of the representations corresponding to that subtask. Figure 4.3 illustrates the total PFC nodes which receive support from each subtask channel in thalamus.

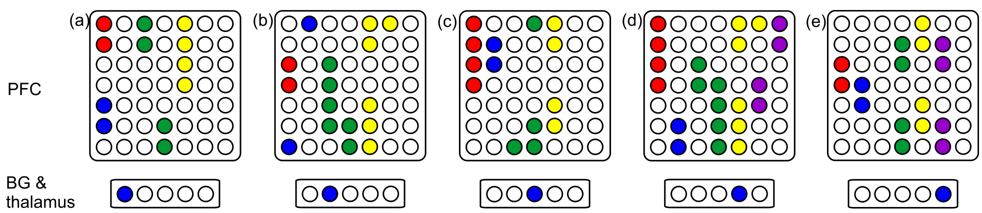

Figure 4.3: Detail of the PFC nodes which receive excitation from each of the five 'subtask channels' in thalamus, and their corresponding channels in basal ganglia and thalamus. (a) **add teabag**; (b) **add coffee**; (c) **add water**; (d) **add sugar**; (e) **add milk**. Note that as some nodes are involved in multiple representations (e.g., top left goal nodes), they receive excitation from multiple channels in thalamus. Comparison with figure 4.1 illustrates the individual representations for which components of these sets of nodes are active. Again, sensitivity of nodes to particular features is indicated by colour: red = goal; blue = subtask; yellow = action; green = object; purple = rank order.

Comparison with figure 4.1 shows that each basal ganglia channel supports all nodes involved in encoding each stage of each subtask, regardless of the overall task in which they appear, though the context in which the subtask is selected is disambiguated by the overall PFC representation. Again, this satisfies Botvinick & Plaut's (2004) requirement that representations or schemas should capture key similarities, as well as differences, between the same action or subtask performed in different contexts.

4.3.2 Incorporating the motor loop

As outlined in chapter 3, we now incorporate a motor BGTC loop to the model to perform action selection itself. We have proposed that the representations held by PFC have a direct biasing influence on action selection in the motor BGTC loop via a corticostriatal projection. In the current model, the motor BGTC loop follows an architectural scheme closely based on the original TC-GPR model (Humphries & Gurney, 2002), where a typical 'channel-wise' organisation is maintained throughout the loop, including cortex, standing in contrast to the associative loop, with its subset-based selection scheme. There are four actions represented by the motor loop: **pick up**, **put into**, **pour into** and **scoop**. Thus, the motor loop consists of four distinct channels, each representing one of these actions. The selection of a channel is interpreted as the initiation of the corresponding action, where selection is defined by the crossing of a selection threshold by motor cortex output (Hanes & Schall, 1996).

Generalisable action representations

Note that the actions represented in the motor loop are *object nonspecific*, in that they may be performed with or upon different objects. This increases efficiency by allowing re-use of a single action channel across different objects. The issue of resolving object related ambiguity which results from these object nonspecific actions relies on the representation of the currently fixated object. As discussed above in section 4.2.3, the motor loop performs selection based on affordance information pertaining only to the currently fixated object, and is thus concerned with action selection only at the point of fixation. This binds actions to a particular object, disambiguating the selection of the object nonspecific action itself (Ballard, Hayhoe, Pook, & Rao, 1997).

4.4 Formal model description

A schematic depiction of the full model architecture and main processes is shown in figure 4.4. Notation describing the internal connections of the associative loop is detailed in chapter 2, and may also be found in list form in appendix A. A formalisation of the additions to the model is given below.

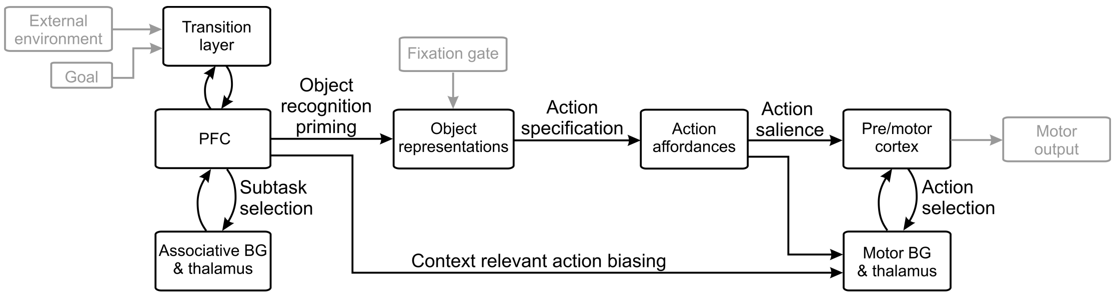

Figure 4.4: Schematic diagram of full model architecture. See text below and chapter 3 for a detailed discussion of the precise structure, function and neuroanatomical basis of each region and projection.

4.4.1 Associative loop

We introduce new notation for the new additions to the associative loop described in general terms in chapter 3. Primarily, this notation relates to the transition layer and its afferent and efferent connectivity. Otherwise, all structure and related notation remains the same. Consistent with chapter 2, unless notation indicates otherwise, connectivity between nuclei is channel-wise. Notation is summarised in Appendix A.

Transition layer

The transition layer, Γ, consisted of $N^{\Gamma} = 22$ nodes, where N^{Γ} is the total number of hard-coded, stable representations required for each of the tasks (corresponding to those illustrated in figure 4.1). Consistent with all other nuclei, each of these nodes consisted of a single leaky integrator neuron, as defined by equations 2.1-2.3.

Each transition node received three contributions to its total activation; a signal indicating the current goal, projections from PFC indicating the current cognitive representation of task context, and a representation of the external environment indicating the current state of the objects therein.

Goal signal

The goal signal ϕ consisted of a vector of $N^{\phi} = 2$ elements, each corresponding to the strength of a particular goal; here, **make tea** and **make coffee**. This contribution to the input of transition node Γ_j may be written as:

$$u_j^{\phi\Gamma} = \sum_{i=1}^{N^{\phi}} W_{ij}^{\phi}\phi_i, \tag{4.1}$$

where W_{ij}^{ϕ} defines the sensitivity of transition node j to goal i.

PFC

PFC projected to the transition layer via the weight matrix $W^{x\Gamma}$. PFC influences this region on the basis of individual nodes, rather than by the subset-based relationship discussed above with basal ganglia and thalamus. If y_i^x is the output of PFC node i,

$$u_j^{x\Gamma} = \sum_{i=1}^{N} W_{ij}^{x\Gamma} y_i^x \tag{4.2}$$

describes the influence of PFC on each transition node Γ_j.

Environment

A simple representation of the visible external environment was encoded by a binary matrix ξ, where each element represented a possible state of each object involved in the tasks (see table 4.2). Values of 1 indicated the status of each object, and values of elements in ξ were updated upon the selection of an action by the motor loop. For instance, after successful execution of the **scoop sugar** action, elements in ξ representing **spoon:held** and **sugar:in-spoon** would show a value of 1. The overall pattern of activity indicated the current status of all objects.

After an update to the environment matrix, ξ imposed a *transient* influence on transition

	held	on-table	in-spoon	in-packet	in-kettle	in-bottle	in-bowl	in-cup
teabag:	0	1	0	0	0	0	0	0
spoon:	0	1	0	0	0	0	0	0
packet:	0	1	0	0	0	0	0	0
kettle:	0	1	0	0	0	0	0	0
bottle:	0	1	0	0	0	0	0	0
bowl:	0	1	0	0	0	0	0	0
cup:	0	1	0	0	0	0	0	0
grounds:	0	0	0	1	0	0	0	0
clear-liquid:	0	0	0	0	1	0	0	0
milk:	0	0	0	0	0	1	0	0
sugar:	0	0	0	0	0	0	1	0
black-liquid:	0	0	0	0	0	0	0	0
brown-liquid:	0	0	0	0	0	0	0	0

Table 4.2: Table illustrating the environment matrix ξ. Each cell in the table represents a corresponding element in ξ. The value displayed in each cell indicates the value of the corresponding element at the initiation of each trial. This matrix denotes the visible status of relevant objects for the task, providing processed sensory information to the transition nodes in order to account for external influences on sequencing. As actions are selected by the motor loop, the values of elements are changed accordingly. For example, after selection of the action **pick up spoon**, the element representing **spoon:held** is set to 1, and the element representing **spoon:on-table** set to 0. Note the inclusion of *black-liquid* and *brown-liquid*; these account for the visual representation of tea and coffee once made, without and with milk, respectively. Thus, once the teabag or grounds *and* water have been added to the cup, the element representing **black-liquid:in-cup** would be set to 1.

nodes via the weight matrix W^{ξ}. This weight matrix was hard-coded such that environmental states preferentially excited transition nodes which corresponded to suitable contextual representations. For instance, environmental states including **spoon:held** and **sugar:in-spoon** would preferentially excite transition nodes corresponding to PFC representations T7, C8a and C10b (see figure 4.1). Note that release actions (‘**put down**’) have not been explicitly modelled (see section 4.2.3). We implemented the assumption that each **pick up** action is preceded by a **put down** of any currently held object, and additionally that each **put into** action effectively encompasses a **put down** by a corresponding update to matrix ξ whenever a **pick up** or **put into** action was selected.

Let ξ_i be the value of element i in ξ, the influence to transition node Γ_j may then be described by,

$$u_j^{\xi\Gamma} = \xi_{on} \sum_{i=1}^{N^{\xi}} W_{ij}^{\xi} \xi_i \tag{4.3}$$

where $\xi_{on} = 1$ for a set number of timesteps after a change to the environment, otherwise

$\xi_{on} = 0$, reflecting the transient influence of ξ on the transition nodes.

Inhibitory recurrence

A mutual inhibitory influence was included in the transition layer to minimise ambiguity resulting from simultaneous activation of multiple nodes. This may be written as

$$u_j^{\Gamma\rho} = -W^{\Gamma} \sum_{i=1, i \neq j}^{N^{\Gamma}} y_i^{\Gamma}, \tag{4.4}$$

where W^{Γ} is the strength of this lateral inhibition and y_i^{Γ} is the output of transition node i.

Summary

The total contribution to the activation of the transition nodes may then be summarised by,

$$u^{\Gamma} = u^{\phi\Gamma} + u^{x\Gamma} + u^{\xi\Gamma} + u^{\Gamma\rho}. \tag{4.5}$$

New influences on PFC

In this amended architecture, rather than an abstract salience signal, PFC received external influences from the transition layer. Equation 2.7 is thus amended, whereby the term representing external salience is removed and replaced by a new term defining the influence from the transition layer. This may be written as,

$$u_j^{\Gamma x} = \sum_{i=1}^{N^{\Gamma}} W_{ij}^{\Gamma x} y_i^{\Gamma}, \tag{4.6}$$

where $W_{ij}^{\Gamma x}$ is the synaptic strength from transition node i to PFC node j, and y_i^{Γ} the output of transition node i. Equation 2.7 defining the total input to each PFC node is thus replaced by,

$$u^{x} = u^{\rho} + u^{Vx} + u^{\Gamma x}, \tag{4.7}$$

where u^{ρ} and u^{Vx} are the contributions from intrinsic PFC recurrence and thalamus, respectively (see equations 2.4 and 2.5).

Other amendments

The remainder of the associative loop was subject to only minor changes, including an increase in the dimensions of the recurrent network representing PFC (now 7x7 nodes), and an increase in the number of channels comprising basal ganglia (now 5). Additionally, the specific weight matrices delineating the projection patterns between PFC and basal ganglia differed in order to support the representations illustrated in figure 4.1, though the principles of PFC subset based connectivity remained constant. Weight matrices delineating the new projection scheme are detailed in appendix C.

4.4.2 Motor loop

The motor loop retained the typical channel wise organisation of the original GPR model. Consistent with this previous work, each channel of each modelled region was represented by a single leaky integrator neuron, as defined by equations 2.1-2.3.

Pre/motor cortex

Cortex of the motor loop retained the traditional channel wise organisation of the TC-GPR model, which was maintained throughout the loop. This region received a direct projection from motor regions of thalamus and from the action affordance region outlined in chapter 3 and formalised below. Action affordances and thalamus both projected to pre/motor cortex in a one-to-one fashion. If $W^{\nu M}$ is the synaptic strength from motor regions of thalamus, and $W^{\Lambda M}$ that from the affordance region, the inputs to pre/motor cortex channel may be described by

$$u^M = W^{\Lambda M} y^\Lambda + W^{\nu M} y^\nu \tag{4.8}$$

where y^ν is the output of (motor) thalamus and y^Λ that from the action affordances region.

Putamen

Putamen is generally accepted to comprise motor regions of striatum (Alexander et al., 1986). This region is organised into selection and control pathways (Gurney et al., 2001a), incorporating D1- and D2- expressing neurons, as described in chapter 2 for the associative loop. Consistent with the nature of inputs to the motor loop in the original TC-GPR model (Humphries & Gurney, 2002), putamen received excitatory projections from pre/motor cortex and directly from action affordances. Additionally, associative loop influences on the motor loop were mediated by a corticostriatal projection from PFC to putamen. As our PFC representations are feature-based, and the motor loop is concerned with only one of those encoded features - actions - only those PFC nodes encoding actions projected to putamen. More specifically, each channel in putamen received excitatory influences from only those PFC nodes encoding the *corresponding* action. This is consistent with evidence showing a relatively minimal cross-territory corticostriatal projection from associative to motor regions (Calzavara et al., 2007).

Let the set of nodes in PFC representing action i be denoted A_i, and let the associated index set be J_i, such that $J_i = \{k \in \mathbb{Z} : x_k \in A_i\}$, where x_k is the k^{th} PFC node. If W^{xp} is the synaptic strength of PFC afferents to putamen, the contribution from PFC to putamen channel i may then be written as,

$$Y_i^{xp} = W^{xp} \sum_{k \in J_i} y_k^x. \tag{4.9}$$

Inputs to putamen are once again modulated by a factor representing tonic dopamine levels, λ_M. If W^M is the synaptic strength of the projection from pre/motor cortex and $W^{\Lambda p}$ that from the action affordances, the total modulated input to D1- and D2- expressing regions may, respectively, be written as

$$u^s = (1 + \lambda_M)(Y^{xp} + W^M y^M + W^{\Lambda p} y^{\Lambda}) \tag{4.10}$$

$$u^c = (1 - \lambda_M)(Y^{xp} + W^M y^M + W^{\Lambda p} y^{\Lambda}). \tag{4.11}$$

where y^M is the output of pre/motor cortex and y^{Λ} the output of the affordances region.

Dorsolateral STN

Dorsolateral (dl) regions of STN are regarded as comprising its motor related extent (Parent & Hazrati, 1995b; Joel & Weiner, 1997). dlSTN received influences from the same sources as putamen, consistent with the orginal GPR model. Thus, STN received a projection from both PFC and pre/motor cortex. Unlike putamen however, the projection to dlSTN from all nodes in PFC was uniform. This ensured a suitable balance of activity in the selection and control pathways, but remained uniform across channels in order to retain focus on the role of the interloop corticostriatal projection for selection. dlSTN also received inhibitory projections from motor regions of GPe. If W^{Md} is the synaptic strength of the pre/motor cortical projection, W^{xd} the strength of the projection from each PFC node, and W^{gd} that of the GPe → dlSTN pathway, the total input to dlSTN may be written as

$$u^d = W^{xd} \sum_{i=1}^{N} y_i^x + W^{Md} y^M - W^{gd} y^g. \tag{4.12}$$

Ventrolateral GPe

Motor regions of GPe are regarded as occupying its ventrolateral (vl) extent (Parent & Hazrati, 1995a; Haber, 2003). As in the associative loop, vlGPe receives projections from its corresponding STN and D2- expressing regions of striatum. dlSTN projections are again diffuse and may be written as a sum of outputs across all channels:

$$Y^d = \sum_{i=1}^{n^M} y_i^d \tag{4.13}$$

where y_i^d is the output of each dlSTN channel i and n^M is the number of channels in the motor loop. If W^{cg} is the synaptic strength from D2- regions of putamen and W^{dg} that from dlSTN, inputs to vlGPe may be expressed as

$$u^g = w^{dg} Y^d + W^{cg} y^c, \tag{4.14}$$

where y^c is the output of putamen (D2).

Ventrolateral GPi

Like GPe, ventrolateral regions of GPi are regarded as its motor territory (Sidibé et al., 1997; Haber, 2003). Inputs to vlGPi follow the same scheme as the original GPR and the associative loop described above. Three inputs are integrated here: inhibitory influences from D1- expressing regions of putamen and vlGPe, and a diffuse excitatory projection from dlSTN according to equation 4.13. If W^{sh} is the synaptic strength from putamen, W^{dh} that from dlSTN, and W^{gh} is the strength of the vlGPe-vlGPi pathway, the total input to vlGPi may be expressed as

$$u^h = W^{dh}Y^d - W^{sh}y^s - W^{gh}y^g. \tag{4.15}$$

where y^s and y^g are the outputs from D1 expressing regions of putamen and vlGPe, respectively.

Ventrolateral nucleus of thalamus

Ventrolateral thalamus (VL) is regarded as its primary motor region (Haber & Calzavara, 2009). Inputs to VL follow the standard architecture of the GPR model, consisting of an excitatory projection from pre/motor cortex and an inhibitory influence from vlGPi. Let W^{Mv} be the synaptic strength of the projection from pre/motor cortex, and W^{hv} that from vlGPi, then

$$u^v = W^{Mv}y^M - W^{hv}y^h \tag{4.16}$$

describes the total input to VL, where y^M and y^h are the outputs from pre/motor cortex and vlGPi, respectively.

4.4.3 Intermediate regions

Object representations

This region was introduced in chapter 3. Its activation reflected an internal representation of the currently fixated object, and, by virtue of defining the action affordances received by the motor loop, the object to be acted upon. Each object was represented by a single leaky integrator neuron, and each received three sources of input: a task-directed priming influence from PFC, a lateral inhibitory influence, and a modulatory influence by a fixation 'gate'. This gating influence reflected the presence of objects in the environment, and took a value of 0 or 1. A value of 0 indicated an object was not currently present in the environment, and thus could not be fixated. If $W^{x\Theta}$ is the weight matrix representing the synaptic strength of PFC afferents to the object representation layer, Θ, then inputs to object representations from PFC may be described by,

$$u_j^{\Theta x} = \sum_{i=1}^{N} W_{ij}^{x\Theta} y_i^x. \tag{4.17}$$

where y_i^x the output from PFC node i. If W^{Θ} is the strength of lateral inhibition in the object representation region, inhibitory recurrence may be described by,

$$u_j^{\Theta\rho} = -W^{\Theta} \sum_{i=1, i\neq j}^{N^{\Theta}} y_i^{\Theta}, \tag{4.18}$$

where y_i^{Θ} is the output of object node i. The total inputs to the object representation region may therefore be summarised by,

$$u^{\Theta} = f\left(u^{\Theta x} + u^{\Theta\rho}\right), \tag{4.19}$$

where f denotes the value of the fixation gate.

Action affordances

Action affordances, as discussed in chapter 3, reflect the specification of suitable candidate actions based on the currently fixated object. Their total afferent input is received from the object representations region, via the synaptic weight matrix W^{Λ}, and is defined by

$$u_i^{\Lambda} = \sum_{j=1}^{N^{\Theta}} W_{ji}^{\Lambda} y_j^{\Theta}, \tag{4.20}$$

where y_j^{Θ} is the output from object representation node j.

4.4.4 Parameters

The synaptic weight matrices, $W^{x\Gamma}$, W^{ξ}, $W^{\Gamma x}$, and $W^{x\Theta}$ are detailed in appendix C. Amendments to the weight matrices W^{ρ} and W^{V} are also detailed in this appendix. W^{Γ} was set to 0.5, W^{Θ} to 1, W^{xd} to 0.01, W^{xp}, W^{X}, $W^{\Lambda M}$ and $W^{\Lambda p}$ to 0.3, W^{XV} to 0.15 and W^{M} to 0.2. W^{ϕ} was set to 1 between each goal and the first transition node of the corresponding sequence; otherwise W^{ϕ} was 0. In order to focus on the role of interloop corticostriatal projections in disambiguating affordance information, W^{Λ} was set to 0.3 for all object-affordance projections in the present simulations. All other associative loop parameters were consistent with those outlined in chapter 2, and motor loop weights consistent with those detailed in Humphries & Gurney (2002), and are also detailed in appendix C. The output gradient m_{Γ} for the transition layer Γ was set to 10, and the threshold ϵ_{Γ} set to 0.5. This resulted in rapid activation and deactivation of transition nodes upon the initiation and cessation of transient influences from the environment representation ξ. Values of m and ϵ for the object representations and action affordances were set to 1 and 0 respectively. Output parameters for the motor loop were faithful to the TC-GPR model (Humphries & Gurney, 2002), as were all time constants used.

Selection of an action was defined by pre/motor cortical activity traversing the selection threshold $\theta = 0.9$. After an action was selected by the motor loop, environment representation ξ was updated and allowed to influence the transition layer for 200 simulated timesteps, resulting in phasic input to PFC from the transition layer, thus imposing the requirement for self-maintenance in PFC. Noise, drawn from a Gaussian distribution with mean 0 and variance 0.001 was added to the activation of transition nodes for symmetry breaking at neutral choice points.

4.5 Results

4.5.1 Sequential selection

Initially we tested the model's ability to perform both tea and coffee tasks, and examined its spontaneous performance rates of milk-first and sugar-first versions of the coffee trial.

Input to the model consisted of activation of the goal signal, ϕ. We ran 200 trials for both tea- and coffee-making tasks.

The model was able to consistently perform all three sequences without error, and milk-first and sugar-first versions of the coffee task were completed with roughly equal frequency. This demonstrates the ability of the functional architecture we have proposed to integrate several types of information in order to mediate goal-directed action sequences. It is notable, given the difficulties previous models have had in mediating such complex sequences (see Botvinick & Plaut, 2004), that component actions may be recruited in a flexible order, independently of the action just performed, independently of the ultimate goal, and multiple times in a single sequence. The model is also able to maintain a record of the performance of previous actions (adding sugar) which do not result in a visible change to the environment.

There are complementary roles of associative and motor loops in this process. The associative BGTC loop is responsible for forming representations of task context and mediating their temporal order via the transition layer. The transition layer integrates information regarding the current internal contextual representation and changes to the external environment, and subsequently supplies activation to the appropriate representation in PFC. On the basis of this activation and in concert with basal ganglia, PFC selects and maintains representations of the current temporal task context in working memory. These may be maintained for an arbitrary duration, until new changes to the environment trigger further transition node activity.

The motor loop is primarily responsible for action execution. This loop integrates activation from two sources in order to perform suitable action selection: a cognitive representation of the required action, originating in PFC, and affordance information specifying multiple candidate actions for selection, compatible with the currently fixated object. Cognitive influences from PFC provide a task related disambiguation of these affordances, ultimately resulting in contextually appropriate action selection.

Example activity: motor loop

Example outputs from the object representations region and pre/motor cortex for a coffee-making trial are illustrated in figure 4.5, below, shown as simulation timepoints at which the output of each node or channel exceeded the selection threshold, θ. Sequential selection is observable as the consecutive activation beyond this threshold by the outputs of each action channel in pre/motor cortex, indicated by the timepoints illustrated below. Action selection is driven in part by activation of object representations, which is possible via the specification of action affordances (see figure 4.4). Note that this causative influence is reflected in an offset in the dynamics of the two regions, where activation of object representations precedes that of cortical channels (figure 4.5).

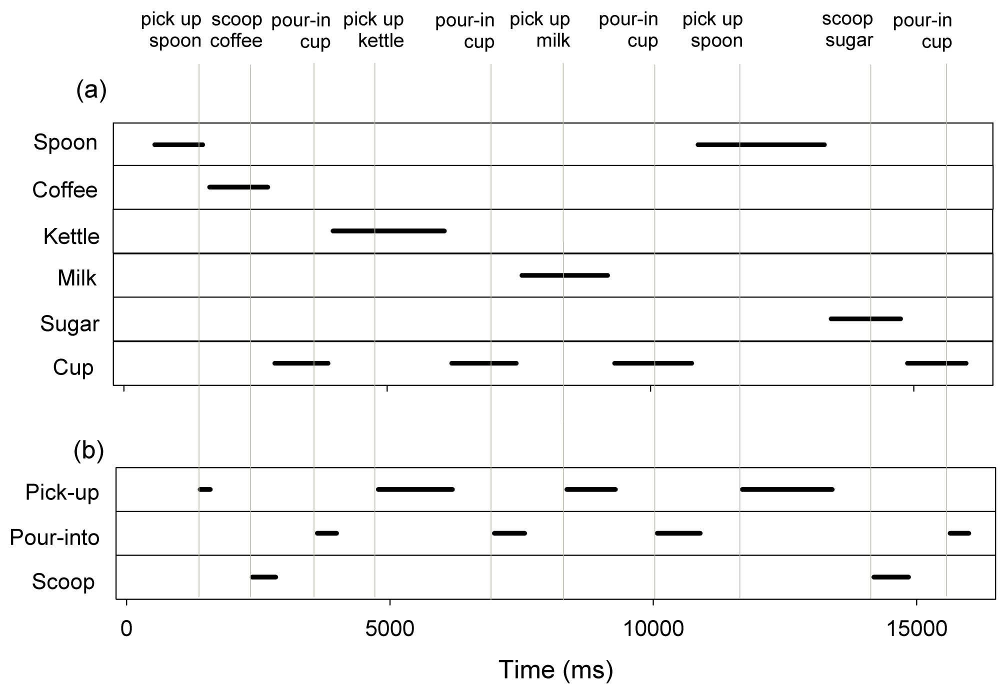

Figure 4.5: Time course of above-threshold activation of key nuclei for a typical coffee-making trial. (a) Charts show timepoints at which the output of each object node in the object representations region exceeded $\theta = 0.9$, indicating fixation of that object. (b) Charts show timepoints at which the output of each channel in pre/motor cortex exceeded $\theta = 0.9$, indicating selection of the corresponding action. Selection times are indicated by the grey vertical lines. Note that fixation of the relevant object occurs in advance of action selection. Durations of each 'selection' were drawn randomly from the range 800-2500ms and demonstrated the ability of the associative loop to maintain PFC representations for arbitrary durations until action selection was complete.

PFC dynamics

The time course of activation of the PFC nodes encoding the goal, subtasks and actions for the coffee-making task are illustrated in figure 4.6. It is this activity which drives action selection in the motor loop, via task-directed influences on visual search resulting in activation of the object representations region, and direct contextual biases on action selection via the corticostriatal projection from PFC to putamen. Note the unique time courses of PFC nodes encoding information at different levels of the task hierarchy, demonstrating the sensitivity of our feature based representations to different aspects of the task which vary at different timescales.

In section 4.2.2, we indicated that our stable representations may be conceived of as schemas. The functional equivalence of different features of these representations and schemas at different levels of the IAN schema network is clear from comparison of the dynamics of PFC nodes (figure 4.6) and the schema activation profile from Cooper and Shallice's (2000) model of coffee making within the contention scheduling system, reproduced in figure 4.7. Selection and deselection of feature nodes, visible in figure 4.6(a), is caused by transition node activation triggering a switch in the overall selected PFC representation. The time course of transition node activation is shown in figure 4.6(b); comparison with 4.6(a) demonstrates maintenance of a selected representation occurs in the absence of transition node activity.

Note that both our PFC output profile in figure 4.6(a) and that of the schema network in Cooper and Shallice's model (figure 4.7) show sensitivity to task features on different timescales. The primary functional difference between the models, however, is the top-down flow of activation in the IAN model from higher to lower level schemas, whereas in our model, all features, akin to low level schemas, are mutually supportive in order to create stable 'high level' schema-like representations in PFC. However, if we consider the functional relationship of our two BGTC loops rather than that of the various features in our PFC representations, the PFC nodes encoding actions may be considered cognitive representations of action schemas; our channels in motor loop, conversely, may be considered motor

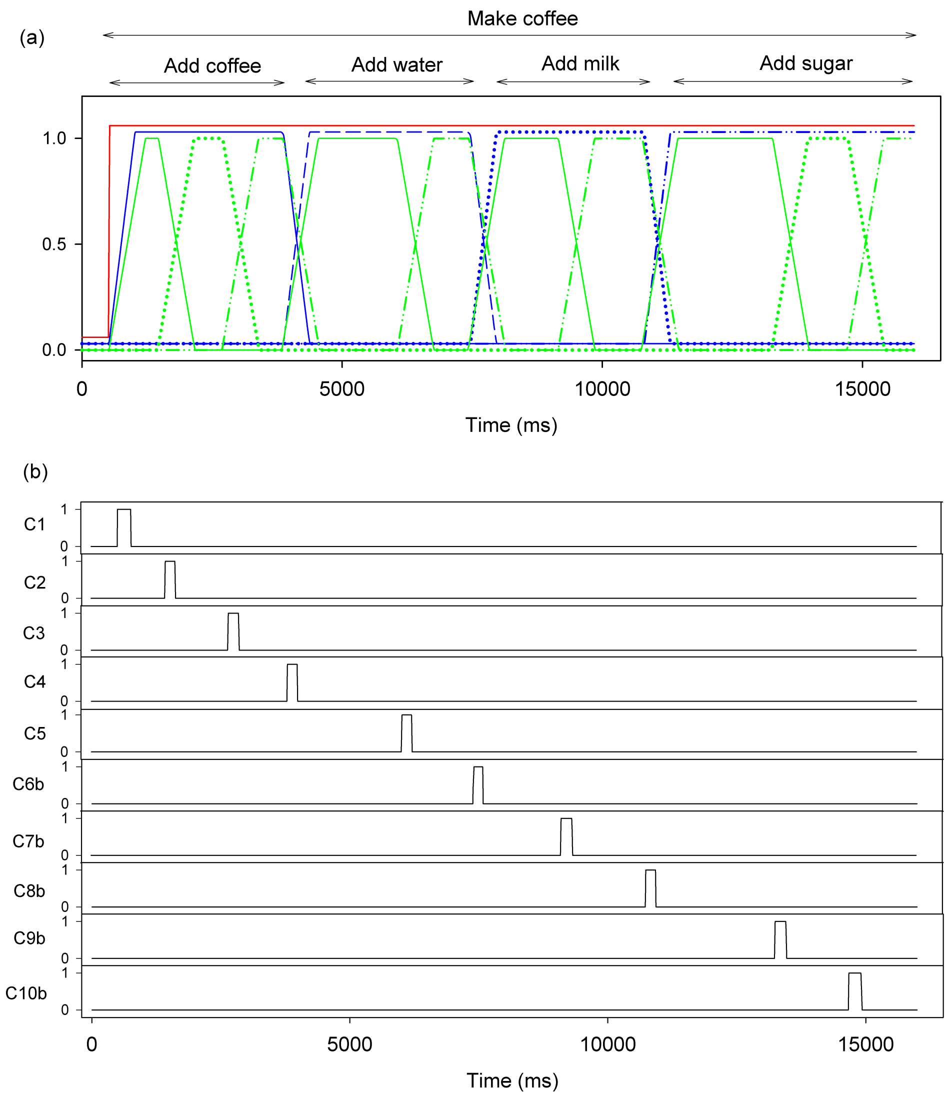

Figure 4.6: (a) Time course of activity of PFC nodes encoding goal (red), subtasks (blue), and actions (green). For purposes of visibility in the current plot, traces have been slightly offset (output peaked at 1 for all nodes), and moving averages are shown for subtasks and actions using window sizes of 500 and 700ms, respectively (in reality, outputs were almost binary in nature, with more rapid selection and deselection of features than shown here). (b) Time course of transition node output, for transition nodes C1-C10b, which provided excitation to the corresponding coffee-making PFC representations C1-C10b (see figure 4.1). Note transience of transition node output, and maintenance of PFC node activation during transition node quiescence.

representations of the *same* schemas. The activation of motor representations of action by cognitive ones may be considered an analogue of the flow of activation from higher to lower level schemas in the schema network.

Figure 4.7: Time course of schema activation during the coffee-making task in the IAN schema network model. Reproduced from Cooper & Shallice (2000), with permission.

Intriguingly, the dynamics of our PFC network may also be seen to be functionally equivalent to those of Botvinick and Plaut's (2004) SRN hidden layer, which itself encodes temporal task context. Beyond being said to encode the same information however, we are able to visualise this equivalence using the same means they adopt for the analysis of the dynamics of their hidden layer using multidimensional scaling (MDS) analysis. MDS is a form of principal components analysis, which extracts from a multidimensional dataset (here, the 49-dimensional PFC representation at each timepoint) two abstract dimensions which preserve the majority of the information encoded by the data at each point. The subsequent two-dimensional dataset may then be easily visualised. Inspired by Botvinick & Plaut's (2004) discussion of the dynamics of their hidden layer, we performed an MDS analysis of the output of our PFC nodes over the course of two coffee trials. In the first, the **add milk** subtask was performed before the **add sugar** subtask, in the other these subtasks were performed in the reverse order.

Results of the MDS analysis of PFC representations during each of the subtasks are displayed in figure 4.8. These graphs show the trajectory of PFC activation in the extracted two-dimensional state space. That our PFC representations were hard-coded allows us to conclude that contextual differences provided by the **rank order** nodes result in slightly different trajectories for each valid sequence; Botvinick & Plaut (2004) were unable to make such an accurate semantic interpretation of their equivalent analysis. However, these results

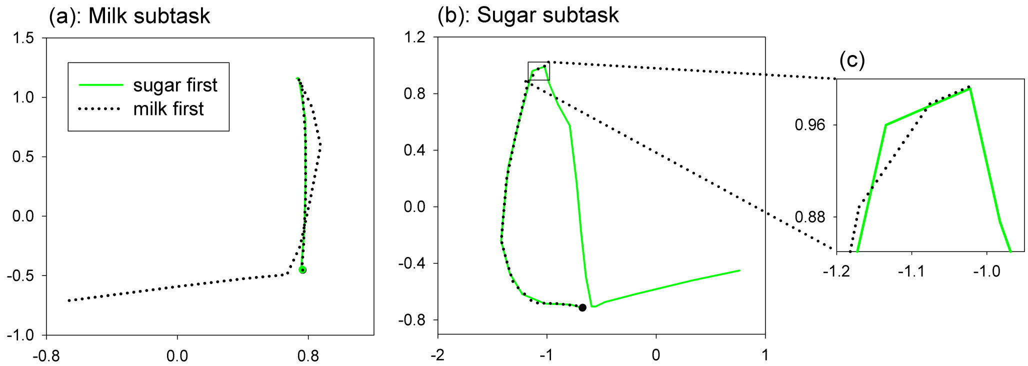

Figure 4.8: MDS analysis results show the trajectory of PFC activity through the extracted two-dimensional state space during (a) the **add milk** subtask and (b) the **add sugar** subtask, for milk-first (black dotted line) and sugar-first (green solid line) versions of the coffee-making task. Start points of each trajectory are indicated by a filled circle. Trajectories are highly similar for significant portions of the subtasks for each version, reflecting significant overlap in representations encoding each version. However, slight differences resulting from the activity of different rank order nodes is visible in (c). Divergence at the end of the subtasks reflects the initiation of transitions to different subsequent representations.

demonstrate a functional equivalence of the representations encoded by the activity of our PFC nodes and by the hidden layer of the SRN adopted by Botvinick & Plaut (2004), in terms of the encapsulation of key similarities and differences between variations of a task. It is important to note that this was observed despite the more explicit degree of localism in our representations, which was of a similar nature to that employed by Cooper & Shallice (2000).

4.5.2 Lesion studies

We have demonstrated above the ability of the model to produce the correct sequences. However, we wished to confirm that the associative loop, in its updated, more complex instantiation, was still performing the functional roles we initially set out for it in the general introduction, and that the success of the model was not a result of an implementational nuance. Additionally, we wanted to confirm that the motor loop was indeed sensitive to both affordance information and cognitive influences from PFC, and was not performing selection based on only one of these influences; if it were, a more complex version of the model with more objects and actions might be unable to perform to the high standard we have observed in the present study.

Associative loop

Pallido-thalamic and thalamo-cortical projections

We re-ran simulations 2.7.8(i) and (ii) from chapter 2, lesioning the associative GPi-thalamus projection and the thalamocortical projection, respectively. To assess the precise effects of the lesions on PFC functionality, we compared the activity of PFC nodes during the first four stages of a successful coffee-making trial, with the equivalent activity after lesioning. A 'stage' is terminated by the selection of an action by the motor loop; thus in a successful coffee trial, the first four stages of the task consist of the performance of the **add coffee** subtask, and the first action of the **add water** subtask.

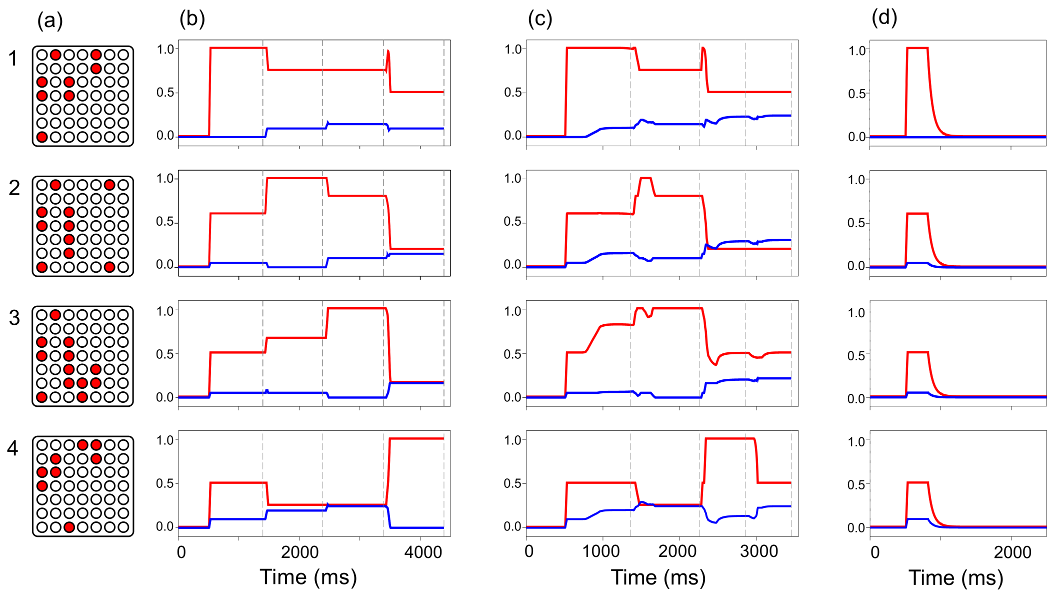

Figure 4.9: Output of PFC representations C1-C4 (illustrated in column (a)) for (b) healthy performance of the first four stages of the coffee task; (c) with a lesion to the pallido-thalamic projection; and (d) a lesion to the thalamocortical projection. Red traces show the mean output of the representations illustrated in A; blue traces show the mean output of all other PFC nodes. Grey dashed lines indicate times of selection of an action in motor loop. Both lesion types result in impaired performance, indicating the necessity for both types of projection in appropriate selection and maintenance of representations.

Specifically, we examined the activity of the four representations corresponding to these first four stages of the task. These representations, labelled C1-C4 in figure 4.1, are reproduced for clarity in column (a) of figure 4.9. We took the average output of the nodes encoding these four representations, over the first four stages of the task. This resulted in four activity profiles, showing the activation level of each representation. These profiles are illustrated by the red traces in the column (b) of figure 4.9. For each representation we examined, we also took the average output for all other nodes in PFC (unfilled nodes in column (a)); the resulting profiles are shown in blue.

As would be expected, the representation encoding the first stage of the task is maximally active during this stage; the blue trace shows that all other nodes in PFC are quiescent during this stage. The representation encoding the second stage is maximally active during the second stage, and so on. Column (b) thus shows the expected profile of activity for the correct performance of the task. Columns (c) and (d) show the average output for the same nodes after lesioning of the pallido-thalamic and thalamocortical projections, respectively. Comparison with the correct activation profiles in column (b) demonstrates that associative loop function is significantly impaired in both cases, though in notably different manners.

After a pallido-thalamic lesion, PFC activity is erratic, particularly beyond the first stage of the task. Multiple activation of representations encoding task stages two and three is observable at $t \approx 1500ms$, and dynamics are observed almost continuously from $t \approx 2000ms$, indicating a loss of stability. Similar effects were observed for the analogous simulation in the investigative model in chapter 2 (see simulation 2.7.8i). The result of a thalamocortical lesion is illustrated in column (d) of figure 4.9. Here, though the correct representation is selected via influence from the transition nodes, a lack of excitatory input from thalamus results in the inability to sustain a selected representation. Upon removal of the influence from the transition nodes, the activity in cortex decays to zero and no action is subsequently selected by the motor loop.

PFC recurrence

We also tested the role of PFC recurrence in the functionality of the associative loop. We lesioned all recurrence, and initiated a new coffee trial. Figure 4.10(a) shows the activation of PFC nodes while receiving external input from the transition nodes at trial initiation, indicating activation of the correct representation. However, after external influences are removed, the representation decays due to a loss of stability resulting from the lesioned connections. The resultant degraded pattern of activity is displayed in figure 4.10(b).

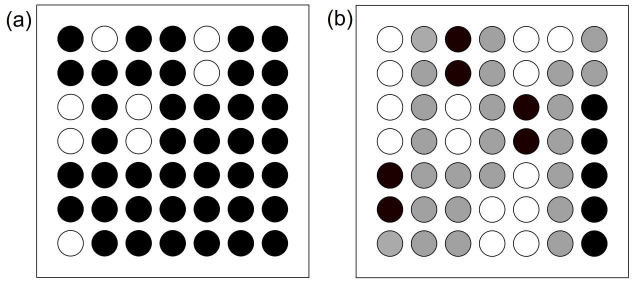

Figure 4.10: PFC node output after lesioning PFC recurrent connectivity, while (a) initially receiving strong input from transition nodes, and (b) after the release of this input. White nodes are strongly activated; black nodes are inactive. Instead of maintaining the selected representation, activity in PFC degrades after removal of input via recurrence in the TC loop.

Interloop corticostriatal projections and affordances

Chapter 3 introduced the concept of candidate actions in the motor loop, initially specified by action affordances and biased by a contextual influence from PFC to cause selection of the correct action. We tested this functionality in the current model by lesioning the PFC → putamen corticostriatal projection and, in a second simulation, the projections from the action affordances to the motor loop.

Figure 4.11(a) shows pre/motor cortical activity at task initiation after an interloop PFC → putamen lesion. Low levels of activation are seen for the appropriate action for the first stage of the trial (**pick up**). However, without the additional biasing influence directly to striatum, this level of activity is insufficient to drive selection in basal ganglia and subsequent positive feedback in the thalamocortical loop. As such, no action selection subsequently takes place. For comparison, pre/motor cortical activity during successful performance of the **add coffee** subtask without lesioning is shown in panel (b).

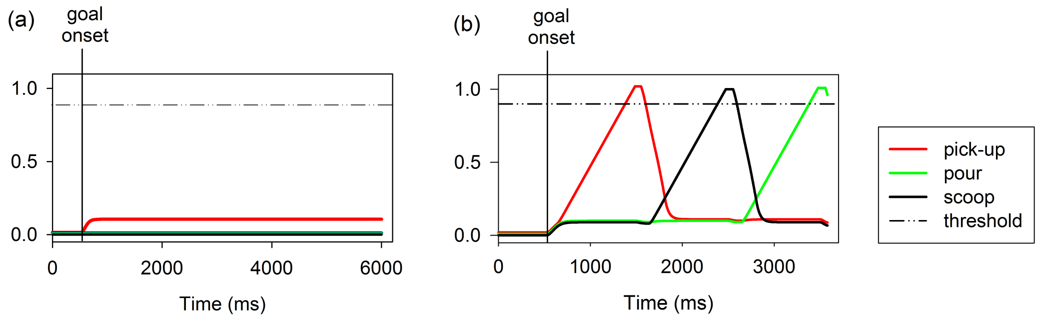

Figure 4.11: (a) Pre/motor cortex output for coffee simulation with lesioned interloop corticostriatal weights. Low level activation of the correct action resulting from affordances is observed, but the lack of an additional bias via corticostriatal projections results in an insufficient level of activity to allow selection. As no selection takes place, the first stage of the task is not completed and no sequencing is observed. (b) Pre/motor cortex output for the first three stages of a coffee trial without lesioning, for comparison. Successful performance of the **add coffee** subtask is indicated by the traversal of the selection threshold by each of the required actions in sequence (activity in the object representations region disambiguates the target object for each action, but is not shown here).

Figure 4.12 displays activity in vlGPi and pre/motor cortex at task initiation after lesioning the affordances to the motor loop. While 'selection' of the correct action takes place in basal ganglia on the basis of corticostriatal influences from PFC, observable as an inhibition of vlGPi activation for the selected channel, no selection takes place in cortex as no direct excitation is received by the thalamocortical loop.

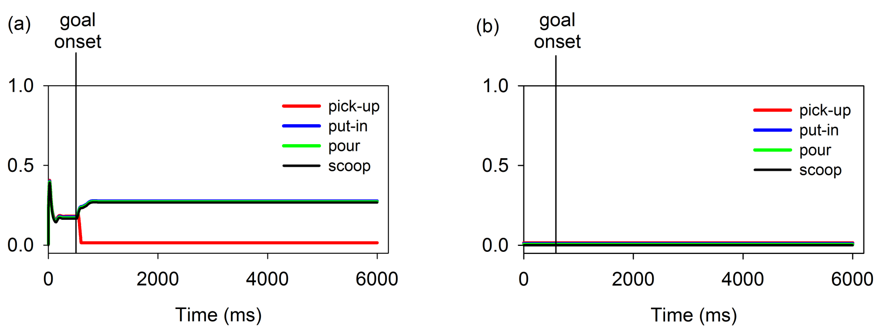

Figure 4.12: (a) vlGPi and (b) pre/motor cortex output after lesioning affordances to motor loop. While the correct channel is disinhibited in GPi, resulting from the influence from the corticostriatal projection, without the active action specification to pre/motor cortex, no selection is possible. See figure 4.11(b) for observed selection behaviour in pre/motor cortex during successful performance of the **add coffee** subtask, for comparison.

4.5.3 Habits as strong affordances

The lesion simulations presented above demonstrate the necessity of both sensory (affordance) and cognitive (contextual) information for correct action selection in a goal directed sequence, and, thus, that the motor loop successfully integrates both sources of information to perform contextually relevant action selection in the initial simulation.

Where these two contributions to motor loop output are compatible and suitably balanced, contextually-relevant object-directed behaviour is the result, as the results in section 4.5.1 demonstrate. However, by increasing the strength of affordances, it is possible to demonstrate behaviour that may reasonably be considered a pure 'sensori-motor' response or habit, based on the fixation of a particular object or other stimulus, heavily associated with a particular action.

We altered the balance of sensory and cognitive inputs to the motor loop, such that affordances had a relatively increased influence in determining motor output. This may be interpreted as a general 'distractedness', and sensitivity to affordance information over task-related influences on behaviour (see appendix C for amendments to weights). Additionally, we increased the strength of the projection from the object representation node representing the cup to the action affordance node representing the pick up action, reflecting a particularly strong tendency to perform a **pick up cup** action upon fixating the cup. With this new pattern of weights, we initiated a new coffee-making trial.

Results are shown in figure 4.13, which details the timepoints at which output of object representation nodes (a) and pre/motor cortex channels (b) exceeded the selection threshold $\theta = 0.9$, during the first three stages of the trial, equivalent to the first subtask, **add coffee**. Consistent with figure 4.5, fixation of objects occurs in advance of action selection. The first two stages of the task proceed as normal, where the spoon is first fixated, then picked up, and the coffee is fixated then scooped. However, upon fixation of the cup, the strong influence to the **pick up** affordance causes strong activation of the **pick up** action in the motor BGTC loop. This activation is sufficiently strong to out-compete activation of the correct

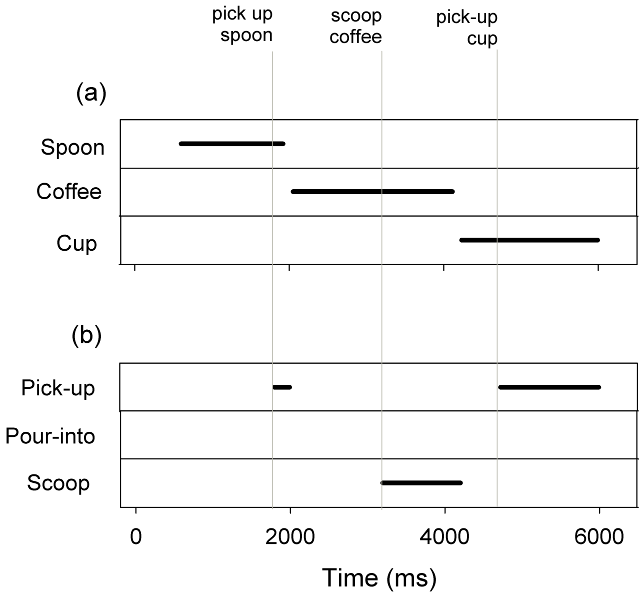

Figure 4.13: Time course of above-threshold activation of key nuclei for the first three stages of a coffee-making trial in which a strong affordance is implemented for the **pick up** action, associated with the cup. (a) Charts show timepoints at which the output of each object node in the object representations region exceeded θ = 0.9, indicating fixation of that object. (b) Charts show timepoints at which the output of each channel in pre/motor cortex exceeded θ = 0.9, indicating selection of the corresponding action. Selection times are indicated by the grey vertical lines. Upon fixation of the cup, rather than selection of the **pour into** action, as would be contextually appropriate, selection of the **pick up** action is performed, indicating a purely sensori-motor or habitual response to the cup.

action, **pour into**, which receives support from PFC, via the corticostriatal projection to putamen, as normal. As such, the incorrect action is selected and **pick up cup** is performed. Here then, a particularly strong affordance is sufficient to initiate an action without a corresponding top-down influence to putamen from PFC. However, it should be noted that such a strong affordance may still be 'overridden' by a sufficiently strong influence to a competing action from PFC, resulting in a top-down inhibition of a habit or over-learned response.

4.5.4 Waiter scenario

A primary distinction between the competing IAN (Cooper & Shallice, 2000) and SRN (Botvinick & Plaut, 2004) models discussed above was the type of representation used to encode contextual information. The SRN model emphasised the power of distributed representations to encode key similarities between different versions of a task or subtask. The authors claimed that this posed specific difficulties for the IAN model, due to the necessity that such variations must be represented by independent schemas. In particular, Botvinick & Plaut (2004) suggested that the schema network would be unable to capture in an efficient way the necessary contextual representations for the performance of what they termed a 'quasi-hierarchical task'; they suggested the example of a waiter having to make several coffees each with different amounts of sugar. The authors emphasised the necessary dependence on context for successful completion of such a task, and argued the localist scheme utilised in the schema network would fail to represent in a useful way the important similarities between these different versions of a single task.

Botvinick & Plaut (2004) presented a simulation based around this task. Successful performance of the task required their coffee-making task to be completed with zero, one, and two sugars, respectively. We replicated this simulation to examine our model's ability to perform different variations of a task.

We designed three alternative sets of representations for the tea-making task; each set of representations included a distinct intermediate level goal node to indicate the number of sugars to be added, in addition to the overall goal node. For the **two-sugar** version of the task, additional **rank order** nodes were included to indicate the sugar adding task history[1]. These representations are illustrated in appendix C. We found that the model was consistently able to complete all three versions of the task without error, according to the specific instruction given.

Botvinick & Plaut (2004) argue that the inclusion of distinct 'goal nodes' required by a schema network to achieve this task would be an 'uncomfortable' measure, given the inability to account for task similarity at the highest level of the schema network hierarchy. However, we effectively utilise this scheme, by including three distinct intermediate-goal nodes to reflect the different versions of the task. Despite this semi-localist structure, how-

[1]*Additional context nodes were also included to explicitly indicate that the termination of the* ***add water*** *and first* ***add sugar*** *subtasks signified the end of the task for the zero- and one-sugar versions of the task. These nodes were included primarily for mechanistic reasons, helping to resolve ambiguity in transition nodes, and serving primarily to help separate heavily overlapping representations. We suggest that in a larger network, more nodes encoding each of the basic features would increase the distance between the original representations and render such additions unnecessary.*

ever, the model is able to perform all versions of the task, while the overall representations continue to capture a large degree of the similarity between each version. This similarity is captured by distinct features from that of the 'goal' itself; the important functionality afforded by overlapping representations is retained despite localist representation at the level of goal. This emphasises the point that overlap between representations may exist at many different levels, and not necessarily at the level of goal; indeed, Botvinick and Plaut's model may well have learned to represent the distinct 'goals' separately on each version of the task as our model does, as from their distributed representations, we cannot deduce how each item of information is represented.

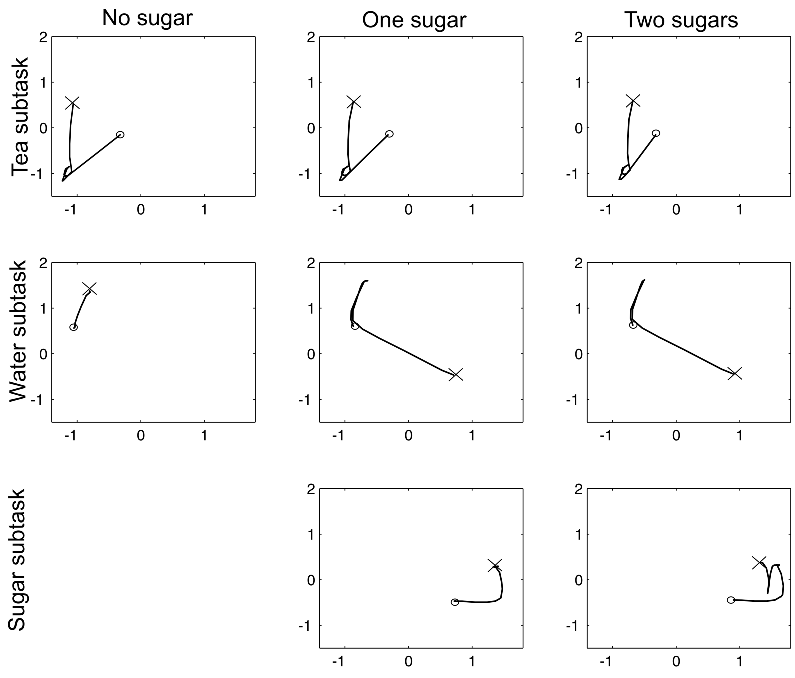

Figure 4.14: MDS plots of the trajectory of PFC through each stage of each subtask in the current model. Similar trajectories are observed for each version of each subtask. Left-hand columns show activity for the no-sugar versions of the task, middle columns for one-sugar, and right columns for two-sugars. Circles and crosses indicate the beginning and end of each subtask, respectively. Compare with figure 4.15.

To illustrate the functional equivalence of the two models, the MDS plots illustrated in figure 4.14 demonstrate the similar trajectories taken by our model in each version of the **add**

teabag, **add water** and **add sugar** subtasks. In particular, the trajectories taken for the **add teabag** subtask are almost identical in each version of the task. Slight differences are the result of the activation of different intermediate goal nodes in each version. In the no-sugar version of the task, a shorter trajectory is seen for the **add water** subtask. This reflects the absence of a switch into the subsequent **add sugar** subtask, compared with the one-sugar and two-sugars versions. Similarly, the trajectory of the **add sugar** subtask is longer in the two-sugars version of the task, and notably, 'doubles back' on itself, reflecting the repetition of the same actions as the second sugar is added.

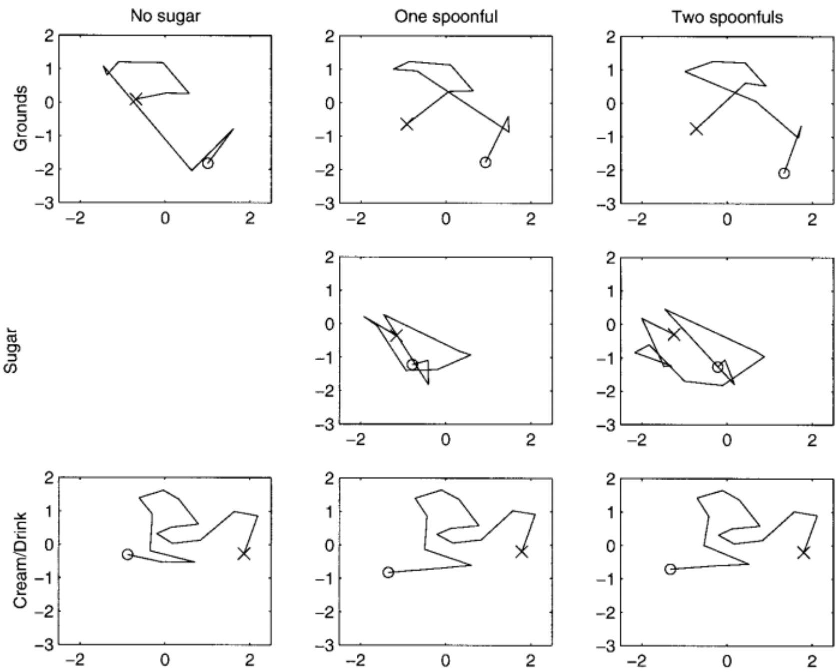

Figure 4.15: MDS plots of the trajectory of the hidden layer through each stage of each subtask in the SRN model (Botvinick & Plaut, 2004). Again, similar trajectories are observed for each version of each subtask. Left-hand columns show activity for the no-sugar versions of the task, middle columns for one-sugar, and right columns for two-sugars. Circles and crosses indicate the beginning and end of each subtask, respectively. (Copyright (2004) by the American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is Botvinick, M. & Plaut, D. (2004) Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review, 111(2)*, 395-429. The use of APA information does not imply endorsement by APA).

Our analyses compare favourably with equivalent MDS plots produced by Botvinick and Plaut (reproduced in figure 4.15), which also emphasise the similarity of trajectories for each version of each subtask. Again, this indicates the equivalent dynamics of the two models, despite a localist representation of the goal in our model. Thus, we suggest that the differences in the trajectories in Botvinick & Plaut's SRN model reflect the same information as that which is encoded by our distinct goal and context nodes. A perfectly acceptable alternative to our scheme would be to include just a single goal node ('make tea with x sugars') for all three versions of the task, in a more strongly localist implementation. Either scheme is likely to be suitable for the task, as each is perfectly able to capture important contextual similarities between the three versions of the task. This further supports the notion that the explicit hierarchies of schema based models and implicit hierarchies embodied in the distributed representations of SRNs may be functionally equivalent, particularly where schema models are able to utilise the same component or low-level schemas for multiple tasks. This in itself effectively results in the same overlap, or capture of similarity, emphasised by SRN models.

4.6 Discussion

4.6.1 Summary of findings

In this chapter we have shown that the architecture designed in chapter 3, based on a series of neuroanatomical and computational arguments, is able to successfully mediate the production of multiple action sequences sharing both constituent actions, and constituent subtasks. We have shown that these component actions may be performed independently of the preceding action, and multiple times in a single sequence. Composite actions may also be recruited for the completion of distinct subtasks. As has been noted in previous modelling literature, such functions are not trivial (Dominey, 1995; Botvinick & Plaut, 2004), and previous modelling attempts have struggled to capture some of these nuances of sequential action (Rumelhart & Norman, 1982). In the present study, cognitive representations of temporal task context allow the model to maintain an implicit or explicit record of the preceding actions and ultimate goal, in turn supporting the composition of low level actions into sequences in a flexible manner.

In applying a model based on the architecture developed in chapters 2 and 3 to the performance of these two typically familiar and well learned tasks, we have highlighted a large degree of functional equivalence between previously published competing models, whereby the time course of activity in individual PFC nodes reflects that of distinct schemas and goals at different levels of Cooper & Shallice's (2000) IAN model (see figures 4.6 and 4.7), whereas the trajectory of the PFC dynamics as a single system illustrates similar processing in the current model as that observed in the hidden layer in Botvinick & Plaut's (2004) SRN model (see figure 4.14). This is discussed further with respect to a reconciliation of model differences in section 4.6.3.

4.6.2 Significance of novel architecture

Here, we have shown the applicability of the 'subset-selection' scheme, introduced in chapter 2, to the mediation of two highly similar but distinct tasks, with complex demands on sequencing and selection, and reiterated the importance of the architecture for selecting and maintaining the required representations. Lesion studies showed that PFC recurrent, thalamocortical, and pallido-thalamic connectivity are all required for the successful selection and maintenance of the necessary representations in the associative BGTC loop. It is also notable that we have implemented 22 stable representations. Traditionally, examinations of the storage capacity of recurrent networks based on the Hopfield prescription suggest that capacity sits around 0.15N, where N is the number of nodes in the network (Hopfield, 1982). Here, taking into account the nodes within basal ganglia, the loop has a total of 79 nodes, suggesting a capacity of 11.85 distinct patterns. Clearly, we have exceeded this theoretical capacity limit. That we have not compromised the potential capacity of a recurrent system of 79 nodes is supportive of the novel interpretation of the associative BGTC architecture we have taken here, and may have implications for understanding the functional advantages of the structure of the BGTC system. For example, the basal ganglia as a 'central selection mechanism' as described by Redgrave and colleagues (1999) may in fact be adaptive beyond the reasons addressed by the authors; such a mechanism may allow the system to reach higher capacity limits. However, recent work has shown recurrent networks able to support up to N stable patterns (Wu et al., 2012). It is for future investigations to examine whether this would be possible with the current BGTC loop architecture, though the results presented here are promising.

Specialisation of prefrontal regions and nodes

In the current study, we have implemented particularly simple PFC representations of temporal context, encoding individual features using individual nodes; actions are represented separately from objects, which are represented separately from goals, and so on. While numerous electrophysiological studies have found prefrontal cells selective for particular stimuli (e.g. Rao, Rainer & Miller, 1997), much of this selectivity has been shown to be modified by other task features (see Tanji & Hoshi (2008) for a recent review). This suggests that individual PFC cells have a greater integrative function than we have accounted for here. However, we maintain that both functional specialisation and representational localisation in PFC are likely. Indeed, Funahashi and colleagues (Funahashi et al., 1993) explicitly suggest that individual neurons encode *partial* information required for the execution of a particular action, 'such as one target location or one movement direction' (p171), rather than more comprehensive information relevant to the whole movement or task.

It is also likely that the conjunctive specialisation observed in experimental work reflects a similar functionality to the single-feature based organisation that we have emphasised here. Indeed, Miller & Cohen (2001) state that an important role of PFC in cognitive control is the 'active maintenance of patterns of activity that represent goals and the means to achieve them', as we have attempted to implement in PFC here. Notably, one feature in particular that we have included has been documented several times; these are neurons encoding the **rank order** of stimuli (Barone & Joseph, 1989; Funahashi et al., 1993, 1997). Such neurons have also been the focus of modelling efforts which emphasise their role in supporting the learning of new sequences (Salinas, 2009). The role of these neurons was critical to flexible sequencing in our model; we suggest that these neurons have been documented so frequently due to their vital importance in ensuring the correct sequencing of actions. More specifically, however, a prediction that arises from this model is that such neurons are likely to be observed only in those studies where the temporal order of stimuli is *variable*, in that stimuli or actions may occur in different sequences. Where sequences are consistent, no explicit sensitivity to temporal order would be expected, as sequence information is implicit in the representation of the task stage itself.

The transition layer itself may be conceived of as encoding the particular 'feature' of sequence knowledge. The externalisation of this feature from the contextual representation itself was important for retaining the stability of PFC representations; indeed, a computational imperative exists for an external mechanism to generate controlled sequencing of distinct representations which are intrinsically stable (Rutishauser & Douglas, 2009). Interestingly, sequencing mechanisms in particular appear to have caused difficulty for previous models, which have tended to rely on mechanisms such as inflexible lateral inhibition (Rumelhart & Norman, 1982), or recurrent dynamics (Dominey et al., 1995; Botvinick & Plaut, 2004) which cause problems for sustained activation.

While the associative loop itself was able to resolve competing inputs and thus perform selection independently (see chapter 2 for a discussion), in the current model the loop rarely receives such competing inputs, as only a single transition node is generally active at any one time. Thus, in the present study, the transition layer is arguably performing an important selection function. It is of particular interest that distinct regions in PFC have been shown to be selectively involved in maintenance or selection (Rowe, Toni, Josephs, Frackowiak, & Passingham, 2000; Tanji & Hoshi, 2008). This raises the possibility that our transition layer may be performing a similar function to the 'selective' prefrontal area 46, whereas our model of the associative loop itself may be performing maintenance functions akin to prefrontal area 8 (Rowe et al., 2000). We suggest that this very specialisation in PFC is observed as a result of the computational incompatibility of internal sequencing and stability indicated by Rutishauser & Douglas (2009). A prediction arising from this is that selective damage to different regions of PFC may result in different patterns of degradation of sequential behaviour; this idea is explored in detail in the next chapter.

Translation of information from cognitive to motor 'co-ordinates'

In chapter 3 we examined the nature of the salience signal in the original GPR model (Gurney et al., 2001a, 2001b). We concluded that this signal reflected information regarding the specification of candidate actions based on the objects or other stimuli currently perceived. We suggested that for goal directed tasks, such information should be enhanced and/or disambiguated by a cognitive representation of the intended action, in order for context appropriate action selection to take place. Here, we have shown that the model indeed requires the convergence of these influences in order to perform such contextually sensitive sequential action selection. The necessity for this combination of influences stems from two aspects of the model. Firstly, an affordance may not be sufficiently strong to trigger selection in of itself, and thus the corresponding motor representation may require additional excitation; secondly, multiple specified affordances may require contextual disambiguation if they are activated with sufficiently similar strength.

These functional influences of the two interloop pathways are consistent with evidence suggesting that damage to the ventral visual stream can result in the inability to use object knowledge to guide action (Goodale & Milner, 1992), and that lesions to the prefrontal cortex can result in *utilisation behaviour* where action is disproportionately guided by external stimuli rather than intention (L'hermitte, 1983). Indeed, results presented in section 4.5.3 support the latter assertion, whereby a particularly strong affordance can 'override' goal related influences originating in PFC. The current model thus makes explicit the functional relationship of these two pathways and their roles in goal directed, sequential action selection.

It is also of note that our interloop corticostriatal projection is relatively sparse in terms of the volume of PFC from which it originates and to which it projects in putamen, arising only from those nodes concerned with the cognitive encoding of the required *action* itself, and projecting to a single channel in putamen. This is consistent with evidence showing that interloop corticostriatal projections are not as widespread as those within loops (Calzavara et al., 2007; Haber & Calzavara, 2009). We contend that this is due to the propagation through different functional territories of only certain categories of information; perhaps those most important for influencing the current action selection. If we are correct in this inference, the observed pattern of connectivity would be far less likely if prefrontal representations were fully distributed, as the entire representation would be important for decoding the correct action. This observation thus provides further support for the likelihood of a feature based or semi-localist organisation of cognitive representations of context.

4.6.3 Comparison with existing models

Throughout this chapter we have referred to the existing debate between two sets of authors regarding the form of the underlying functional substrate necessary for the performance of these and other well-learned tasks (Cooper & Shallice, 2000; Botvinick & Plaut, 2004). While these models have adopted significantly different approaches to understanding the mediation of sequential behaviour, one notable feature they have in common is the lack of contact with biological data; neither take into account a great deal of evidence from neuroanatomy or neurophysiology in order to guide model design. Here, we have used neuroanatomical evidence as a primary constraint on model development. The resulting architecture has additionally reconciled aspects of the two competing models, effectively creating a 'hybrid' model displaying a degree of functional equivalence to both.

The architecture we employed imposed particular requirements on the nature of the representations needed for the task. Information pertaining to objects and actions was necessary in order to guide selection of the appropriate action along two converging pathways. Additional information regarding the current subtask, goal and, where applicable, rank order, was necessary to ensure a correct transition to the next representation on completion of the current stage of the task. In this chapter, we have shown how the activity of the resulting feature nodes in PFC has commonalities with the activity of the equivalent underlying substrate in both models, due primarily to the feature-based organisation of our PFC representations. Activity in nodes representing different levels of the task hierarchy show qualitatively similar profiles over time as equivalent goal and schema nodes in Cooper & Shallice's (2000) IAN model. However, MDS plots demonstrate the functionally equivalent processing in our PFC and the hidden layer of Botvinick & Plaut's (2004) SRN model. As such, we argue that these two approaches are easily reconcilable. Both sets of authors do in fact acknowledge that their differences are not catastrophic (Cooper & Shallice, 2006a, 2006b; Botvinick & Plaut, 2006b, 2006c), with Cooper & Shallice (2006a) stating that,

> 'recurrent networks and interactive activation networks may be reconciled through the

> mapping of nodes within the interactive activation network to discrete point attractors ... within the recurrent network ... one can be optimistic about the development of a model which functions at one level according to the principles of Botvinick and Plaut, and at another according to our... principles' (p906).

Indeed, this is what we have effectively achieved here. However, it is worth emphasising that the current model was not constructed with a reconciliation at this level in mind; rather, it has emerged naturally as a result of the implementation of a biologically plausible architecture. Beyond making this reconciliation explicit, we suggest that the IAN and SRN models are, to a large extent, functionally equivalent in their existing instantiations. We believe that the opaque nature of the emergent representations in the SRN model obscures the fact that the same key information is likely to be represented on the same timescales in both models. In particular, the representations of goals, which provoked a great deal of debate between the two sets of authors, are equivalent in their role in directing behaviour to that of schemas, where both may be simply regarded as key 'features' in the overall representation of context. This is made transparent by our neurally inspired model, but is essentially not a novel function.

Botvinick & Plaut (2004) argue that the localist representations of goal and other task features limit model performance by failing to capture important similarities between tasks. On this, we note two important points. Firstly, that we have shown this to be inaccurate. Where schemas are implemented as nodes within a network, this contextual similarity inevitably emerges *despite* the localist representation of individual features. The ability of the model to perform the 'quasi-hierarchical' task of making several drinks with different amounts of sugar demonstrates this principle. Additionally, as Cooper & Shallice (2006a) point out, the IAN model does incorporate 'overlapping' representations where higher level schemas make use of the same subschemas for different tasks. Secondly, evidence now suggests that neural representations are unlikely to be fully distributed (Bowers, 2009; Quiroga et al., 2005; Gross, 2002). We again suggest that an intermediate, semi-localist representation scheme is most plausible, and captures sufficient similarity for flexible task performance.

4.6.4 Significance for understanding neural data

The seven stages of the tea task were represented in PFC by seven corresponding representations (T1-T7 in figure 4.1). Activation of each of these representations in the correct sequence was required for successful performance of the task. However, due to differing degrees of overlap between each representation, this effectively resulted in *partial* activation of *multiple* representations at any one time. For example, all representations related to the tea task, T1-7, share the same representation of goal. As such, all these representations show partial activation for the duration of the whole task, due to the continued activation of this particular feature throughout the task.

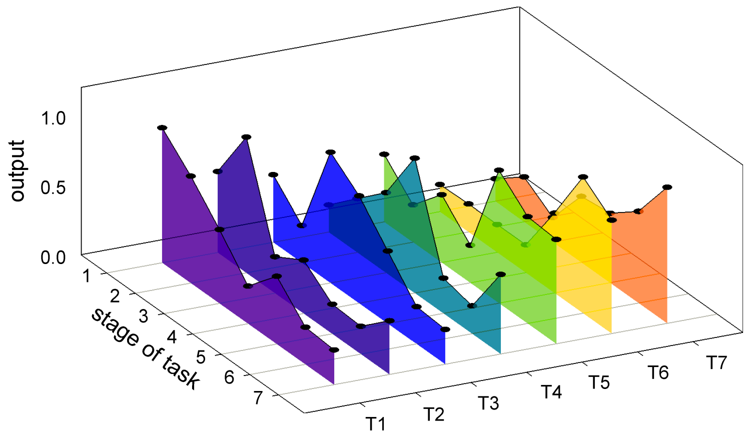

Figure 4.16: Average output for PFC representations T1-T7 (see figure 4.1) for each stage of a successfully performed tea-making task. Note that all representations show a degree of activation during the entire task, due to overlapping features in the representations. This pattern of results is qualitatively similar to neural data found by Averbeck and colleagues (2002; see figure 4.17), though a different interpretation of the underlying mechanisms was given. See text for details.

We took the average output of each representation involved in the tea-making task for each stage of the task. The resultant traces are plotted in figure 4.16. Notice that, from task initiation (stage 1), all representations are partially active. Trivially, each representation's activation peaks during the stage in the task that it has been designed to encode. However, note also that intermediate levels of activation of a particular representation are often observed adjacent to the point of peak activity; rather than a sudden increase in activity at the corresponding stage of the task, the representations generally show a gradual increase and then decrease in activation during the preceding and subsequence stages of the task. This is particularly evident for representations T1, T6 and T7.

This profile resembles neural activity recorded in macaque PFC during a sequential drawing task (Averbeck et al., 2002). In this study, monkeys were taught to draw geometric shapes, such as triangles. Isolable 'ensembles' of prefrontal neurons were found to preferentially encode particular stages of the task, where each stage generally consisted of drawing one side of a shape. These stages were equivalent to the performance of a single action in our current model. Activation of all ensembles was observed from the initiation of the task (figure 4.17). This early activation of each ensemble was interpreted as low level 'preparatory' activation of each representation required for the task, and taken as supportive of competitive queueing accounts of sequencing (e.g., Houghton, 1990), which emphasise the parallel activation of representations of all sequence components from the initiation of a sequence. Here, however, we are able to proffer an alternative possible explanation for the observed results. Rather than a low level activation of entire prefrontal representations, we have shown that qualitatively similar patterns may be observed as a result of high level activation of *partial* representations, as a result of overlap, to a greater or lesser extent, with the currently selected representation. In other words, the activation of one representation effectively results in the partial activation of others, due to varying degrees of overlap between the representations required for a task. Here, the apparent gradual onset and offset of each representation over multiple task stages is in fact a result of the tendency for greater overlap between representations that encode adjacent stages of the task.

The similarity of our results to neural profiles shown by electrophysiological data lends some support to our interpretation of PFC representations as feature based. However, the contrasting interpretation of the neural data given by Averbeck and colleagues (2002) - that all sequence components are activated in parallel - indicates that multiple possible mechanisms may in fact give rise to the same trends in the data. It would be of interest for future neurophysiological work to attempt to distinguish these possibilities; a suitable task for study might involve a sequence wherein temporally adjacent actions share few semantic features, and those separated in time share more. Our model would predict that a cognitive

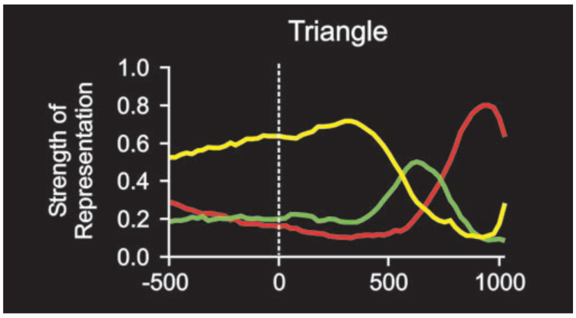

Figure 4.17: Monkey PFC 'ensemble' output during a sequential drawing task showed early and parallel activation of all ensembles correlating with each stage of the task. This was interpreted as evidence supportive of competitive queueing accounts of sequencing. In the present study, we see similar patterns of PFC activation (see figure 4.16), though the underlying mechanisms contrast with the interpretation given by Averbeck et al. See text for details. (Adapted from Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20), 13172-13177. With permission. Copyright (2002) National Academy of Sciences, U.S.A.)

representation of the more temporally distal action would show a stronger activation earlier in the sequence, whereas the hypothesis posed by Averbeck and colleagues (2002) would predict the opposite pattern. Understanding the true nature of this pattern of activity has tremendous implications for how we view the organisation of representations of context, and in turn, the organisation of sequential behaviour.

On goals and habits

We noted in chapter 1 that we do not consider routine tasks such as tea-making to be truly habitual. However, undoubtedly these tasks are well-learned, and performed with more ease than novel tasks, requiring significantly fewer cognitive resources to perform successfully. Indeed, behavioural evidence has been found to suggest the existence of schema-like processes underlying performance on routine tasks in neurologically impaired individuals (Forde et al., 2004). Truly habitual behaviour, however, is unlikely to require any great involvement of associative neural regions such as PFC (Redgrave et al., 2010) and a vast amount of evidence has been gathered to suggest that as behaviour becomes automatised, its performance relies less on processing in associative and more on processing in motor regions of the BGTC hierarchy (Graybiel, 2008; Yin et al., 2009; Seger & Spiering, 2011).

With this in mind, results from our model presented in section 4.5.3 suggest that a habit may be understood as an *overlearned affordance*; fixation of a particular stimulus might trigger an action without the support of a corresponding contextual or goal based influence if an affordance is sufficiently strong. It may be that in routine tasks, the affordances related to the objects required for the tasks themselves *have* been overlearned to some degree and may thus require little contextual disambiguation. However, given that it is not difficult to override the initiation of a tea- or coffee-making sequence upon the sight of a kettle or a teabag, it is unlikely that these affordances are sufficiently strong that they require active inhibition, and are therefore unlikely to be truly habitual. Furthermore, frontal lobe damage has repeatedly been shown to affect performance on routine tasks (Schwartz et al., 1991, 1995; Humphreys & Forde, 1998); this is contrary to what might be expected if these tasks were primarily supported by affordances alone, which are accepted as relying on parietal regions of cortex (Riddoch et al., 1989; Fagg & Arbib, 1998; Cisek, 2007). Additionally, evidence from studies of distraction also suggests that these sequences cannot truly be regarded as habit in the strong sense described by Dickinson (1985), as a higher number of mistakes are made under conditions of distraction, even on routine tasks (Morady & Humphreys, 2009).

Of particular interest is work by Aarts & Dijksterhuis (2000) which conceives of habits as automatic behaviours which nonetheless require the presence of a related goal. This type of behaviour perhaps lies somewhere between truly goal-directed and strongly habitual behaviour; we contest that such goal-directed automaticity accounts for routine behaviours and requires greater involvement of executive structures than strongly habitual actions. This would account for the fact that routine sequences may be carried out with minimal cognitive effort, but would not necessarily assume that this is a result of inflexible, overlearned affordances. Rather, we suggest that such behaviour results from well learned, efficient prefrontal representations of task context. It is possible that, early in learning, multiple items of information required for a task must be maintained individually. With learning, more information may be encoded within a single, internally stable, prefrontal representation, which may reasonably be called a schema. This might, over time, free up working memory capacity, allowing other tasks to often be carried out in tandem.

4.6.5 Limitations

Plasticity

Conspicuously, we have not included plasticity in the current instantiation, so are unable to comment on the possible mechanisms underlying the emergence of the representations supporting sequencing, the association of contextual representations with objects and actions, and the emergence of temporal order knowledge.

In a preliminary study of plasticity however, we implemented a two-stage learning procedure in which learning of composite actions was followed by the construction of sequences. In the action learning stage, the weight matrices W^{Λ} and W^{xp} were subject to Hebbian and reinforcement-learning inspired algorithms, respectively (see figure 4.18). These successfully adopted the correct connectivity in order to associate representations of action in PFC, and object representations with particular actions in the motor loop. In a second, sequence learning stage, the projections W^{ξ} and $W^{x\Gamma}$ to the transition nodes from the environment representation ξ and from PFC were subject to learning in a supervised, Hebbian learning procedure. Again, suitable connectivity emerged in order to mediate appropriate sequencing. Notably, where attempts were made at learning all projection patterns simultaneously, affordances corresponding to actions that were performed more frequently in the sequence (e.g., **pour into cup**) were overlearned, interfering with learning of those which appeared less frequently (e.g., **put into cup**). While it is not unreasonable that more frequently encountered actions should be represented more strongly, the degree of interference encountered suggests a possible necessity for an early 'motor babbling' stage of skill development, which must precede more complex sequence learning.

While this preliminary work showed promising results in terms of the ability of the model to adopt appropriate patterns of connectivity, it leaves open the question of the means by which the PFC representations themselves form. Additionally, the action learning stage simply associated existing representations of context with existing representations of action; this raises the question of which arises first, and whether the pre-existence of one type of representation affects the manner of formation of the other. These are important questions

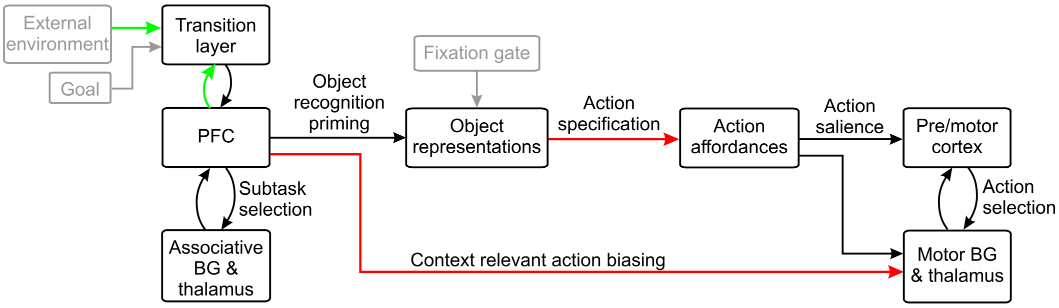

Figure 4.18: Loci of plasticity in a preliminary learning study. Projections illustrated in red were learned in an early action learning stage. Those illustrated in green supported sequence knowledge and were learned in a later sequence learning stage.

to be addressed in future instantiations of the model.

Sensory influences

We have included only very limited explicit sensory influences, in terms of the 'fixation gate' on object representations. This influence is entirely passive, indicating only the presence of an object in the immediate environment and subsequent 'availability' for fixation. It is likely, however, that different objects and other stimuli are more or less salient in themselves, and thus more or less likely to be fixated irrespective of concurrent goals. This is likely to have a significant effect on the performance of sequential tasks, particularly under conditions of distraction or in impaired populations (Forde et al., 2004). Additionally, different affordance strengths associated with each object are likely to have a significant effect on performance, which we have not addressed here. Future work should include explicit representations of external stimuli and some representation of their intrinsic salience, as well as the likelihood of differing affordance strengths.

Additional BGTC loops

We introduced three new regions in the current model: the transition layer, the object representation region, and action affordances. It is likely that each of these is contained within a distinct BGTC loop; indeed, neuroanatomical evidence points to the existence of such circuits, if we assume that our object representations and action affordances reflect processes in inferotemporal and parietal cortices (Middleton & Strick, 2000). Subtle effects on selection in these regions are thus not currently captured, which may have significant effects on performance in a more comprehensive model. Furthermore, where interloop corticostriatal projections impose a *bias* on the selection of motor representations, we include a direct excitatory effect from PFC on object representations. It is likely, however, that PFC has similarly soft biasing influences on visual search and perception as we have included in our motor loop. This is likely to combine with specific salience values related to external stimuli in order to produce much more sophisticated selection processes in object representations, which is likely to effect the subsequent processing of affordances, thus having a direct effect on action selection.

4.6.6 Summary

In the present chapter we have established the viability of the theoretical architecture outlined in chapter 2 to mediate multiple, related, goal directed action sequences, and discussed its implications for understanding the organisation of information for such sequences in the brain. In the following chapter, we examine the effects of localised disruption to the model in order to validate the model architecture against behavioural constraints, and we discuss the resulting behaviour with reference to everyday action slips and action disorganisation syndrome.

Chapter 5

The simulation of action slips and action disorganisation

5.1 Introduction

As discussed in the general introduction, a great deal of data have been collated on the patterns and occurrence of human error commission during sequential performance, both in healthy volunteers (Reason, 1979, 1984) and in various patient populations (Buxbaum et al., 1998; Humphreys & Forde, 1998; Schwartz et al., 1998). Much of this has focused on errors made during routine tasks described as 'activities of daily living' (ADLs), familiar, well learned action sequences of which tea- and coffee-making are common examples. Traditionally, models of sequential processing have been examined under a degree of disruption in order to examine their ability to account for patterns in these error data. This is a well accepted test of the validity of any such model (e.g., Henson, 1996), and has been a focus of the competing IAN (Cooper & Shallice, 2000) and SRN (Botvinick & Plaut, 2004) models discussed extensively in the previous chapter.

In chapter 4, we established the practical viability of the theoretical architecture developed in chapter 3 for sequential performance. We showed that a computational model utilising this architecture, based on the known neuroanatomical organisation of the BGTC loop system and a number of computational constraints, was able to perform two routine-type sequences sharing several actions and subsequences, and had important potential implications for the interpretation of existing neuroanatomical and neurophysiological findings. In this chapter we aim to examine the performance of the model under various levels of noise in order to simulate the effects of distraction on normal performance, and more severe disruption that might correspond to deficits observed in disorders of action selection. In doing so, we continue to refer to the ongoing debate over the underlying mechanisms, and discuss the possible relationship between the perspective of schemas and hierarchies, and non-hierarchical emergent representations.

5.1.1 Accounting for human error in routine action

In the general introduction we discussed findings from human behavioural studies and neuropsychology that detailed typical patterns of error commission in healthy volunteers and sufferers of action disorganisation syndrome (ADS). To briefly recap, a series of diary studies examining action slips, or actions 'not as intended' (Reason 1979, 1984, 1990) indicated that healthy participants commonly committed errors during the performance of routine actions. These tended to consist of the perseveration, omission or intrusion of well formed subsequences, and occurred at natural branch points in performance. Additionally, object substitutions were common, where the correct action was performed with an erroneous object.

ADS sufferers tend to show greater disruption to routine action. Most notably, behaviour is more incoherent, with a key feature being the performance of *single action errors*, contrasting with the tendency in healthy controls to make errors at the level of full subsequences. Single action errors may manifest as the omission of a single action, or the performance of *independent actions*: single actions performed outside the boundary of clear subsequences. Independent actions include, but are not limited to, *toying* behaviour, where objects may be aimlessly picked up and put down (Cooper et al., 2005). Perseverations and intrusions consisting of both independent actions and complete subsequences are also observed in ADS sufferers, as are object substitutions.

Various explanations have been proffered to account for the patterns of errors seen in the data for both action slips and ADS, many of which revolve around the disruption or faulty activation of action schemas (Humphreys & Forde, 1998; Norman, 1981; Schwartz et al., 1991, 1998). In the general introduction, we discussed two particular hypotheses which have been examined in computational modelling work. The suggestion that ADS results from an imbalance of top-down and bottom-up influences on the activation of action schemas (Schwartz et al., 1991) was tested with the IAN model (Cooper & Shallice, 2000). An alternative interpretation that a general deficit in cognitive processing resources was responsible for ADS symptoms (Schwartz et al., 1998) was examined by Botvinick & Plaut (2004) in the SRN model (see also Cooper et al., 2005). The IAN and SRN models reproduced a number of features typical of healthy and pathological error commission lending some support to the hypotheses, though as discussed, both models failed to replicate particular error types, and neither took a neuroanatomically focused approach to understanding routine sequential action, focusing rather on cognitive (Cooper & Shallice, 2000) or mechanistic (Botvinick & Plaut, 2004) explanations. Moreover, evidence from patients has suggested that the processing resources deficit is insufficient to account for patterns of errors in ADS (Forde & Humphreys, 2002).

An hypothesis which has received less attention in the modelling literature is given by Humphreys, Forde and colleagues (Forde & Humphreys, 2002; Forde et al., 2004; Humphreys & Forde, 1998; Morady & Humphreys, 2009) which describes the disruption of temporal order knowledge as a predominant cause of errors in ADS. Moreover, throughout their work the authors have also suggested a possible breakdown of the stored schemas encoding the component actions themselves in ADS (Forde & Humphreys, 2000; Forde et al., 2004; Humphreys & Forde, 1998; Humphreys et al., 2000). In experimental tests of these hypotheses, the authors found that patients showed both impaired action *and* order knowledge (Humphreys et al., 2000). However, they also suggested a dissociation between these processes may be possible (Forde et al., 2004; Humphreys et al., 2000). This potential decoupling of component action representations and their associated temporal order is consistent with suggestions made by Sirigu et al. (1996) and Partiot et al. (1996), though predictions made by the latter regarding specific brain areas encoding these processes have not been confirmed (Humphreys & Forde, 1998).

5.2 Current study

Where the hypotheses that ADS results from a weakening of top-down influences on action selection (Schwartz et al., 1991) or from a general reduction in cognitive resources (Schwartz et al., 1998) have been simulated in computational models, to the best of our knowledge no model has tested the proposals that the disruption of action schemas or temporal order knowledge are responsible for these error patterns. The architecture of the current model lends itself well to the testing of these hypotheses, given the functional separation of the action schemas themselves, manifested in PFC representations, and temporal order knowledge, encoded in the projections between the transition layer and PFC.

In the general introduction, we set out a goal for the current chapter to examine how the structure, organisation and dynamics of PFC representations account for human error data. Here, we break this into three separate issues which we aim to address by the simulation of disruption to temporal order knowledge and action schemas:

1. Does the disruption of the same functional mechanism underlie action slips observed in normal performance, as well as the more severely degraded performance seen in ADS?

2. Can disruption to a single mechanism account for all error types and rates observed in ADS, or might damage to different processes be responsible for particular patterns in the data?

3. If disruption to a single mechanism is responsible for ADS, can this mechanism be identified as the disruption of either temporal order knowledge or discrete action schemas?

5.2.1 Categorisation of errors

Errors have proven difficult to classify in previous work, with multiple coders required to identify error types in human performance and some disagreement in their analysis (Schwartz et al., 1998). Different studies of both human and model performance have used slightly different systems in order to describe and classify patterns of errors. With this in mind, in the following discussion, we attempt to define a system which classifies errors committed by our model in a way that allows comparison of the model data with as much of the existing work as possible. Where previous studies have inconsistent interpretations of a particular error type, we highlight this and attempt to examine our result with reference to each viewpoint, drawing attention to the implications such inconsistency has for understanding human behaviour, and the risks of generalising across studies using different analyses.

1. **Omissions**

   In previous work, omissions may be classified where either a full subsequence or single action was omitted (Cooper & Shallice, 2000; Schwartz et al., 1998). In the present study, we classify an omission error wherever a key item is not added to the cup (i.e., where the 'crux' action of a subtask is not performed, Schwartz et al., (1991)). This may take the form of the omission of a full subtask; for example, where the **add water** subtask might be absent in its entirety. However, if the kettle is picked up but the contents not poured into the cup, this is also counted as an omission. The picking up of the kettle was then classified as an independent action, as one performed outside of the context of an intact subtask.

2. **Sequence errors**

   Across previous work, what is classified as a 'sequence' error has varied according to the individual author and, indeed, study. In light of this, we present four different classifications of sequence error, and later summarise our results according to these different views.

   - **View i:**

     The first view of sequence errors is consistent with the first of two sets of analyses performed by Schwartz et al. (1998). This included so called *anticipation-omission* errors, whereby an action is performed without having first performed a prerequisite action. An oft-cited example of such an error is pouring cream from a closed container without having first opened it. A suitable analogous error in our model is attempting to pour from the spoon without having first scooped either the sugar or the coffee. Also included in view i are reversals - where two actions or subtasks are performed in the wrong order - and perseverations of either whole subtasks or component actions.

- **View ii**

  This view is based on the second analysis performed by Schwartz et al. (1998). This view included only anticipation-omissions and reversals.

- **View iii**

  This view is based on the analysis of Humphreys and colleagues (Humphreys & Forde, 1998). Here, perseverations are not considered to be sequence errors, but classified separately. Anticipation-omissions and reversals are included in this definition. Finally, a sequence error was recorded when a subtask was performed earlier in the sequence than would be expected according to healthy participants' typical descriptions of the current task. In our model, this might manifest as adding the sugar to the cup without having first added the water. Conversely, if the water is added after the sugar, this is classed as a reversal.

- **View iv**

  Finally, we examine sequence errors according to the view adopted by Cooper & Shallice (2000). Here, perseverations of whole subtasks were included, as were reversals and performing subtasks out of the standard order. However, anticipation-omissions and perseverations of single steps were not included.

3. **Additions**

   Addition errors have generally not been well defined in previous work. Botvinick & Plaut (2004) describe them as 'actions that do not appear to belong to the assigned task' (Botvinick & Plaut, 2004), which conceivably encompasses *intrusion* errors, where an action or subtask belonging to a different task is mistakenly performed. However, Botvinick & Plaut (2004) give an example from their add-sugar subtask, of scooping sugar with a sugar bowl lid, then putting down the lid, as an addition error. This is arguably an object substitution (see *semantic errors*) - where the sugar bowl lid is substituted for the spoon - followed by a discontinuation of the subtask, rather than an addition. Similarly, examples given by Schwartz et al. (1998), suggest an addition is an anomalous action not even belonging to any particular task.

For clarity, in the present study we only include intrusion errors in this category. Such an intrusion error might consist of adding milk during the tea-making task. Conversely, where similar actions to that described by Botvinick & Plaut (2004) are performed, these are classed as object substitutions.

4. **Semantic errors**

   Semantic errors consist primarily of object substitutions: actions which are appropriate in context, but that are performed with an erroneous object, albeit one which tends to be semantically related to the target object. We also include object omissions in this category, an example of which often given in previous work is the stirring of coffee with a finger instead of a spoon. Although in previous work (Humphreys & Forde, 1998) object omissions have been classified separately, we consider this to be a distinct class of substitution and thus a semantic error. Also, as is clear in the results sections, such errors accounted for a very small fraction of the total errors and due to the underlying mechanisms it was more meaningful to group these error types in our analysis.

5. **Quality/spatial errors**

   These consist of performing correct actions, but involving inappropriate quantities or volumes of objects, or performing them at inappropriate points in space. Though the current model can produce errors which have been classed as quality errors in previous models - pouring cream four times in a row, (Botvinick & Plaut, 2004) - we consider this to be a perseveration error; we interpret a true quality error to consist of a prolonged single pour. Due to design choices, our model has no real capacity to produce these types of errors, as it includes no representation of quantity or location; thus, we do not consider these types of errors in our analysis.

Subtask based and single action errors

In the following simulations, we frequently refer to subtask based or single action errors. To clarify our discussion of this dichotomy with regard to the five error types listed above, subtask based errors may consist of omissions, perseverations, reversals or intrusions of entire subtasks. Single action errors consist of omissions of single actions, object substitutions, or the performance of independent actions outside the boundaries of subtasks as they are defined in chapter 4. Independent actions themselves may also be further categorised as perseverations of a previously performed action, intrusions from another task, reversals of single actions or anticipation-omissions.

5.3 Disrupting sequence knowledge

In the first simulation, we examined the effects of disruption of temporal order knowledge. In order to achieve this disruption, we added noise to the transition layer. This affected sequence 'knowledge' by increasing the probability that erroneous transitions between PFC representations would be made, via spurious activation of incorrect transition nodes. Note that no noise was injected into the PFC or the associative loop itself; once a representation was selected, it remained intrinsically stable as in the minimal noise simulations in the previous chapter, until sufficient influence from the transition layer caused a new transition in the expressed representation in PFC.

5.3.1 Parameters

Noise was drawn randomly from a normal distribution with mean $\mu = 0$. To examine the effects of increasing levels of noise, we tested the model at 9 levels of noise. These levels corresponded to the variance σ of the noise distribution, which took values of 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75 and 1. As in the previous chapter, 200 trials were run for each of the tea- and coffee-making tasks at each level of noise. Noise was applied to the activation a of each transition node at each simulation timestep. All other parameters and weights remained consistent with simulation 4.5.1.

5.3.2 Coding of action errors

Where sequences were executed with *only* subtask based errors - i.e., the omission, perseveration or intrusion of full subtasks - these were extracted by an automated process. However, in order to examine the prevalence of each type of single action error, we chose to examine a sample of 100 trials at each noise level by eye. While automatic extraction of single action errors was trivial at low levels of noise, this manual approach was taken due to the complexity of model behaviour at high noise; attempts at automated classification of the resulting errors resulted in some error mis-classification. Examples of typical erroneous sequences at various levels of noise are given in table 5.1 to illustrate the increasing complexity of output with increasing noise and our typical error categorisations.

$\sigma = 0$	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 1$
pick-up spoon	pick-up spoon	pick-up spoon	pick-up spoon
scoop coffee	scoop coffee	scoop coffee	scoop coffee
pour-into cup	pour-into cup	pick-up kettle	pick-up spoon
pick-up kettle	pick-up kettle	pour-into cup	scoop coffee
pour-into cup	pour-into cup	pick-up spoon	pick-up spoon
pick-up spoon	pick-up spoon	pick-up spoon	pour-into cup
scoop sugar	pick-up milk	pick-up spoon	pick-up kettle
pour-into cup	pour-into cup	scoop sugar	pour-into cup
pick-up milk		scoop sugar	scoop sugar
pour-into cup		scoop sugar	pour-into cup
		pour-into cup	

Table 5.1: Correct coffee-making sequence ($\sigma = 0$) compared with examples of erroneous sequences at low, medium and high levels of noise. $\sigma = 0.2$: This example includes the initiation and abandonment of the **add sugar** subtask, resulting in the classification of an omission error (sugar is not added to the cup), and an independent action (the spoon is picked up but not used). $\sigma = 0.5$: This example shows the successful execution of the **add sugar** subtask (blue), but this execution includes two perseverations of both the **pick up spoon** and **scoop sugar** actions in the process. Additionally, the **add coffee** subtask is not completed (red), giving an omission error and an independent action. Finally, the **add milk** subtask is omitted in its entirety, resulting in a second omission error. $\sigma = 1$: Here, both the coffee and the sugar were in fact successfully added to the cup, though in a disorganised manner, with the former including perseverations (blue), and the latter achieved through an object substitution (green), using the kettle to scoop the sugar. Again, the **add milk** subtask is omitted altogether.

5.3.3 General error profile

Figure 5.1 shows the total number of correctly performed trials for each level of noise for the tea- and coffee-making tasks. Notice that, for the coffee-making task, the model has spontaneously chosen to perform either **add milk** or **add sugar** subtasks first with approximately equal frequency. Notice also that the model is fairly robust, with substantial numbers of errors only occurring at $\sigma \geq 0.2$, and that increasing numbers of errors appear gradually with increasing noise, demonstrating a graceful degradation of behaviour.

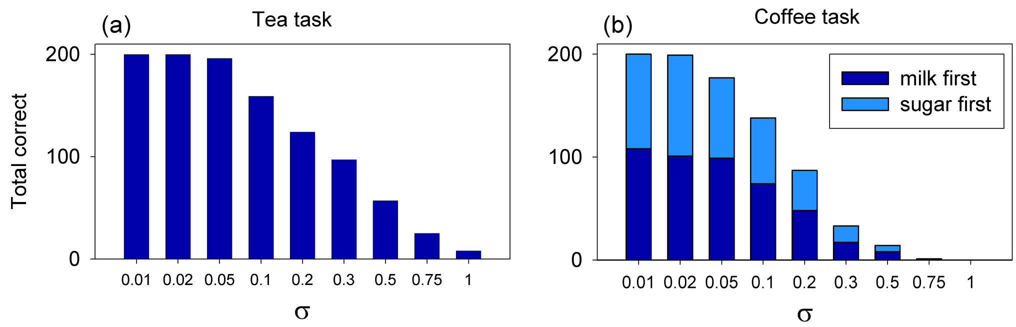

Figure 5.1: Total number of correctly performed trials at each level of noise for (a) tea- and (b) coffee-making tasks, out of 200 trials run at each noise level.

5.3.4 Normal performance

To enable a consistent comparison with previous work (Botvinick & Plaut, 2004), we divided the results into categories of normal and impaired performance based on the mean number of errors per trial; performance was classed as impaired if more than 0.5 errors per trial were observed on average, again based on a sample of 100 trials at each level of noise. We found that noise levels of $\sigma \geq 0.2$ produced impaired behaviour for both tasks.

Also consistent with Botvinick & Plaut (2004), for an initial analysis we examined the rates of subtask based errors and single action errors appearing at each level of noise, as defined above (section 5.2.1). As discussed earlier, Reason's diary studies (1979, 1984) indicated that subtask based errors are dominant in normal performance, whereas the commission of single action errors is more likely in impaired populations (Schwartz et al., 1991). As such, where performance was classified as 'normal' on the basis of the total number of errors made ($\sigma < 0.2$), we expected to see predominantly subtask based errors. Indeed, this pattern was observed. Figure 5.2 shows the distribution of subtask based and single action errors across noise levels. Note that trials containing *only* subtask based errors comprise around half of all erroneous trials up to $\sigma = 0.2$. At greater levels of noise, trials containing only subtask errors comprise a decreasing percentage of the overall number of erroneous sequences.

At levels of noise resulting in unimpaired performance, most errors consisted of the per-

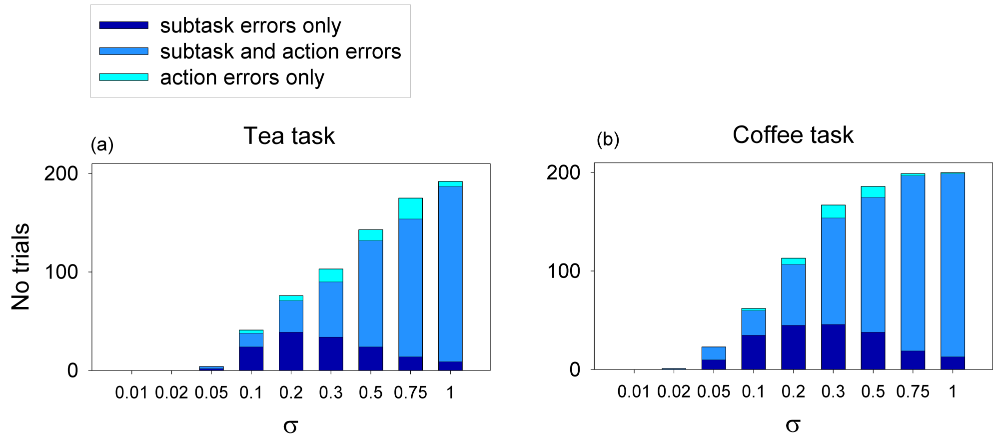

Figure 5.2: Number of trials displaying subtask based errors only, subtask based and single action errors, and single action errors only, at each level of noise for (a) tea- and (b) coffee-making trials.

severation, omission or intrusion of full subtasks, consistent with observations of 'action slips' performed at branch points in the sequences (Reason, 1979, 1984). The particular instances of these errors were in fact rather stereotypical; on those tasks where only subtask based errors were made, and no single action errors observed, several instances of specific erroneous sequences were performed. Figure 5.3 shows the distribution of these particular sequences for tea- and coffee-making tasks. These sequences tended to include adding milk to the tea (a subtask intrusion from the coffee-making task), forgetting or repeating the sugar subtask in either of the tasks, or forgetting the milk subtask in the coffee task. It is notable that repetitions of whole subtasks in these stereotypical sequences tended to be of the recurrent type, where a subtask was repeated after an intervening subtask (Sandson & Albert, 1984); specifically, instances of **add sugar** → **add milk** → **add sugar** were commonly observed in both tea- and coffee-making tasks. The significance of these common error types is discussed in section 5.3.6.

5.3.5 Impaired performance

At levels of noise resulting in impaired performance, the model performed in excess of 0.5 errors per trial, on average, based on a sample of 100 trials at each noise level. As discussed above, a primary feature of impaired performance, particularly with regard to ADS, is the frequent performance of single actions which do not clearly contribute to the completion

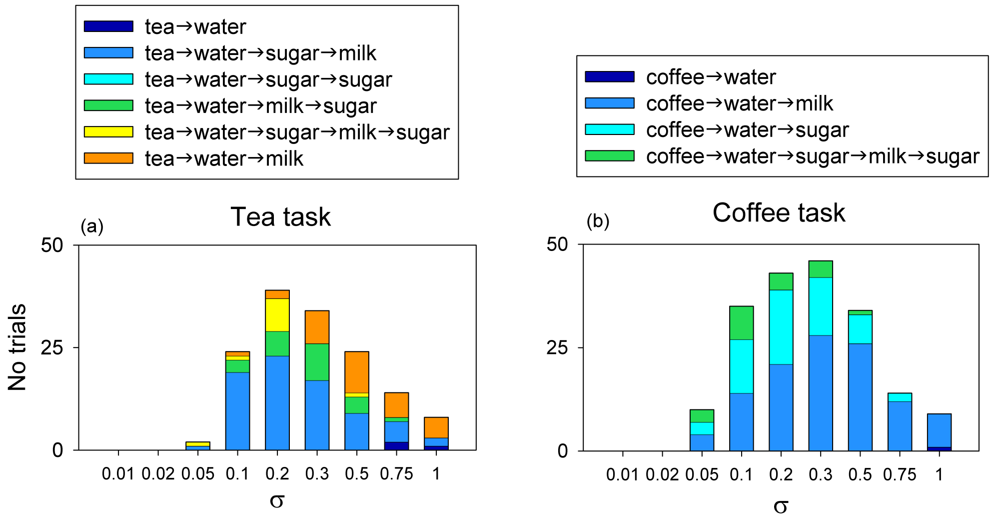

Figure 5.3: Rates of stereotypical erroneous sequences containing *only* subtask based errors. (a) tea-making trials; (b) coffee-making trials. Keys denote sequence performance at the subtask level; thus the first entry, 'tea → water', refers to trials where the **add teabag** and **add water** subtasks were performed correctly, but the **add sugar** subtask was omitted in full. Comparison with figure 5.2 indicates that the majority of trials showing subtask errors only (dark blue in figure 5.2) may be accounted for by these stereotypical erroneous sequences.

of the overall task or component subtasks. We showed in 5.3.4 that this class of error is indeed dominant in our model at medium/high levels of noise, both for tea- and coffee-making tasks. We now turn to a discussion of the particular types and frequencies of errors made by the model, and their implications for understanding the cognitive representation of sequences in healthy and impaired populations.

Independent actions

The occurrence of single or independent actions outside the context of complete sequences or subsequences has been described as a 'general fragmentation of behaviour' (Botvinick & Plaut, 2004). This is generally regarded as an important component of ADS. For example, Schwartz and colleagues (Schwartz et al., 1991, 1995) described the performance of two patients, HH and JK, classified as suffering from ADS after brain damage, including to the frontal lobes. Patient HH performed up to 31% independent actions that were not clearly contributing to the completion of the assigned task (Schwartz et al., 1991); patient JK performed around 16% such actions (Schwartz et al., 1995). We saw similar levels of fragmentation, with independent actions accounting for 15-33% at the two highest levels of noise.

(a)

This image has been removed from the electronic version of this thesis due to copyright restrictions.

The original image may be found in figure 2C of Schwartz, M., Reed. E., Montgomery, M., Palmer, C., and Mayer, N (1991). The Quantitative Description of Action Disorganisation after Brain Damage : A Case Study. *Cognitive Neuropsychology*, 8(5), p381-414

(b)

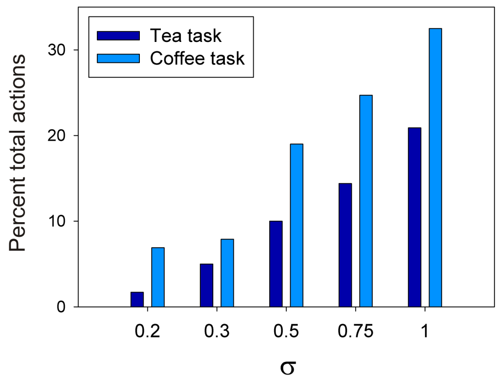

Figure 5.4: (a) Independent actions as a proportion of all actions performed by patient HH over several sessions. The proportion of independent actions produced decreased over time with HH's recovery (adapted from Schwartz et al, 1991; with permission). (b) Independent actions as a percentage of all actions performed in the current simulations, based on 100 trials at each noise level. The increasing proportions of disjointed actions with increasing noise reflects the trend in (a) suggesting increasingly fragmented behaviour with increasing impairment.

In patients, the proportion of independent actions performed was variable across tasks, and in the case of patient HH, was reduced with practice and recovery (Schwartz et al., 1991), suggesting a general vulnerability to fragmentation with increasingly severe impairment. Correspondingly, we saw a general increase in independent actions as noise increased; figure 5.4(b) illustrates this trend for both tasks. The nature of the individual single action errors comprising these independent actions are now examined in detail.

Summary of error rates

The error rates produced by the model for the error types discussed in section 5.2.1 are summarised in table 5.2. These rates were based on a sample of 100 trials at each level of noise. Note that not all errors produced by the model are explicitly represented in this table. Those not included here were predominantly independent actions which were not further classifiable as a sequence, perseveration, addition or semantic error; such as occasions when subtasks were abandoned part-way through. As these actions did not contribute to the completion of the task, we did consider them to be erroneous, however, we do not examine them as a distinct error category here, as their primary effect - contributing to the 'general fragmentation of behaviour' - has already been considered above.

	Omission	Sequence i	Sequence ii	Sequence iii	Sequence iv	Addition	Semantic
Tea	29%	20%	4%	5%	5%	26%	2%
coffee	42%	29%	3%	8%	10%	0%	2%

Table 5.2: Summary error data showing mean error rates for noise levels $\sigma \geq 0.2$. These levels of noise resulted in impaired performance, and error rates displayed here were based on a sample of 100 trials at each noise level. Summaries describing error types are given in section 5.2.1.

The general trends we observe in error types are reflective of the human data in several ways. In particular, the graph in figure 5.5 highlights the overall trends in relative error rates at our two highest levels of noise (analogous to more severely impaired patients) across the four main error types, and their similarity to trends observed in patient data from Humphreys & Forde (1998) and Schwartz et al. (1998).

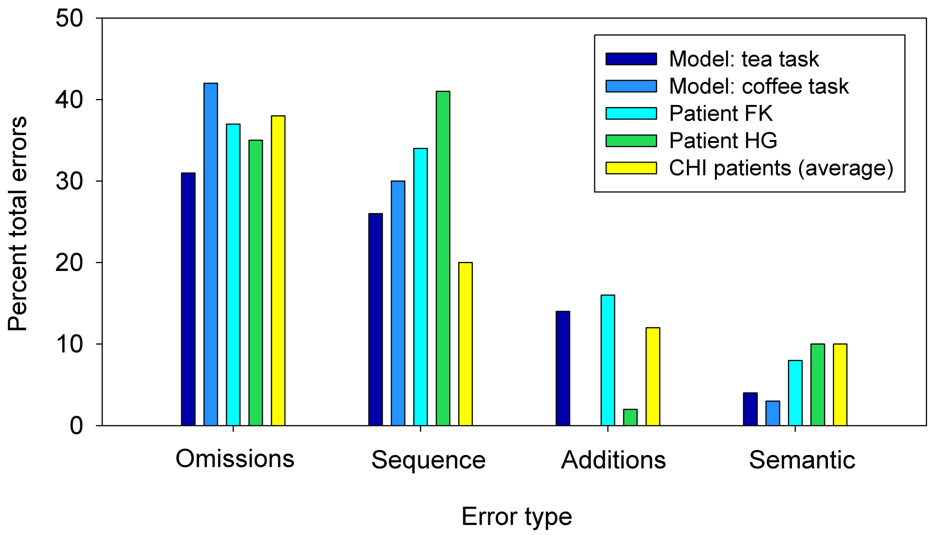

Figure 5.5: Comparison of mean error rates for the current model across $\sigma = 0.75$ and $\sigma = 1$, in tea- and coffee-making tasks, against existing patient data. Here, all 'sequence' errors are plotted according to view i, including perseverations. Data for patients FK and HG from Humphreys & Forde (1998; table 4, p789); data for closed head injury (CHI) patients ($n = 30$) from Schwartz et al. (1998, table 6, p19).

Omissions & the omission rate effect

Omissions have generally been reported to account for around 30-40% of errors in ADS, which is a relatively consistent finding across labs, patients and - with certain exceptions - tasks (Schwartz & Buxbaum, 1997; Buxbaum et al., 1998; Schwartz et al., 1998; Humphreys & Forde, 1998). Moreover, omissions tend to account for the single most common error type committed in ADS (though see discussion on sequence errors below). An 'omission rate effect' has also been noted (Schwartz et al., 1998; Morady & Humphreys, 2009), whereby omissions in patients tend to account for a relatively higher proportion of all errors with increasing disorder severity.

Omission rates in our simulations showed a strong resemblance with those rates observed in ADS patients (see figure 5.5). We found 31% and 42% of all errors were omissions in the tea- and coffee-making tasks, respectively, averaged over the two highest levels of noise. Also, from table 5.2, it is clear that omissions outnumbered any other single error type, for both tea and coffee tasks. Though only mean rates are displayed in the table 5.2, this was the case at all noise levels that resulted in impaired performance.

Figure 5.6(a) shows the number of omissions made in our sample for the tea- and coffee-making tasks, divided by the total number of subtasks in the task. This value was examined in order to account for the number of opportunities for an omission in each task. We found that the coffee-making task was far more prone to omission errors than the tea-making task, even when taking into account the additional subtask involved for the coffee task. Examination of example trials suggested that this discrepancy was due to the flexibility involved in the ordering of the **add milk** and **add sugar** subtasks in the coffee-making task. This flexibility resulted in a greater opportunity for the model to become confused as to its action history within a single trial, essentially 'forgetting' that it had not previously added an item. The mechanisms underlying this effect are discussed in detail in section 5.3.6.

Given that fewer omissions were seen at all levels of noise for the tea-making task, to illustrate relative omission rates, we created a standardised omission rate measure. We took the proportion of omissions at the lowest impaired noise level ($\sigma = 0.2$) as a baseline omission rate, and divided the omission rate at each noise level by this baseline rate. Thus, at noise $\sigma = 0.2$, the standardised omission rate for both tasks was 1. This allowed us to clearly examine the magnitude of the omission rate increase relative to each task's *own* baseline omission rate. These standardised omission rates are illustrated in figure 5.6(b). Intriguingly, while the total number of omissions increased with noise for both tasks, we only found a trend for an increasing omission *rate* for the tea-making task. In contrast, the coffee-making task showed relatively constant omission rates across all levels of noise, though a slight trend towards a decrease was observed. This finding was unexpected given previous reports of a strong omission rate effect (Schwartz et al., 1998; Morady & Humphreys, 2009).

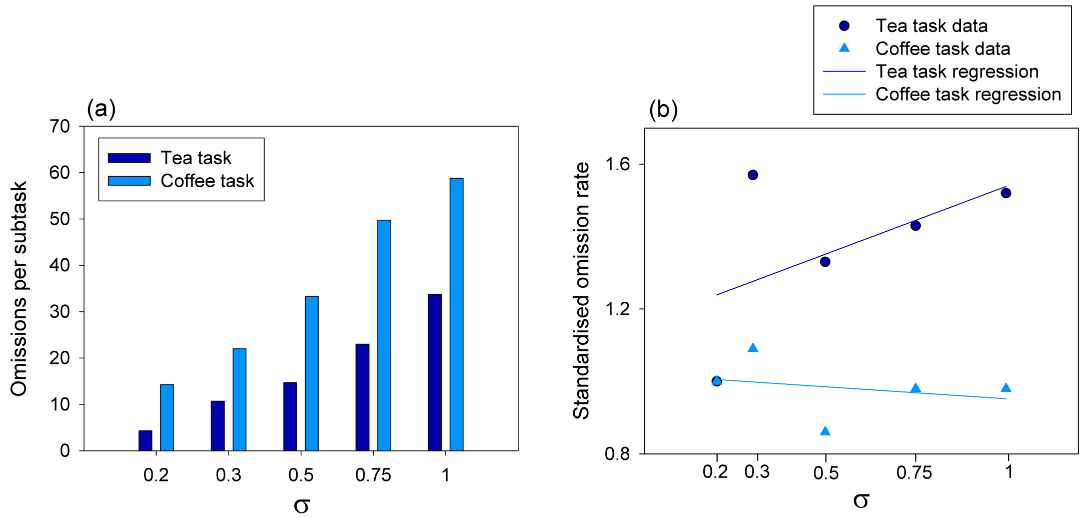

Figure 5.6: (a) Comparison of average omissions per subtask and (b) standardised omission rates (points) and linear regressions thereof, for tea- and coffee-making tasks. Linear regression analyses showed a trend towards an increasing omission rate for the tea-making task only. While this was not significant, very few data points significantly reduced the power of this analysis, so this result should be interpreted with caution.

The lack of a strong omission rate effect, however, appeared to be due to the dominance of subtask based errors at low levels of noise, coupled with the opportunities for intrusion errors in each of the two tasks. Many of the erroneous tea-making trials at low noise, for example, involved intrusions of the **add milk** subtask. Such an intrusion was representative of an action slip, due to its strong dependence on context: both the **add water** and **add sugar** subtasks may be correctly followed by the **add milk** subtask in the coffee-making task. As both the **add water** and **add sugar** subtasks are also involved in the tea-making task, this results in two extremely similar contexts in which the **add milk** should and should not be performed. A small amount of disruption is then sufficient to confuse these very similar contextual representations and result in an intrusion error. However, there is no equivalent opportunity for intrusion from the tea- to the coffee-making tasks. As a result, such intrusions are not observed in the coffee-making trials, resulting in a relative over-representation of omissions at lower levels of noise. This effect is not constant over all noise levels, however, due to the infrequency of intrusion errors with increasing noise in the tea-making task (see ‘additions’ below). This pattern implies that with suitable opportunity for such errors, an omission rate effect might also be observed in the coffee-making task.

Sequence errors

Our sequence error rates according to each of the alternative views outlined in section 5.2.1 are summarised again in table 5.3. It is clear from this summary that perseverations, particularly of single actions (included in view i, but not ii, iii or iv) contribute a great deal to the overall number of sequence errors. This is consistent with data presented by Humphreys and Forde (Humphreys & Forde, 1998; Forde & Humphreys, 2002), which demonstrated that perseverations accounted for a greater proportion of all errors than did sequence errors as defined by view iii.

	view i	view ii	view iii	view iv
mean tea	20%	4%	5%	5%
mean coffee	29%	3%	8%	10%

Table 5.3: Summary sequence error data for different classifications, as described in section 5.2.1.

Note that we roughly replicate the data reported by Schwartz et al., (1998), where perseverations were included in the general analysis of overall error proportion (see figure 5.5). In this analysis, sequence errors were reported to account for around 20% of all errors in patients; here we see rates of 26% and 30% for the tea- and coffee-making tasks across the two highest noise levels, using the same classification.

Humphreys & Forde (1998) reported sequence error rates of 16% and 10% for patients FK and HG according to view iii, which did not include perseverations; we saw lower rates of 5% and 9% for the tea- and coffee-making tasks, respectively. Of note, however, is that when sequence errors and perseverations were considered together (consistent with view i), rates for FK and HG increased to 34% and 41% respectively, which accounted for similar or even greater proportions of errors than omissions, in contrast to the commonly reported finding that omissions tend to account for the majority of error types. This, in combination with the very different sequence error rates arising from the four different views presented above, indicates that common findings regarding relative rates of sequence and omission errors are not in fact consistent.

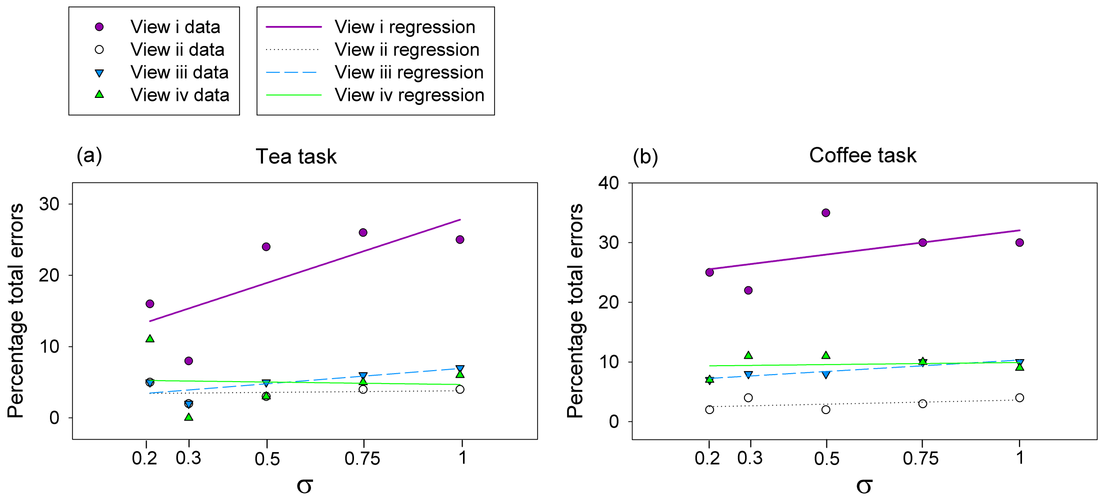

Figure 5.7: Sequence errors as a percentage of all errors according to the four different views discussed above, for (a) tea and (b) coffee trials. Linear regression analyses indicated a generally consistent sequence error rate, with the exception of view i. A slightly increasing error rate here is due to the inclusion of single action perseverations in this view, as discussed in the text.

Notably, while the total number of sequence errors produced increased with noise, we tend not to see sequence errors increasing as a *proportion* of all errors (figure 5.7), consistent with previous findings (Schwartz et al., 1998). A slight increase was seen for view i however, resulting primarily from increased perseverations of single actions at high noise.

These contrasting results highlight that generalisations about sequence errors against which model data has been compared (Botvinick & Plaut, 2004) are not necessarily informative, and greater consideration should be given to the component error types contributing to this category, particularly perseverations. This is discussed further with respect to the particular mechanisms underlying sequence errors in section 5.3.6.

Additions (intrusions)

As mentioned earlier, while previous definitions of 'addition' errors appear to have encompassed a broad range of errors, precisely what is meant by an addition has not been well defined. As such, we include only intrusion errors in this category in the present study. At higher levels of noise, we saw rates of additions similar to those reported in the previous literature (figure 5.5). For the tea-making task, over the two highest noise levels additions accounted for 14% of all errors; Humphreys & Forde (1998) reported rates of additions at around 16% for patient FK (though patient HG made fewer additions, at around 2%). Similarly, Schwartz and colleagues reported an average of 12% additions in their patient population (Schwartz et al., 1998).

We observed a tendency for a decreasing proportion of addition errors to be observed with increasing noise (see figure 5.8), where additions accounted for up to 34% and 45% at $\sigma = 0.2$ and $\sigma = 0.3$. Such a trend has not, to our knowledge, been reported in previous studies. Notably, additions of entire subtasks comprised a greater proportion of the total additions at lower noise. Such additions predominantly consisted of performing the **add milk** subtask during the tea-making task. Occasionally, additions involved the adding of coffee grounds in the tea task; this usually occurred where the **add sugar** subtask was expected, and coffee grounds were scooped instead of sugar. Intrusions of intact subtasks in this manner are relatively common in healthy participants, thus it is not surprising that they comprise a larger proportion of errors at low noise. The proportion of single action additions is, however, more consistent across noise levels, and may be more representative of the broader, but less well defined 'additions' recorded in human behaviour studies.

In contrast to the data, we saw no additions in the coffee-making task. However, this is again a result of the dependence on context for additions, as discussed above with respect to the lack of omission rate effect for the coffee-making task. There are far fewer opportunities for intrusion errors in the coffee-making task, where only actions from the **add teabag** subtask might be inserted. Additionally, these actions are performed in a rather different context to any seen during the coffee-making task, significantly reducing the chances that such additions will occur.

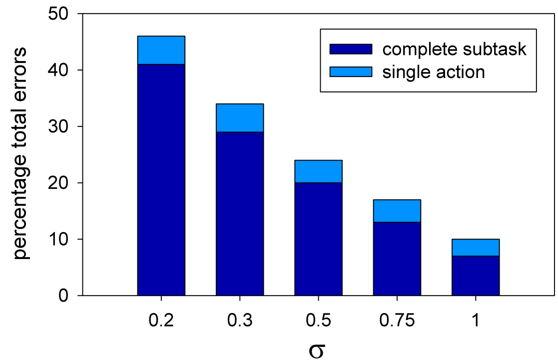

Figure 5.8: Additions of whole subtasks and single actions as a percentage of all errors. Results correspond to the tea-making task only, as no additions were observed in the coffee making task.

It would be of interest to examine patient performance of two very similar tasks - such as tea- and coffee-making - to observe the relative rates of addition or intrusion errors compared with existing studies, which have conversely focused on the performance of unrelated tasks. While the effects of the presence of distractor objects have been studied in previous work (Humphreys & Forde, 1998; Morady & Humphreys, 2009), the performance of similar tasks has not, to our knowledge, been examined. We would predict that intrusions would be far greater in frequency in such a study, particularly with less severely impaired patients, than other action additions, which may consist of anomalous actions not drawn from any particular task. Indeed, Forde et al. (2004) found that patient FK did make a number of intrusion errors which were triggered by task context. Interestingly, the presence of semantically related distractor objects did *not* worsen patient performance (Forde & Humphreys, 2002), suggesting any intrusions or additions are not necessarily a result of sensory, 'bottom-up' interference as has been previously suggested, but related more to overall task context, consistent with our present results.

Semantic errors

While the model as it stands includes no representation of semantic similarities between objects, and as such there is little capacity for true semantic error, we did observe occasional errors of this apparent nature. We saw a number of object substitutions, predominantly involving scooping the sugar with the wrong 'container' - either the kettle or the milk bottle rather than the spoon. We also observed occasional object omissions, where, for example, the contents of the kettle were poured but not directed into the cup.

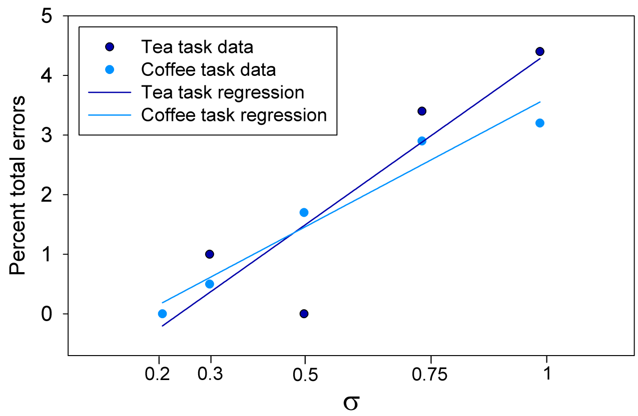

Figure 5.9: Semantic errors as a percentage of all errors for tea and coffee simulations. Linear regression analyses demonstrated a slight increase with increasing noise, but note low overall rates generally.

These errors showed a similar pattern across tea- and coffee-making tasks, occurring at low rates and showing a slightly increased occurrence with increased disruption. However, they accounted for a smaller amount of our total errors than seen in experimental papers (e.g., Humphreys & Forde, 1998). There are several possible reasons for this. As mentioned, we do not include any representation of semantic similarity between distinct objects, nor do we include any active role of object 'salience' on fixation; rather, fixation is guided only by the current PFC representation. A more active sensory influence might result in higher rates of semantic errors, as a result of 'bottom-up' influences or biases on *fixation*, rather than on action. Furthermore, each of our affordances was of equal strength; the inclusion of affordances of different strengths would be expected to increase rates of semantic errors. For example, in the case of a weak influence from PFC an affordance might 'override' any conflicting task related information (simulation 4.5.3).

The relationship between semantic knowledge and semantic error commission is not straightforward, where semantic errors may be common even where semantic knowledge is intact (Schwartz et al., 1991, 1995), or rare where semantic knowledge is impaired (Forde & Humphreys, 2002). Neither is the relationship between semantic error rate and task congruent distractors a simple one; semantic error rates have been shown to *decrease* in the presence of distractor objects (Forde & Humphreys, 2002). Accordingly, our results suggest that semantic errors have multiple possible root causes; clearly such errors do not always arise as a result of the semantic similarity between specific objects - if so these errors would not occur in our model. The fact that we observe such errors at all indicates that these errors may arise from an alternative mechanism, namely semantic similarity between *contexts*. The infrequent nature of these errors in the present model, however, suggests that disruption of sequence knowledge - causing a confusion between *stages* of the task - is unlikely to be the only mechanism involved.

5.3.6 Mechanisms underlying errors

We expected that the predominant mechanism resulting in errors in this simulation would be the confusion of temporal order knowledge caused by activation of erroneous transition nodes due to noise. In many cases, particularly at low levels of noise, this was indeed the case. However, more subtle effects were also observed. This resulted in three primary mechanisms causing erroneous performance. Importantly, these mechanisms tended to underlie different types of error. Here we outline these three most common causes of erroneous performance.

Confusion of context

Contextual confusion, akin to the mechanism for error described by Botvinick & Plaut (2004), was the primary mechanism for most subtask based errors. Examination of transition node activity illustrates this general point. Each representation illustrated in figure 4.1 had a corresponding transition node, which, when activated, provided excitation to that representation in PFC. Thus, transition nodes T1-T7 provided excitation to tea-making representations T1-T7 in figure 4.1, respectively. Transition nodes C1-C10a and C1-C10b conversely activated the corresponding coffee-making representations. In a successful trial, then, we would expect to see consecutive activation of the transition nodes corresponding to each relevant PFC representation as the model progressed through the task (see, for example, figure 4.6(b)).

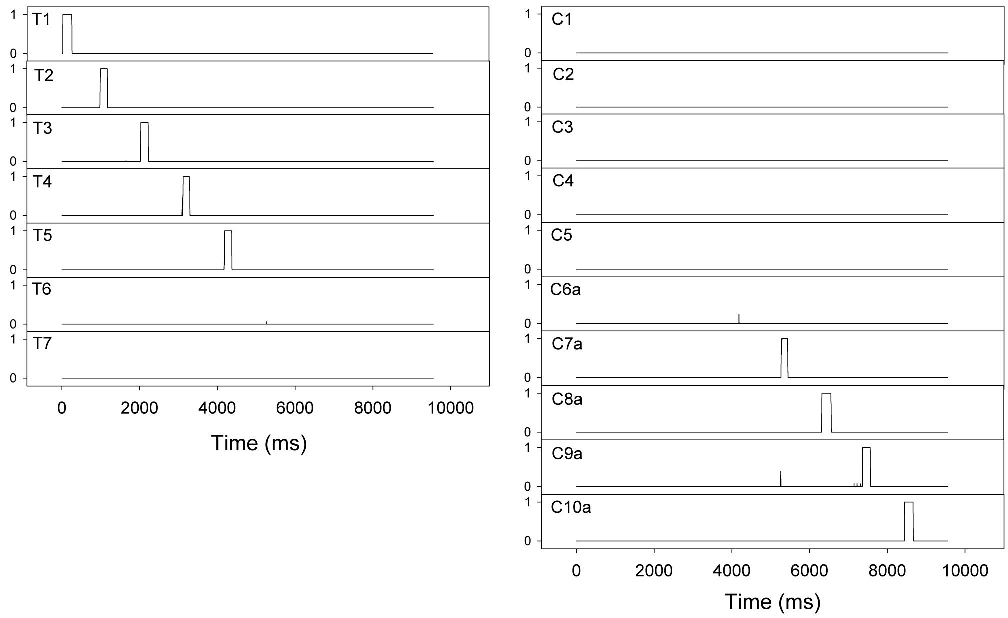

Figure 5.10: Transition node activity during a tea-making trial in which milk was added. Noise causes erroneous activation of transition node C7a rather than T6, triggering a contextual representation from the coffee-making task. The sequence proceeds from this point as if performing the coffee-making task.

The time course of transition node activity during a trial in which an intrusion error was committed (milk was added to the tea) is illustrated in figure 5.10. In this figure we can see the expected pattern of activity for the first 5 stages of the task, where transition nodes T1-T5 are activated in sequence. However, where we would expect to see activation of transition node T6 (corresponding to representation T6, which triggers a **scoop sugar** action), instead we see activation of *coffee*-making transition node C7a, causing the activation of representation C7a in PFC.

Importantly, both transition nodes T6 and C7a activate PFC representations which correspond to the performance of the same action, **scoop sugar**. However, activation of C7a results in a different overall representation of *context*; the model now has an internal representation corresponding to the coffee-making task. This has a temporally delayed effect on outward behaviour. The sequence is continued as if it were the coffee-making task; more specifically, an instance in which the sugar is added before the milk. As a result, on the completion of the **add sugar** subtask, activation of the **add milk** subtask is triggered and an addition error occurs. The effect of this is analogous to 'forgetting' that the drink being prepared is tea, and adding milk as though preparing coffee. Importantly, and as observed by Botvinick & Plaut (2004), while behaviourally, an error occurs after the performance of the **add sugar** subtask, at a cognitive level, the error occurs *during* the subtask. Of note is evidence that distraction of normal participants during the performance of routine subtasks is more likely to result in error than distraction between subtasks (Botvinick & Bylsma, 2005), as the present results would suggest.

This confusion of context occurs as a result of the feature-based representations we have implemented. Each PFC representation and environment representation selectively provides excitation to the transition node corresponding to the subsequent stage of the task. Since these representations are organised in a feature-based fashion, this results in a 'gradient' of excitation provided to any transition nodes which receive activation from overlapping representations. For example, representation T5 preferentially activates transition node T6. Likewise, representation C6a preferentially activates transition node C7a. However, given the large degree of overlap between representations T5 and C6a (see figure 4.1), representation T5 incidentally provides a significant degree of excitation to transition nodes C7a. This is also true of representations of the external environment, where, for example, the representation of **black-liquid:in-cup** (see table 4.2), applicable after successful completion of the first two subtasks, provides excitation to transition nodes T6, C7a and C7b. As the transition layer approximates a WTA network, a small amount of noise added to this region is sufficient to cause the erroneous activation of transition nodes C7a, rather than the correct T6. This is what we see here.

Similar activity underlies the repetition and omission of subtasks; common occurrences at low noise were the repetition of the **add sugar** subtask after an intervening **add milk** subtask, or the omission of the **add sugar** subtask. Both of these examples were generally due to the confusion of the 'sugar first' (representations C6a-C10a) and 'milk first' (representations C6b-C10b) versions of each task, which again, shared a majority of features.

At low noise, such confusions of context are most commonly observed at points where the current representation overlaps heavily with another, as described. However, low levels of noise are insufficient to drive activation of transition nodes which receive little or no excitation from the current representation in PFC. Increasing levels of noise, however, result in a greater likelihood of confusion of more disparate contexts, where fewer features are shared. This accounts for the more fragmented nature of errors at high noise, where actions are performed outside of intact subtasks; actions in different subtasks tend to have underlying representations which differ more than those within the same subtask.

This type of activation did result in single action errors at high noise, however. A notable example was the premature termination of the task after execution of the **pick up spoon** action, resulting in both an omission error (no sugar added to the cup) and a single action error (execution of the **pick up spoon** outside of the context of an intact subtask). This was due primarily to ambiguity from the environment, where an empty spoon in the hand was consistent with having just performed the **pick up spoon** or **pour sugar** actions. This indicates the important roles of both internal representations of context as well as the current environmental state in determining the correct next action.

Uninhibited activation of transition nodes

Figure 5.11 shows transition node activity during a coffee-making trial in which the model was observed to pour water from the kettle twice in immediate succession. We might assume from the outward behaviour of the model that the observed error was the result of an ineffective transition in the PFC representation, resulting in the maintenance of the currently selected representation for a second stage of the task.

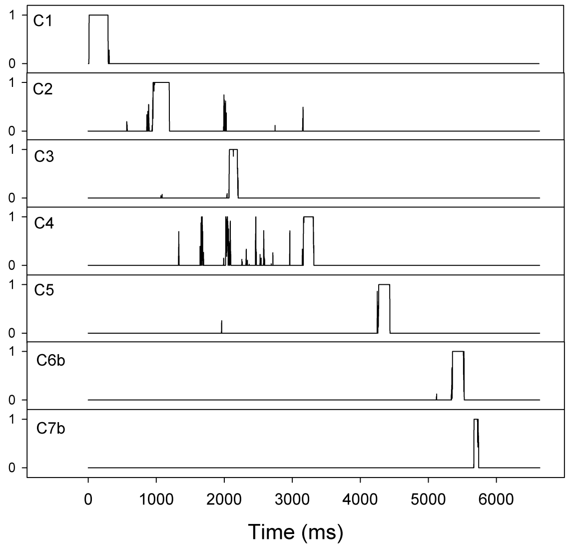

Figure 5.11: Transition node activity during a coffee-making task in which the **pour into cup** action is performed twice in immediate succession. Premature activation of transition node C7b causes rapid switching of the PFC representation, resulting in a cognitive-level omission of the action corresponding to representation C6b, and performance instead of the action corresponding to representation C7b. This is the same as the previously selected action corresponding to representation C5. This manifests behaviourally as a continuous perseveration.

Transition node activity in fact reveals a different mechanism. At around t=5500ms, we see activation of the transition node C6b, activating the corresponding PFC representation C6b, which in turn should trigger a **pick up milk** action. However, note the rapid onset of activity in transition node C7b at around t=5800ms. This premature onset of C7b, caused by the combined influence of the current PFC representation C6b and noise, results in a transition to the subsequent PFC representation C7b *before* the **pick up milk** action has been successfully executed. This prematurely selected representation C7b corresponds to a **pour into cup** action, which is duly executed by the motor loop. However, as the **pick up milk** action was not performed, the kettle remains held. The result thus *appears* to be a perseveration of the previous action at the behavioural level, but at a mechanistic level, the error bears more resemblance to an anticipation-omission error. While this mechanism did also underlie anticipation-omission errors, as well as a number of object substitutions, it is important to note that the apparent behaviour of the model was not necessarily representative of the underlying cognitive dynamics. To our knowledge, no previous modelling work has been in a position to provide similar insights regarding this apparent discrepancy between cognition and erroneous behaviour. A behavioural prediction arising from this observation is that the initiation of perseverative actions would be slightly temporally delayed; as no studies to our knowledge have noted such a phenomenon however, future observational studies would be required to explore this possibility.

Interference from simultaneously activated transition nodes

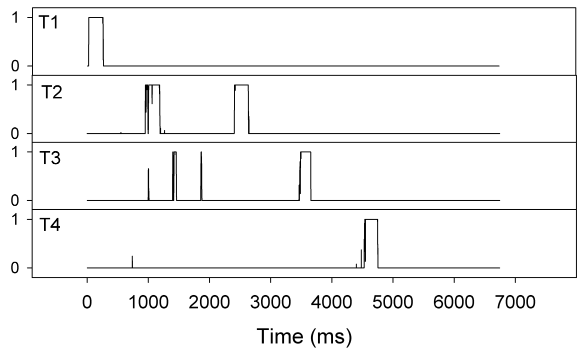

Figure 5.12: Transition node activity during a tea-making task in which the teabag is picked up twice in succession before adding to cup. This 'toying' behaviour results from inefficient activation of representation T2, which itself is caused by interference from multiple transition nodes. This leads to a 'decay' back to the previously selected representation T1, and a perseveration of its corresponding action.

A further mechanism for error is illustrated in the transition node activity in figure 5.12. In this example of an erroneous tea-making trial, the **pick up teabag** action was executed twice in succession. On this occasion, the error was caused by noisy activation of transition node T2 (triggering the representation **put (teabag) into cup**), and simultaneously, some activation of node T3 (triggering the representation **pick up kettle**). This caused a degree of interference in the signals sent to PFC, ultimately resulting in the reactivation of representation T1 (**pick up teabag**).

This reactivation of representation T1 is most likely a result of an unclear input to PFC. This noisy input causes PFC representation T2 to be insufficiently activated to compete with the currently selected representation T1. As such, representation T2 does not reach the required level of stability, and upon the removal of excitation from the transition layer, the pattern of activation in PFC decays back to the more stable representation T1. This mechanism tended only to result in immediate perseverations of single actions. It is interesting to note, however, that both this interference effect and the previously discussed premature switching are significant causes of such perseverations, which, on a behavioural level, appear to be the same error.

Rates of error

It is not immediately obvious why the particular rates of error types differ, given that all are ultimately caused by disruption to temporal order knowledge in the present study. However, we suggest that omissions and perseverations are increasingly common because multiple different errors at a cognitive level may result in these behavioural manifestations; omissions were seen to be a result of at least two of the mechanisms we describe above, and perseverations in general were seen to be a result of all three. Other errors, such as anticipation-omissions, additions, object substitutions and repeats of entire subtasks tended to be the result of just one. Again, this is a novel insight which previous modelling attempts have not explicitly revealed. Furthermore, mistakes involving full subtasks are less common with increasing noise because any misplaced subtask will be increasingly subject to errors within itself. Independent actions are also a result of this increasing difficulty in completing a single subtask.

5.3.7 Discussion

We have accounted for a surprising amount of the human observational data by the simple disruption of sequence knowledge. In particular, a tendency towards omissions and sequence errors was observed over additions and semantic errors, as has consistently been observed in previous patient studies. Figure 5.5 illustrates these qualitative similarities for our two highest noise levels to patients FK and HG (Humphreys & Forde, 1998) and closed head injury patients from Schwartz et al., (1998).

The present simulation is supportive of Humphreys, Forde and colleagues' (Humphreys & Forde, 1998; Forde et al., 2004) hypothesis that disordered sequence knowledge is responsible for many of the effects of ADS. In particular, we noticed that our data at high noise bore a particularly close resemblance to the error profiles of patient FK (Humphreys & Forde, 1998; Forde & Humphreys, 2002), for whom the explanation of disrupted temporal order knowledge was originally suggested (see figure 5.5). Furthermore, throughout the simulations we found a greater propensity for error later in the task, during performance of the milk and sugar subtasks. This was a tendency of FK which was noted by the authors.

However, there were some important differences between our data and those from human behavioural studies. For example, we saw particularly low rates of semantic errors, as well as unusually high levels of additions at low/medium levels of noise. Furthermore, due to the specific implementation of the model, we are currently unable to account for quality or spatial errors. It is possible, however, that various additions or amendments to the model would allow replication of further aspects of the data. For instance, incorporation of semantic properties of objects and their similarity might increase rates of semantic errors, particularly if additional distractor objects were included. Explicit modelling of orienting actions and object salience might affect the type and frequency of addition and substitution errors due to more explicit sensory influences. Likewise, an inclusion of a representation of spatial information would introduce the opportunity for spatial error. A replication of the current simulation in a more comprehensive model would begin to shed light on these questions; however, whether the disruption of temporal order knowledge alone would result in expected rates of these error types remains to be seen.

Continuum of disruption or multiple deficits?

In section 5.2, we questioned whether the symptoms of ADS and the action slips common in normal performance might be due to the same underlying deficit. We suggest that, in certain cases, ADS may be considered an extreme version of the effects of disruption in healthy controls, given that the disruption of temporal order knowledge results in both action slips at low noise and more severe fragmentation at high levels of noise.

However, some of the errors made at high noise were due to certain mechanisms that we rarely observed at low noise (interference and premature switching). Thus, while some of the deficits of ADS may be accounted for by the same mechanisms observed in normal performance, other mechanisms may also be at work, albeit as a result of the same root deficit in both cases, this being the enhancement or attenuation of sub-threshold activation in the transition layer by noise. Additionally, given that the model did not produce all error types observed in human behavioural studies, we remain unable to comment on the significance of these errors for understanding the organisation of sequential action in healthy and impaired populations.

5.4 Disruption of schemas

In the second half of this study, we attempted to simulate the disruption of action schemas themselves, rather than the knowledge of their temporal order. As indicated throughout this and the previous chapter, we have conceived of our stable PFC representations as schemas. To simulate their disruption, addition of noise was unsuitable, as the intrinsic stability of the representations meant extremely high levels of noise were required to destabilise them. As an alternative, we lesioned a randomly selected proportion of intrinsic PFC connections. The proportion of lesioned projections denoted the severity of the disruption. We ran seven simulations of 200 trials each for both tea- and coffee-making tasks, lesioning 1, 2, 5, 8, 10, 12 and 14% of randomly selected intrinsic PFC connections, respectively.

5.4.1 General error profile

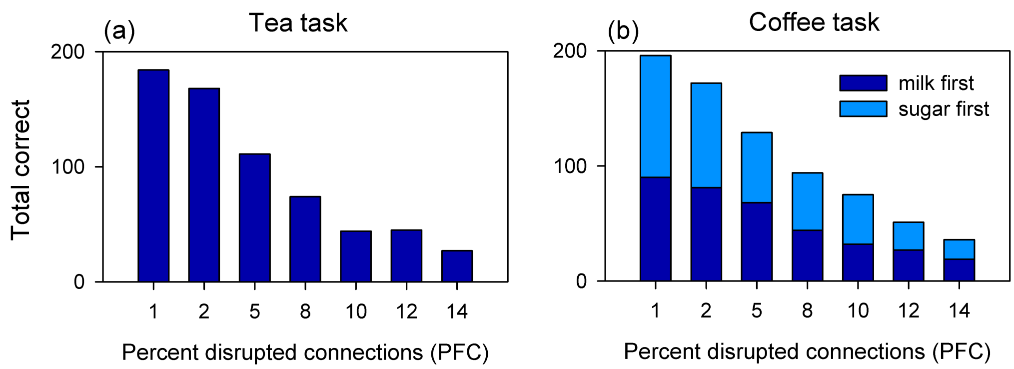

Figure 5.13: Total number of correctly performed trials for the (a) tea- and (b) coffee-making tasks, out of a total of 200 trials at each level of disruption.

This simulation again resulted in a fairly gradual increase in the number of erroneous trials with increasing disruption, as shown by the number of correct trials at each level of disruption illustrated in figure 5.13. However, the present study gave rise to a significantly different profile of errors from that which was observed under disruption of temporal order knowledge. Whereas before, we classified 'normal' performance as that which resulted in a mean of less than 0.5 errors per trial, the current simulation tended to produce error rates much higher than this even at low levels of disruption. The tea-task, for example, averaged 1.04 errors per trial at just the lowest disruption level (based on a sample of 100 trials taken from the tea-task simulation). This implies that for the tea-making task, even at levels of disruption which result in predominantly correct performances of the sequence, performance is still 'impaired'. The coffee task, in contrast, appeared to be generally more robust, only showing errors of more than 0.5 per trial at 5% of connectivity lesioned or more. However, those errors that were performed at lower levels of disruption tended towards single action errors, rather than the subtask based errors we would expect of normal performance, and as was observed in the previous simulation, suggesting that the current manipulation cannot capture the effects of distraction on normal performance.

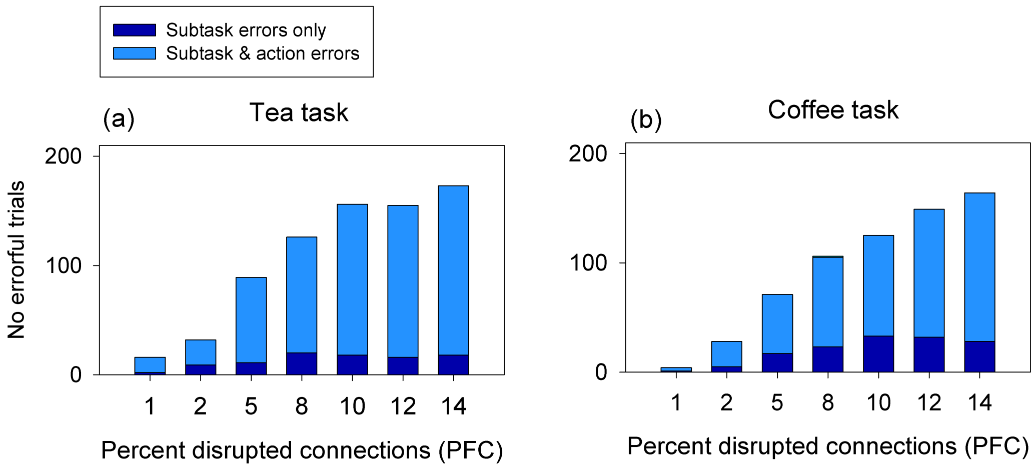

Figure 5.14: General error profile for tea (left) and coffee (right), showing the number of trials displaying only subtask based errors, and those with subtask based and single action errors. Note the profile is dominated by trials with both subtask and single action errors at all levels of disruption.

The high average number of errors per trial at low disruption for the tea-making task is surprising, given the general performance profile in figure 5.13, which indicates a small number of errorful trials. Examination of figure 5.14, however, shows that trials containing single action errors dominate at all levels of disruption. Trials containing single action errors tend to display a greater overall number of errors than those trials displaying only subtask errors. Indeed, examination of a sample of 100 trials revealed a small number of highly error-prone trials at low disruption.

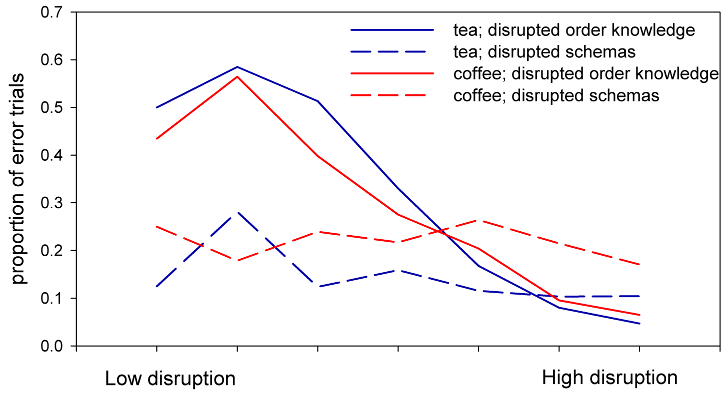

Figure 5.15: Proportion of error trials displaying just subtask based errors over all levels of disruption, compared with subtask based error rates for the simulation presented in section 5.3. Data from the previous simulation was taken from noise conditions $\sigma \geq 0.05$, as lower noise resulted in virtually no errors.

Moreover, examination of the relative rates of trials with subtask based errors only indicates no significant peak of subtask based errors at low/medium levels of noise, as we found in the previous simulation, and as would be suggested by the human behavioural data. Rather, the proportion of erroneous trials displaying only subtask based errors remained relatively constant across all levels of disruption, and did not show the distinctive decrease with increasing noise as observed in the previous simulation. Furthermore, at no disruption level did the number of trials displaying only subtask based errors outnumber or even equal those in which single action errors were also made. This stood in stark comparison to the previous simulation in which, at low levels of noise, trials containing only subtask based errors comprised up to 58% of erroneous trials. In the current simulation, the maximum proportion of erroneous trials comprising subtask only errors was just 28%. This markedly different error profile is clear in graph 5.15, which compares the proportion of subtask based error trials from the seven highest levels of noise from the previous simulation with the current simulation. This indicates the noticeably different subtask error rates resulting from the different types of disruption in the two simulations. Whereas the previous simulation accounted for 'normal' performance well, the result presented here suggests that the disruption of action schemas is unable to account for normal, everyday slips of action.

5.4.2 Impaired performance

Brief examination of model output at levels of disruption resulting in impaired performance indicated that a strong tendency to perform many continuous perseverations of single actions accounted for a large majority of the total errors committed, and was chiefly responsible for the high numbers of total errors at low levels of disruption. A more detailed examination and classification of a sample of 100 trials at disruption levels of 2, 8 and 14% indicated that perseverations consistently accounted for around 70-80% of all errors for the tea-making task, at low, medium and high levels of disruption. Omissions, in contrast, accounted for around 10%; where omissions did occur, many of these appeared to be the result of the model getting 'stuck' repeating the same action and becoming unable to complete the subtask. Perseverations and omissions together tended to account for approximately 90% of all errors across all levels of disruption. Sequence errors other than perseverations accounted for a very small proportion of all errors, averaging around 1-3%. These trends are illustrated in figure 5.16(a).

The coffee-making task did not exhibit quite as strong a tendency to perform continuous perseverations, though these were more strongly represented here compared to the disruption of temporal order knowledge. Here, they made up to 58% of all errors at the highest level of disruption, in contrast to 30% omissions. These rates were consistent across 'impaired' conditions - here, we sampled at ***5***, 8 and 14% disruption. This is illustrated in 5.16(b). Note, however, that low levels of noise resulting in 'normal' performance for the coffee task showed a different profile, dominated by omissions (not illustrated). It is unclear why the tea- and coffee-making tasks have shown such different profiles in this simulation, particularly at low levels of disruption. However, we propose that it is due to the varying stability of the underlying representations required for each task, which we explore in detail in the following section.

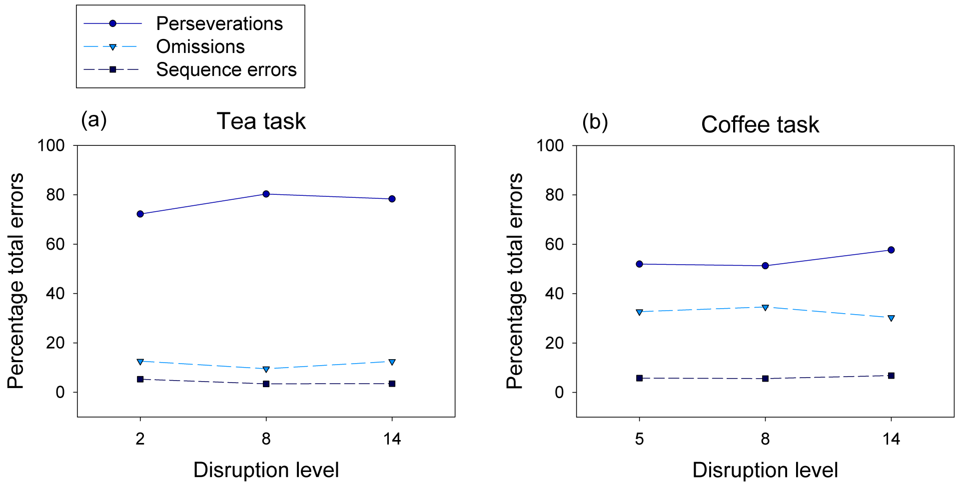

Figure 5.16: Error rates at disruption levels resulting in impaired performance were consistent across increasing disruption for both (a) tea- and (b) coffee-making tasks. Note however, lower disruption was required to result in impaired performance for the tea-making task.

Given the clear mismatch of the results of this simulation with the human behavioural data, for both tea- and coffee-making tasks, and the strong tendency to perform perseverations above other errors, we do not give an in depth discussion of the relative rates of each error type here, but focus instead on the mechanisms underlying the high rate of perseveration and its implications for understanding ADS.

5.4.3 Mechanisms underlying errors

Perseverations

For this simulation, examination of transition node activity reveals little about the underlying mechanisms responsible for the observed errors. Rather, error types may be better understood by examining the likely results of disruption on the attractor space of the system, as discussed in chapter 2. We discussed our PFC representations as stable attractors in the associative loop system's 'state space'. More specifically, we suggested that each basal ganglia channel may define a coarse-level 'subset-attractor space', and intrinsic PFC connectivity defining a fine grained 'state-attractor' within this space. The transition nodes may be thought of as providing the energy required to move the system from one representation - or attractor - to another. Where connections in the associative loop remain intact, we suggest that each attractor has a relatively similar ‘gravity’, where the energy supplied by the transition layer is sufficient to move the system out of any single attractor. However, lesioning of PFC connectivity, as we have done in this simulation, is likely to disrupt the relative gravity of each of these attractors, weakening some and strengthening others. This effect is illustrated is figure 5.17. The greater the proportion of connections lesioned, the more exaggerated this effect is likely to be, possibly to the extent of eliminating some attractors altogether. After lesioning, once the system then makes a transition into one of the resulting ‘deep’ attractors, transition node activity will no longer be sufficient to drive changes to the expressed PFC representation. This results in continued perseverations of the same action.

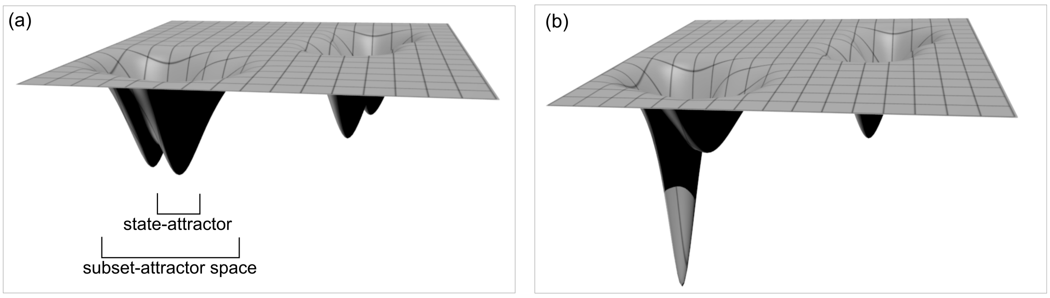

Figure 5.17: (a) Simplified schematic diagram of the hypothesised attractor space of the associative loop system before lesioning. (b) Hypothesised alteration caused by lesioning PFC connectivity. In (b), transition node activity no longer supplies the system with sufficient energy to escape deep attractors, and so continuous perseverations dominate the error profile. Lesioning of a greater proportion of connectivity leads to increasing disruptions of the attractor space, thus a higher likelihood of such perseverations.

We noticed that certain actions in particular tended to be perseverated. The **pick up spoon** and **scoop sugar** actions in particular were heavily over-represented in these errors. Clearly then, while the connections lesioned on each trial were randomly selected, the effects of this disruption were less so. We propose that these two actions were more likely to be perseverated due to the conjunction of their component feature nodes (pick up & spoon; and scoop & sugar) in multiple PFC representations. For instance, these two actions were represented in three different versions of the **add sugar** subtask (tea, coffee [sugar first], and coffee [milk first]). Statistically, any deepened or strengthened attractor will be more likely to correspond to one of these more highly represented actions, leading to the high frequency occurrence of perseverations of these actions.

Other single action errors

Other single action errors would tend to result simply from the weakening of certain representations, rather than 'over-strengthening' as with continuous perseverations. For instance, omissions of single actions would result from an ineffective transition to a state which was de-stabilised by lesioning; in these cases, rather than a decay to the *previous* representation or state-attractor, which would have resulted in a perseveration, PFC would 'settle' into the representation corresponding to the subsequent stage of the task, hence causing an action omission. This phenomenon would underlie anticipation-omission errors, for example, though these were relatively rare. This is probably due to a generalised effect of lesioning, whereby more attractors were disrupted than strengthened, providing few stable states and a resulting tendency to perseverate only one.

Subtask based errors

Though single action errors, especially perseverations, were dominant, particularly at high noise, other errors certainly were observed throughout all levels of disruption. These were, we propose, also due to the alteration of the strength of individual state-attractors, whereby the precise location in attractor space of the disruption resulted in different error types. Subtask based errors, for example, may have generally resulted from the degradation of one representation causing a transition into a contextually different representation, but one encoding the same action. Imagine, for example, that the representation of **pick up spoon** in the context of the tea-making task was heavily disrupted (representation T5 in figure 4.1). Transition node activity evoking this action will thus fail in driving PFC into this representation. However, due to significant overlap between each version of the **pick up spoon** action (representations T5, C6a and C8b in figure 4.1), PFC might settle into a less disrupted version of this action representation. The result will be a confusion of context (see section 5.3.6), but still the expected action itself. It is possible that this ability to access 'backup' versions of corrupted representations underlies the greater robustness of the coffee-making task at low-medium disruption, where two alternative correct versions of the task are effectively available for selection.

5.4.4 Discussion

The current simulation produced a strikingly different profile of errors to that observed for the prior simulation which examined disruption of temporal order knowledge. Most notably, this manipulation did not appear to account for the types of errors generally performed by healthy controls, with no predominance of subtask based errors at low disruption. At high levels of disruption, errors were dominated by continuous perseverations, particularly for the tea-making task. A distinct mechanism was proposed to underlie the errors in this simulation, whereby lesioning of intrinsic PFC connectivity corrupted the 'attractor space' of the associative loop; weakened attractors corresponded to representations which were then omitted, whereas strengthened attractors resulted in a tendency to perform continuous perseverations.

These results suggest that the transient disruption of action schemas is unlikely to be the primary result of distraction in healthy controls, as, even a low levels of disruption, we found a dominance of single action errors, rather than the expected subtask based errors that would be consistent with evidence from, for example, Reason's diary studies (Reason, 1979, 1984, 1990). This might be expected given the type of damage imposed here; the fact that lesioning of connections was necessary - rather than the simple addition of noise - suggests that this type of disruption might be the result of a more long-lasting impairment than that caused by transient distraction in normal performance. Nor is it likely that action schema disruption is a central mechanism in ADS, as it does not appear to account for the relative *rates* of errors commonly observed in human behavioural studies. However, most error types were observed to some degree, suggesting that schema disruption may be a component of a greater pattern of damage in ADS. Indeed, it is interesting to note that Forde and Humphreys (2002) pointed to a dissociation between recurrent and continuous perseverations in two patients, FK and HG, and indeed suggested distinct underlying deficits. The current simulation suggests that those patients who show a tendency to display continuous perseveration have a primary deficit involving disruption to schemas themselves. This is less likely to be the case in those patients who rarely show this type of error. Thus, we would predict that those ADS patients subject to continuous perseverations would show evidence of action schema breakdown on tests such as those utilised by Forde et al. (2004).

5.5 General discussion

5.5.1 Summary of findings

Any model of sequential behaviour should be able to account for errors in human sequential performance. In section 5.1 we asked whether disruption of any particular single process was able to account for the full range of errors observed in ADS, as well as normal slips of action commonly performed by healthy controls. More specifically, we focused on the hypothesis proposed by Humphreys, Forde et al. (Humphreys & Forde, 1998; Humphreys et al., 2000; Forde et al., 2004) that action disorganisation is due to disruption of action schemas and/or temporal order knowledge. Generally, we have provided support for these ideas. In particular, we found that the disruption of temporal order knowledge accounted for several general trends observed in previous work on human error. Importantly, this process is stored separately from action schemas in our model, in contrast to the previous SRN model (Botvinick & Plaut, 2004). Specifically, such disruption was able to account for the predominance of subtask based errors in healthy controls, the dominance of omission errors in patients, and both omission and sequence error rates reflected the equivalent rates in patient data. We also replicated the omission rate effect, though only on one of our tasks.

In contrast, disruption of action schemas themselves did not result in an error profile that was generally representative of the human errors data, suggesting that such a deficit is unlikely to underlie the full range of errors in either healthy controls or in ADS patients. However, its ability to produce a relatively wide range of errors suggests its possible presence alongside other deficits in ADS. In particular, this simulation appeared to account for particular nuances in the patient data, reflecting more closely those patients who show a propensity to commit continuous perseverations, such as HG (Forde & Humphreys, 2000).

Despite the success of these simulations, however, certain error types remain unaccounted for or under-represented in both simulations. Semantic errors, for example, were relatively rare, particularly compared with patients HH and JK (Schwartz et al., 1991, 1995), who showed a strong tendency to make object substitutions. Additionally, the model was not designed to address quality or spatial errors. While a result of implementation, this limits the conclusions we may draw about the generality of either type of disruption across error types. In the following discussion, we consider the insights regarding the organisation of cognitive information resulting from the current study and the relation of the current model to previous modelling work.

5.5.2 Mechanisms of error commission

In the first simulation examining the effects of disruption to temporal order knowledge, noise applied to the transition layer caused the selection of incorrect PFC representations which tended to share features with the desired representation. This results from the overlapping nature of the feature-based representations of context and external environment. This organisation results in a 'gradient' of disruption, whereby higher levels of noise increase the probability of activation of increasingly *dis*similar representations to the intended one. At low levels of noise, this tended to cause a confusion of context, whereby the model would select PFC representations corresponding to the correct *action*, but in an incorrect context. This manifested in behavioural errors at a later stage of the task, a phenomenon which has been observed previously in modelling (Botvinick & Plaut, 2004) and experimental (Botvinick & Bylsma, 2005) work. At higher levels of noise, however, this tended to result in selection of PFC representations which had little in common with the intended one, causing immediate errors and more disjointed behaviour, manifesting as single action errors. We also found that noise caused spurious and/or untimely activity in transition nodes, resulting in interference or premature switching of PFC representations, often resulting in stereotypical single action errors, such as anticipation-omissions, object substitutions and continuous perseverations.

In the second simulation examining disruption to action schemas, damage to intrinsic PFC connectivity - and thus the action schemas such connectivity supports - appears to critically destabilise a set of the total representations, and conversely, other representations may become 'super' stable. In terms of the attractor space of the network, this seems to cause an imbalance of the relative strengths or depths of the attractors corresponding to each encoded representation. Interestingly, at low levels of disruption, this manifested differently in the tea- and coffee-making tasks. This is possibly due to a greater level of robustness of the coffee task to schema disruption as a result of the flexibility of the **add milk** and **add sugar** subtasks. Having two versions of each representation for these subtasks in the context of coffee-making may have allowed for a greater degree of disruption before large numbers of errors occurred, effectively allowing a 'backup' version of the task to be performed if one version was critically damaged. At higher levels of disruption however, the imbalance of the strengths of the attractors seems sufficiently severe that performance resembles that of the tea-making task, whereby fewer representations remain 'intact' and the model is unable to escape the resulting particularly strong or deep attractors in the state space of the associative loop system.

Cognitive vs behavioural identification of errors

Importantly, it is clear from the current study that different types of behavioural breakdown are achievable by disrupting different processes, but also that different types of disruption at the cognitive level may have the same or similar behavioural manifestations. Furthermore, it was notable throughout our study that errors tended to occur which would not intuitively be attributed to the disruption that was imposed. The disruption of temporal order knowledge, for instance, was responsible for a number of errors which have been attributed to different causes in previous work. For example, low-level 'toying' behaviour (the repeated picking up and putting down of objects), and object substitutions have been said to be caused by an imbalance of cognitive and sensory influences on behaviour (Schwartz et al., 1991; Cooper & Shallice, 2000). Here, however, we have shown that damage to regions encoding temporal order knowledge is capable of producing these error types, though not at rates observed in the human behavioural literature.

Our discussion of the precise mechanisms underlying errors in our first simulation additionally highlighted the finding that the behavioural manifestation of an error may be qualitatively different to the actual cognitive dynamics inducing the error. For instance, we showed that premature switching of PFC representations - an 'omission' at the cognitive level - frequently caused, behaviourally, a continuous perseveration error. This example suggests that omissions - particularly of single actions - may not be a problem of selection, but a problem of sufficient maintenance, with the further implication that ADS may not simply be a disorder of selection, but also one of timing. Further examples of discord between cognitive and behavioural interpretations of errors include apparent semantic errors, which themselves were frequently results of these omissions at the cognitive level. This finding, in particular, is consistent with evidence that semantic errors need not rely on an underlying deficit in semantic knowledge (Schwartz et al., 1991, 1995; Forde & Humphreys, 2000) and sheds some light on the reasons for this unintuitive finding.

The observed patterns of errors in human behavioural studies may thus not necessarily reflect the underlying processing, and should be interpreted with caution in future studies. Most importantly, we argue against the arbitrary classification of multiple error types into a single category (in particular, 'sequence' errors). Beyond the confusion resulting from different definitions, this indicates a degree of consistency which is potentially misleading; we certainly did not find that sequence errors as a category correlated with either type of disruption. It is also important for future behavioural studies to examine tasks which allow the distinction of 'cognitive' error types where possible. In the current tea- and coffee-making tasks, for example, the correct performance of the tasks involved the repeated pouring of items into the cup. In a behavioural study, such tasks might result in the confounding of error types, such as the example of continuous perseverations given above. Tasks involving more distinct component actions might provide more opportunity to understand the cognitive processes underlying error types.

Competitive queueing and activation gradients

As discussed, we found evidence to support the hypothesis that the disruption of temporal order knowledge accounts for many of the error types commonly observed in ADS. Humphreys and colleagues (Humphreys & Forde, 1998; Forde & Humphreys, 2002; Forde et al., 2004) interpreted this in terms of a competitive queueing account, whereby disruption of sequence knowledge was conceived of as a degradation of the activation gradient across component action schemas required for the task. Moreover, they proposed that insufficient rebound inhibition upon the currently selected action representation is the cause of continuous perseverations of the type that were common in our second simulation. This account, while appealing, suggests a strong degree of predetermined preparedness of all actions, and implies that each representation is automatically activated as a result of inhibition of the previous action representation. This, we suggest, is a relatively passive account of selection which gives little credit to real-time influences on behaviour.

Our model, however, provides a closely related, but more active view on the process of sequential selection. We suggest that during the performance of the current action, *any* suitable candidate next-action should display a degree of sub-threshold activation. In our model, this results from the overlapping projections from PFC to the transition layer. The amount of excitation an action representation receives at any time may be indicative not of its rank order in the sequence as competitive queueing accounts would imply, but of its *suitability* as a subsequent action. This suitability may be regarded as a measure of the number of features it has in common with the correct or intended subsequent action representation. We term this a *suitability-based* queueing account, rather than the more traditional rank order-based account. This means, importantly, that influences on the cognitive representations of actions are dynamic; each representation receives activation in an 'on-the-fly' manner according, predominantly, to the current overall context, rather than receiving a static gradient of activation from the outset. This in turn means that sequential performance may be naturally more flexible, taking into account relevant and changing information from the environment which may alter the planned course, rather than rigidly performing actions according to a predetermined order.

Importantly, as each individual representation is individually stable, no external gradient of activation must be maintained upon the representations of the actions involved in the sequence. Therefore, once a transition successfully occurs to the subsequent action representation, inhibition on the prior action representation need not be explicitly maintained by some external source in order to prevent reactivation. This avoids the inherent problems of maintaining inhibition on previously selected actions over the course of a sequence, for example, when the same action is required multiple times in a single task.

Though evidence for competitive queueing accounts exists (Averbeck et al., 2002), we discussed in the previous chapter that evidence for the pre-preparedness of action representations in PFC was also consistent with our model, and what we now discuss as a suitability based account. Again, we emphasise the importance of future studies to attempt to distinguish these possibilities. We suggest, however, that our interpretation accounts for a greater degree of the data than the traditional view. Take, for example, the dominance of subtask based errors at low noise, as are common in healthy performance. As Botvinick & Plaut (2004) discuss, and as we have also shown, errors at low noise result from a confusion between very similar *contexts*; this is due to the similarity in their representations, and thus their relative interchangeability. A traditional competitive queueing account, however, would predict that low noise would result in single action errors, where confusion occurs between two adjacent actions in a sequence. In particular, we would expect a dominance of anticipation omissions at low noise, which the human behavioural data do not support.

Humphreys and Forde's account of perseverations as a result of inadequate rebound inhibition (Humphreys & Forde, 1998) may also be reinterpreted in light of the current results. Here, we have suggested that such perseverations may be a result of disruption to the system's 'attractor space', with the result that certain actions, once initiated, are difficult to deselect as a result of insufficient energy supplied by the transition nodes. This is consistent to an extent with Humphreys and Forde's interpretation, but with an important difference which relates to the role of the external environment in influencing the course of action. In competitive queueing accounts, sequencing is insensitive to the current state of the environment, relying on simple rebound inhibition for sequencing. In our model, this plays an active role in selection by providing vital information about the suitability of each possible subsequent action, potentially disambiguating between multiple candidates.

5.5.3 Variability across patients and tasks

Earlier, we highlighted the fact that patients labelled with ADS display different patterns of errors; in particular, errors such as object substitutions and perseverations have varied markedly across candidates. We suggest that proposals of a generalised reduction of processing resources (Schwartz et al., 1998) are inadequate to account for this range of error patterns. Rather, we suggest that individual processes may be selectively damaged, resulting in a propensity to produce particular error types. Here, we have provided some insight about the possible results to two of these mechanisms; namely, disruption of temporal order knowledge and of action schemas themselves. In particular, we have provided support for the suggestion that disrupted temporal order knowledge may be primarily responsible for the more commonly observed features of ADS and action slips in normal populations, and thus may be the general foundation for ADS. It is probable, however, that damage to additional processes are responsible for more nuanced errors; we have shown that damage to action schemas can result in a propensity to commit continuous perseverations. We would predict that damage to other processes would be primarily responsible for semantic errors, as these were infrequent in both simulations.

In addition to different deficits arising from damage to distinct processes, overall error rates differed quite markedly for each task that we examined. For example, more errors were observed for the coffee-making task in our first simulation, whereas the opposite pattern was found for the second, suggesting that different tasks are more or less robust to particular types of disruption. Additionally, rates of particular error types differed for each task. Additions, for example, accounted for a significant number of errors in the tea-making task, whereas they did not appear in the coffee-making task, and higher relative rates of omission errors were seen in the coffee-making task. This is consistent with previous work showing different tasks evoked different error types (Forde & Humphreys, 2002), and indicates that observed error types are not simply a function of the disrupted process, but also of the particular task features.

It is probable, for instance, that different error rates on different tasks reflect the relative cognitive demands and the degree of flexibility in the tasks, as well as the particular opportunities for error during each task. The current model would predict that a longer task is likely to be more vulnerable to omissions, for example, not only as there are simply more actions to omit, but also because the probability of a mistake occurring increases with overall sequence length. Indeed, Humphreys & Forde (1998) reported that the number of subtasks required for the task affected FK's ability to perform it. Equally, the model also predicts that the more flexibility in the rank order of the component subtasks, the more recurrent perseverations and subtask omissions would be observed, due to 'losing track' of the current point in the sequence. Any subtasks which do not have an immediately visible outcome (e.g., **add sugar**) would be expected to be omitted more frequently, as no external cue exists as to the status of the subgoal. Very similar tasks are likely to suffer intrusions from one another, as we have shown with the frequent performance of the **add milk** subtask during the tea-making task. Future experimental work should thus focus on relating the particular task demands to the errors observed, in order to better understand the underlying cognitive processes.

5.5.4 Contention scheduling, supervisory attention & PFC

The effects of disrupted sequential behaviour, particularly on familiar tasks, are frequently discussed in terms of the supervisory attention (SAS) and contention scheduling systems (CSS) (Norman & Shallice, 1986; Shallice, 1982). These systems are commonly thought to underlie controlled and automatic performance, respectively. Routine sequences are proposed to rely primarily on the contention scheduling system, requiring little conscious cognitive control. Damage to this system alone, however, is not regarded as sufficient to account for ADS symptoms, as the SAS is thought to compensate, resulting in a more controlled performance of such sequences. Equally, damage to the SAS alone should rarely result in deficits on routine tasks due to the competence of the CSS for their performance. ADS is, therefore, generally thought of as resulting from more widespread damage involving both systems (Schwartz et al., 1995; Schwartz, 1995; Humphreys & Forde, 1998).

Where the SAS is equated with frontal lobe function (Shallice, 1982), this suggests that frontal damage may be insufficient to result in ADS. Indeed, patients with ADS often have damage extending beyond the frontal lobes. While ADS-like symptoms were observed in our study over damage to just a single process, conspicuously we do not include any kind of error monitoring, a process that is regarded as a key component of SAS function, and one that has been said to be compromised in ADS (Humphreys & Forde, 1998; Forde & Humphreys, 2002). Such monitoring processes may not affect the commission of errors *per se*, but rather the response after errors are committed. In normal participants, for example, it has been shown that errors - particularly single action errors - are often noted and spontaneously rectified. In patients, however, this is not always the case. Additionally, errors may not be corrected even when noticed by the patient (Forde & Humphreys, 2000), indeed suggesting some damage to SAS-like functions.

The omission of an explicit error monitoring process in our model effectively reflects damage to the SAS. However, given that ADS-like effects in the model are produced by the disruption of PFC function, this effectively posits the CSS at least partly in PFC. This region is, conversely, more often associated with SAS function. However, this is in fact consistent with our suggestion in the general introduction regarding 'routine' sequences not as true habit, which would indeed call into question the involvement of PFC, but as goal directed processes which, while utilising well learned and highly efficient stereotypical representations of context - 'schemas' - nonetheless require prefrontal involvement. It is likely, however, that certain components of the CSS do lie outside PFC. These, we suggest, are related more to true habitual performance, and may encode pure stimulus response contingencies, or overlearned action affordances, as demonstrated in chapter 4.

Within the CSS, the fact that different profiles of errors resulted from different types of disruption, and that each profile appears to be more or less strongly represented in individual patients, as discussed with reference to patients FK and HG (Humphreys & Forde, 1998), suggests that action schemas and knowledge of their temporal order for performance are, to some extent, represented distinctly in the brain. Partiot et al (1996) indeed suggested that these processes are dissociated, suggesting that the left and right frontal lobes, respectively, are specialised for dealing individually with these processes. While evidence has not been found for this particular pattern of localisation (Humphreys & Forde, 1998), it is somewhat consistent with evidence that different regions in PFC are selectively involved in the selection and maintenance of working memory representations (Rowe et al., 2000). If our model represents these areas in the transition layer and PFC, respectively, selective damage to these areas would be expected to preferentially result in the distinct error profiles we observed in our simulations. Also of interest is evidence that basal ganglia damage can result in continuous perseverations (Luria, 1965; Sandson & Albert, 1984); while we observed this effect after disruption to stability of representations by lesioning intrinsic PFC connections, similar results would be expected by lesioning projections to and from basal ganglia in the associative loop. The role of basal ganglia in maintaining representations in the current model (see chapter 4) may explain this effect.

5.5.5 Comparison with existing models

Previous models have focused on proposals that ADS is a result of an imbalance of top-down and bottom-up influences on behaviour (Cooper & Shallice, 2000), or of a general reduction in cognitive resources (Botvinick & Plaut, 2004; see also Cooper et al., 2005). The simulation of each of these types of disruption has accounted for a range of errors and effects, and it is possible, perhaps even likely, that ADS patients suffer from each of these problems to a greater or lesser degree. The real power of these models then lies not just in the amount of human error data that is replicated, but in their ability to provide insight into the underlying processes governing sequential performance.

Botvinick & Plaut (2004) effectively examined the 'reduced cognitive load' hypothesis by applying noise to the hidden layer of their SRN model during testing. It is notable that their overall rates of omission errors were quite markedly higher than those observed in the behavioural data. They showed an omission rate of 77% at a medium level of noise; given their replication of the omission rate effect it is assumed that this percentage increased further with greater disruption. It is likely that their high omission rate resulted from their general disruption, which is likely to have affected multiple different processes. As mentioned above, disruption to several processes may cause omissions, suggesting that their inflated omission rate might be a result of this generalised damage.

Due to the nature of their model, however, which incorporates contextual representation, sequencing knowledge, and affordance information in an opaque fashion in the hidden layer, we are unable to gain any further insight into these processes. Though the authors provided an elegant discussion of the activity underlying errors in terms of the confusion of context to which we have alluded in this chapter, we cannot, for instance, conclude exactly which processes are responsible for this activity. The present study, however, allows for an analysis at the mechanistic level as Botvinick & Plaut (2004) provide, but given the distinct separation of functions into neuroanatomically plausible regions, we are able to give a finer account of the particular processes underlying errors; in our case, disruption of temporal order knowledge and action schemas. Conversely, the IAN model presented by Cooper & Shallice (2000) accounts less well for the patterns seen in the human errors literature, with a more erratic overall error profile seen with increasing noise, and an absence of, for example, recurrent perseveration errors. However, given the separation of function, the model is able to give a more precise account of the disruption leading to the errors that are observed. By assigning separate functions to particular regions and connections, this allows a deeper analysis of the nature of the errors that are observed with regard to specific processes. This is made particularly clear in a follow up study which suggested that disruption of separate processes in a similar model differentially accounted for patterns of errors in distinct disorders of action (Cooper et al., 2005).

Reconciliation

As in chapter 4, we believe that our model points to a reconciliation of the IAN (Cooper & Shallice, 2000, 2006a, 2006b) and SRN models (Botvinick & Plaut, 2004, 2006b, 2006c). The partially distributed nature of the representations of schemas we have included, for example, allows the production of many of the human-like error profiles that were also seen by Botvinick & Plaut (2004), and the utility of an explanation of errors at the mechanistic level in terms of degradation of underlying representations. It is important to note that this benefit is maintained despite the hierarchical structure of the representations themselves. However, by separating function into distinct regions of the model, we also retain the explanatory power of the IAN model (Cooper & Shallice, 2000) and by adopting a schema based approach, interpret our results according to cognitive level theories regarding the organisation of action (e.g., Norman & Shallice, 1986; Houghton, 1990). Additionally, given the neuroanatomical basis of the model, we are able to interpret our results in terms of the underlying neural substrate and, importantly, make predictions about the expected deficits given the location of damage within PFC.

The ‘hybrid’ nature of our model also allows us to confront many of the problems discussed by each set of authors. For instance, Cooper & Shallice (2006a) suggest that errors in the recurrent network model occur only when the ‘error’ exists as a correct action elsewhere in the training corpus, citing an over-reliance on the training set. For example, they suggest that the sequence A-B-C is only likely to manifest as A-C (thus containing an omission error) if the sequence A-C is explicitly trained as part of another sequence or subsequence. We have shown, however, that in a model utilising a recurrent network approach, omissions of both single steps and of entire subtasks occur even when the preceding and following actions do not appear together in any valid version of the task. We attribute this more diverse range of errors to the separation of processes in the model, and our emphasis on working memory. We have shown, for example, that instability of working memory representations caused by interference from transition nodes gives rise to certain error types, which the SRN model cannot explicitly account for. Conversely, Botvinick & Plaut (2006b) criticise the fact that some strong patterns in the behavioural literature were not replicated by the IAN model. However, by implementing schemas as overlapping representations and disrupting a process which we believe is more central to ADS, we have shown that this lack of replication is not due to the fundamental organisation of the model in terms of schemas and goals within a hierarchical framework, but is more likely to be due to the particular locus of disruption in the original study.

5.5.6 Limitations

Model and task structure

As mentioned in section 5.2.1, the current instantiation of the model was not designed to account for spatial or quality errors, due to the nature of its implementation. While we do not wish to rule out the possibility that these error types may result from the types of disruption we have investigated here, we are not able to say with any certainty that they would be observed in a more complete model. Furthermore, it is likely that the limited number of semantic errors that we observed is at least partly a result of the fact that we do not include a representation of the semantic similarity of different objects. Without the inclusion of semantic information, we are unable to assess the performance of the model in response to distractor objects, which has been an important element of many behavioural studies to date. Additionally, a more complex task environment, perhaps including tasks of a more greatly differing nature, might result in different rates of action additions, as might the inclusion of more candidate action affordances for each object.

Omitted action types

As mentioned in chapter 4, we omitted both orienting actions ('fixate') and release actions ('put down'). While we incorporated processes and assumptions to account for this, inevitably these simplifications will have an effect on the behaviour of the model. The inclusion of orientation as a distinct type of action, for example, potentially opens the model to an entirely new class of error. Doing so would allow the inclusion of active sensory influences, such as the relative sensory salience of particular objects, on fixation. This might have significant effects on the rate of object substitutions, for example, as well as possible action additions, whereby an incorrectly fixated object might evoke an habitual response, in the manner of utilisation behaviour.

Equally, the inclusion of release actions might result in new opportunities for error. Indeed, we performed a preliminary simulation examining the effects of assuming a **put down** action after each **pour into** action. This measure had the effect of altering the nature of the external environment between subtasks. Due to space limitations, we are presently unable to provide a detailed explanation of this study; however, this had a significant effect on the overall rates of errors, suggesting that the manner in which actions are cognitively 'grouped' or parsed into subtasks has a significant effect on the actor's vulnerability to error at distinct points in a sequential task. This model further suggested that normal participants and patients may parse actions into subtasks in different ways. This is a finding worthy of further study which was not immediately obvious from the current model; thus future modelling work should aim to produce a more comprehensive version of the task which incorporates the action classes that we have omitted here.

Plasticity

Hard-coding all weights in the model had the important advantage of allowing us to retain complete control of the information processed by the model, and allowed a thorough understanding and discussions of the mechanisms underlying various types of error. However, without the incorporation of learning processes, we are unable to comment on a number of features that have been prominent in previous modelling work. Most notably, Botvinick & Plaut (2004) showed effects of relative task frequency on error type; they found that by including more instances of the coffee-making task in the training set, the frequency of intrusion errors from the coffee- to the tea-making task was increased. In the current instantiation of the model, we are currently unable to comment on such effects, but this would be an interesting avenue for future work. The model in its current form presents many opportunities for the implementation of plasticity, as discussed briefly in the previous chapter, and might be utilised to further understand the processes underlying the formation of PFC representations of schemas, as well as the learning of action-outcome and stimulus-response associations. Given the neuroanatomically constrained nature of the model this would allow investigation into the implementation of necessary mechanisms for such learning in the brain, in contrast to the back-propagation methods implemented by Botvinick & Plaut (2004).

5.6 Summary

In this chapter, we have tested the hypothesis that ADS is a result of disruption to temporal order knowledge and action schemas, in a neurally inspired computational model of routine sequences. Additionally, we have examined whether either of these processes may account for action slips in normal performance. In doing so, we have found further grounds for reconciliation between existing, competing models of such sequences, and have found support for the suggestion that the disruption of temporal order knowledge underlies many of the most common findings in human error studies, and we thus propose this is a core deficit of ADS. We suggest that disruption of action schemas may be an additional deficit which may underlie particular nuances of ADS, specifically in those patients who commonly show continuous perseverations. It is probable that these processes are, to some extent, dissociated in the frontal lobes. Though we have not simulated further disruption in the present study, it is likely that other deficits are also present in ADS, such as an imbalance of cognitive and sensory influences on action selection, or damaged semantic knowledge. The extent to which these particular deficits are present is likely to determine an individual's particular error profile. We note, however, that the manifestation at the behavioural level may not be entirely representative of erroneous activity at the cognitive level, and thus caution is advised when interpreting the results of human behavioural studies.

Chapter 6

General Discussion

6.1 Main results and contribution of the research

As in-depth discussions of the work have been presented in each chapter, to which we direct the reader for a detailed analysis of the work, we conclude the thesis simply by summarising the main findings and achievements of the present research and outlining areas of interest for future study.

In this thesis, we aimed to bring together insights from several fields in order to address a series of questions regarding the processes subserving the dynamics of prefrontal representations of context, and the nature of their influence on action selection for the successful performance of routine, sequential action. Such action was highlighted as an area of profound interest given its significance for understanding the organisation and dynamics of cognitive information in the brain, and the nature of its breakdown in certain disorders of action. Two significant modelling contributions were discussed, each taking a distinct focus and providing several insights into this area (Botvinick & Plaut, 2004; Cooper & Shallice, 2000). However, two particular points were noted. Firstly, that the competing models had failed to resolve certain areas of disagreement - most notably, the representational schemes used and the role of goals (Botvinick & Plaut, 2006b, 2006c; Cooper & Shallice, 2006a, 2006b). Secondly, neither model had any significant grounding in neuroanatomy, thus their applicability to the biological system was limited. Given recent findings from neuroanatomy suggesting the existence of several projections between territories of the BGTC hierarchy (Calzavara et al., 2007; Draganski et al., 2008; Geyer et al., 2000; Haber et al., 2000; Haber & Calzavara, 2009; Joel & Weiner, 1994), and based on findings from neurophysiology and neuropsychology (Humphreys & Forde, 1998; Schwartz et al., 1998; Tanji & Hoshi, 2008), we suggested this system was a likely substrate for the organisation of action sequences, hypothesising that connections between loops primarily serve to exert motivational and cognitive influences, respectively, on the selection of contextual representations in working memory, and on motor action selection.

Given evidence suggesting that the preservation of the intrinsic circuitry of the BGTC architecture across functional territories (Middleton & Strick, 2000), and the well supported hypothesis that basal ganglia is specialised for selection (Mink, 1996; Redgrave et al., 1999), we developed of a model of the associative BGTC loop for the selection of cognitive representations of context. This was based on an existing model of action selection in basal ganglia, dubbed the 'GPR' (Gurney et al., 2001a, 2001b; Humphries & Gurney, 2002). Insights into the likely distributed nature of cognitive representations of contextual information suggested the need for a re-conceptualisation of the 'channel' in the GPR for the processing of associative information, and the inclusion of additional selection functionality within cortex. This led to a novel functional architecture which mediated the selection of subsets of prefrontal representations by associative regions of basal ganglia, supported by intrinsic connectivity in PFC. The model presented is successful insofar as it represents the first neuroanatomically constrained model of associative basal ganglia which implemented several processes, identified as necessary in the problem analysis, which have previously been regarded as incompatible; notably, the competitive selection, timely deselection and active maintenance of non-localist representations in PFC, and convergence in corticostriatal projections (Frank et al., 2001).

The incorporation of the associative loop model into a larger theoretical architecture addressed the means by which cognitive representations of task context may influence motor action selection, and the mechanisms underlying their sequencing. Consistent with evidence from neuroanatomy (Borra et al., 2008; Calzavara et al., 2007; Cavada & Goldman-Rakic, 1991), we proposed that context sensitive action selection is mediated by the integration of sensory and contextual influences in the motor territories of the BGTC hierarchy. Here, the sensory influence was implemented in the form of action affordances specified by a representation of the currently fixated object. Fixation of the desired object was driven by the current PFC representation, a process which reflected task-related influences on visual search. Subsequent activation of object-related affordances caused the low-level excitation of a set of candidate actions in motor territories of basal ganglia. A direct corticostriatal projection from PFC exerted a biasing influence upon candidate actions according to the current representation of temporal task context, thus embodying contextual influences on selection. Sequence generation was mediated by a novel transitioning mechanism distinct from the actively maintained representation within PFC. Inclusion of such a mechanism was based on evidence that intrinsically stable representations require an external influence to generate transitions (Rutishauser & Douglas, 2009), and may represent a region of PFC distinct from that responsible for maintenance (Rowe et al., 2000). This mechanism provided excitation to appropriate representations of the next stage of the task based on an internal representation of the current stage, and the state of the external environment, thus incorporating multiple influences on sequencing.

The model was tested on its ability to perform two related sequential tasks. These were typical examples of sequences explored in human behavioural studies (Humphreys & Forde, 1998; Schwartz et al., 1991, 1995), and similar to those investigated by Cooper & Shallice (2000) and Botvinick & Plaut (2004). In this instance, the tasks were designed to tax specific capabilities of the model; for instance, a number of actions were required for multiple separate subtasks and in both tasks, demanding flexibility in terms of the context in which these actions were selected. Actions with outcomes which were not immediately visible were employed, requiring the maintenance in working memory of information delineating task history. Finally, the same actions were required multiple times within a task, demanding a sequencing mechanism which does not rely on continued inhibition of previously selected actions.

This led to an analysis of the required contextual information for the mediation of the tasks, which showed the need for several distinct features and a hierarchical decomposition of PFC representations. In particular, the specific representation of the task, subtask, intended action and target object were required for the completion of each stage of the tasks. Representations of this nature allowed performance of the 'quasi-hierarchical' waiter scenario, which, according to Botvinick & Plaut (2004), should have proven difficult for the current model given its localist representation of the current goal within PFC. Analyses showed similar dynamics to both the existing models of sequencing upon which we have focused, as well as resembling existing neural recording data (Averbeck et al., 2002), thus pointing towards a reconciliation of the differences of these models and a neurally focused account of their processing. Additionally, the successful performance of the model gave support for the hypothesised functional basis of the observed interconnectivity between BGTC loops.

Testing the model under disruption in order to examine its ability to account for data from studies of action slips and ADS lended support to the the twin hypotheses that errors result from damaged temporal order knowledge and/or action schemas. The model was able to account for several trends observed in the human errors literature after disruption of temporal order knowledge, including the tendency for subtask-based errors in healthy controls, the finding that omissions and sequence errors account for the two greatest overall error types in ADS patients, and similar overall rates of each error type. Accordingly, we provided support for the notion that many of the core deficits of ADS may be explained by a disruption of temporal order knowledge. In contrast, we suggested that more nuanced patterns of disruption may be explained by more specific deficits; in particular, that the tendency for certain individuals to be prone to continuous perseverations (Forde & Humphreys, 2002) may be due to the de-stabilisation and over-stabilisation of particular action schemas in those individuals.

These results thus offer an explanation for the processes mediating sequencing that is easily interpretable in terms of existing cognitive theories, but that is grounded in biology. For example, the connectivity between PFC and the transition layer may be seen as implementing the activation gradients of competitive queueing accounts, and stable representations are interpretable as schemas within the contention scheduling system (CSS), as proposed by Norman & Shallice (1986). Importantly, the present account also gives rise to several testable predictions relating to distinct aspects of the model, regarding the sensitivity to particular task features expected in associative regions of basal ganglia and PFC, and the expected behavioural effects of task manipulations in healthy controls and ADS patients. Finally, we have offered a new ‘suitability-based’ interpretation of competitive queueing accounts of sequencing (e.g., Houghton, 1990) which is compatible with existing neurophysiological data (Averbeck et al., 2002), and proposed a framework by which the current account may be tested against the more traditional order-based account.

6.2 Future work

Beyond the achievements of the present research, there are a number of areas in which the current model can be expanded to address a wider range of questions. For instance, we discussed in chapters 4 and 5 the importance of introducing biologically plausible learning mechanisms, so that we may examine whether the correct connectivity may be adopted by the model to produce the behaviour we have examined in the present study. Ideally, such work should be inspired by insights from developmental psychology and neuroscience, looking particularly at the progression of skill development, and their composition into deliberate and directed actions, and the formation of schemas. Further, as discussed in chapter 5, the addition of more complex task environments, spatial parameters, semantic object knowledge and the inclusion of the omitted release and orienting actions would increase the opportunity for particular error types, which may serve to further validate our conclusions regarding the role of temporal order knowledge in ADS. In addition to these suggestions discussed in earlier chapters, there are several potential further applications of the model that are worth noting.

6.2.1 Error monitoring

It is commonly reported that a significant difference in performance between healthy controls and ADS patients is the ability to detect when an error has been made and to remedy performance accordingly (Forde & Humphreys, 2000; Reason, 1990). Cognitive theories stress the importance of the supervisory attention system (SAS) in error detection (Norman

& Shallice, 1986), and propose that simultaneous damage to both the CSS and SAS is required for the appearance of ADS symptoms (Humphreys & Forde, 1998; Schwartz et al., 1995; Schwartz, 1995) whereas in normal performance, the occupation of the SAS, for example, by some other distracting task, results in mild error (Morady & Humphreys, 2009). Given the lack of any such error-detection mechanism in the current model, our work would appear to be consistent with these accounts. However, without comparison of model performance with the inclusion of such a mechanism, we are unable to actively investigate this idea at the present time.

Evidence from neuroimaging points to the involvement of the anterior cingulate cortex (ACC) in performance monitoring for error detection and correction (Carter et al., 1998), consistent with cognitive theories placing the SAS in the frontal lobes. The addition of an ACC-inspired component to the model would allow an exploration of the possible interactions of error-detection with various types of disruption. Such a programme of work may complement any future studies incorporating learning processes in the model, by focusing on an investigation of the means by which the control of action becomes 'routine' with practice, whereby SAS involvement in sequence production is diminished with expertise (Norman & Shallice, 1986), with possible implications for understanding performance monitoring in routine tasks. In particular, the interaction of this process with the learning of stable representations as schemas would be of great interest for future study.

6.2.2 Generalisation: different types of sequence

In the general introduction, we noted that different processes are likely to underlie the processing of different categories of sequence. We emphasised the distinct nature of well-learned, familiar action sequences that we have investigated in the present study, and unfamiliar strings of words, letters or numbers encountered in tasks such as immediate serial recall (ISR). Most particularly, we stressed that the nature of the cognitive representations required to mediate routine tasks are likely to encode stereotypical examples of the task and its required actions as schemas, whereas those representations governing performance of more transient sequences may be more likely to encode rules and generalisable task features beyond the specific elements in any instance of the task.

It would be of interest to examine whether any of the principles underlying the design of the current model may be extended to account for the mediation of less familiar sequences. As an example, neurons encoding 'rank order', such as those that played a critical role in encoding task context in the current model, would have a clear utility in such tasks, by assigning a position in sequence to a particular stimulus (e.g., Funahashi et al., 1993, 1997). However, the means by which stimuli might be attached to such markers is unclear, though the observed sensitivity of PFC neurons to conjunctions of features (Tanji & Hoshi, 2008) may assist in this process. It is unlikely, however, that stable representations, such as those we have exploited here, are formed immediately for unfamiliar sequences. Rather, modelling work suggests the existence of dynamic working memory traces in such tasks (Botvinick & Plaut, 2006a), while an influential theory points to the adaptability of PFC neurons which may represent distinct information according to the individual instance of the task, suggesting rather different mechanisms to those we have implemented here (Duncan, 2001). Understanding how these processes interact within cognitive and motor selection systems to produce a full range of sequential behaviour would be an ambitious but worthy goal of future developments.

6.2.3 Other disorders of action

A strong model of routine sequences which claims to represent the underlying processes should be able to take into account data not simply from a single action disorder, but from any which affect the composite processes the model claims to account for. Indeed, an extension of Cooper & Shallice (2000) found that disruption of distinct processes accounted preferentially for error data from patients suffering from ADS, ideational apraxia and utilisation syndrome (Cooper et al., 2005). Any future developments to the current model should aim to account for such data.

We have already shown in chapter 4 that the strong activation of a particular affordance may override the contextual bias on a competing affordance in motor cortex; this was interpreted as a strong stimulus-response association and representative of a habit according to the definition provided by Dickinson (1985). It is likely that this process is at least partially responsible for utilisation behaviour, possibly via the breakdown of interloop corticostriatal connectivity, or of PFC representations themselves. Future work on the present model should aim to explore this possibility. Moreover, the model of action selection in basal ganglia upon which the present study is based has been shown to display problems with selection, reminiscent of Parkinsonian akinesia, as a result of reduced levels of simulated tonic dopamine in striatum (Gurney et al., 2001b). Again, future work should examine the results of such a manipulation in the present multi-loop architecture. These existing results suggest the exciting possibility that the present multiple loop model may be able to account for key behaviour patterns in multiple disorders of action as well as healthy controls. If so, this provides strong support for the processes that we have suggested underlie routine sequential performance.

6.3 Concluding remarks

In sum, this thesis presents a unified body of research whereby a model grounded in neuroanatomy and sound computational arguments has accounted for a significant amount of experimental data, providing new insights regarding the organisation of cognitive information for sequential performance and its transformation into motor information for action. It also points to a reconciliation of existing competing models and framed existing cognitive theories in neuroanatomical terms. Future work should attempt to address implementational issues which may have affected the nature of the results, and exploit the potential of the model for supporting investigations into the processes involved in the organisation of cognitive information for a broader range of sequential tasks, and in multiple disorders of action.

References

Aarts, H., & Dijksterhuis, A. (2000). Habits as knowledge structures: automaticity in goal-directed behavior. *Journal of Personality and Social Psychology*, *78*(1), 53-63.

Abraham, W., & Robins, A. (2005). Memory retention–the synaptic stability versus plasticity dilemma. *Trends in neurosciences*, *28*(2), 73–78.

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

Alexander, G., & Crutcher, M. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, *13*(7), 266–271.

Alexander, G., & DeLong, M. (1985). Microstimulation of the primate neostriatum. II. Somatotopic organization of striatal microexcitable zones and their relation to neuronal response properties. *Journal of Neurophysiology*, *53*(6), 1417–1430.

Alexander, G., DeLong, M., & Strick, P. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience*, *12*(3), 505–519.

Ashe, J., Lungu, O., Basford, A., & Lu, X. (2006). Cortical control of motor sequences. *Current Opinion in Neurobiology*, *16*(2), 213–221.

Averbeck, B., Chafee, M., Crowe, D., & Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13172-13177.

Averbeck, B., Crowe, D., Chafee, M., & Georgopoulos, A. (2003). Neural activity in prefrontal cortex during copying geometrical shapes. II. Decoding shape segments from neural ensembles. *Experimental Brain Research*, *150*(2), 142–153.

Averbeck, B., Sohn, J., & Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience*, *9*(2), 276–282.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200.

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

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–742.

Balleine, B., Liljeholm, M., & Ostlund, S. (2009). The integrative function of the basal ganglia in instrumental conditioning. *Behavioural Brain Research*, *199*(1), 43–52.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609.

Barbas, H., & Pandya, D. (1987). Architecture and frontal cortical connections of the premotor cortex (area 6) in the rhesus monkey. *The Journal of Comparative Neurology*, *256*(2), 211–228.

Bar-Gad, I., Morris, G., & Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, *71*(6), 439–473.

Barone, P., & Joseph, J. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, *78*(3), 447–464.

Batista, A., & Andersen, R. (2001). The parietal reach region codes the next planned movement in a sequential reach task. *Journal of Neurophysiology*, *85*(2), 539–544.

Bayley, P., Frascino, J., & Squire, L. (2005). Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature*, *436*(7050), 550–553.

Beiser, D., & Houk, J. (1998). Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology*, *79*(6), 3168–3188.

Beiser, D., Hua, S., & Houk, J. (1997). Network models of the basal ganglia. *Current opinion in neurobiology*, *7*(2), 185–190.

Berns, G., & Sejnowski, T. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, *10*(1), 108–121.

Bolam, J., Hanley, J., Booth, P., & Bevan, M. (2000). Synaptic organisation of the basal ganglia. *Journal of Anatomy*, *196*(4), 527–542.

Bornstein, A., & Daw, N. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, *21*(3), 374–380.

Borra, E., Belmalih, A., Calzavara, R., Gerbella, M., Murata, A., Rozzi, S., & Luppino, G. (2008). Cortical connections of the macaque anterior intraparietal (AIP) area. *Cerebral Cortex*, *18*(5), 1094–1111.

Botvinick, M. (2007). Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1615–1626.

Botvinick, M., & Bylsma, L. (2005). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin & Review*, *12*(6), 1011–1017.

Botvinick, M., & Plaut, D. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*(2), 395-429.

Botvinick, M., & Plaut, D. (2006a). Short-term memory for serial order: a recurrent neural network model. *Psychological Review*, *113*(2), 201-233.

Botvinick, M., & Plaut, D. (2006b). Such stuff as habits are made on: a reply to Cooper and Shallice (2006). *Psychological Review*, *113*(4), 917–928.

Botvinick, M., & Plaut, D. (2006c). Postscript: the way forward – comment. *Psychological Review*, *113*(4), 928.

Bowers, J. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*(1), 220-251.

Brom, C., & Bryson, J. (2006). *Action selection for intelligent systems.* [White paper]. European Network for the Advancement of Artificial Cognitive Systems. Retrieved from `http://www.eucognition.org/asm-whitepaper-final-060804.pdf`

Brown, J., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, *17*(4), 471–510.

Buxbaum, L., Schwartz, M., & Montgomery, M. (1998). Ideational apraxia and naturalistic action. *Cognitive Neuropsychology*, *15*(6-8), 617–643.

Calzavara, R., Mailly, P., & Haber, S. (2007). Relationship between the corticostriatal terminals from areas 9 and 46, and those from area 8A, dorsal and rostral premotor cortex and area 24c: an anatomical substrate for cognition to action. *European Journal of Neuroscience*, *26*(7), 2005–2024.

Carter, C., Braver, T., Barch, D., Botvinick, M., Noll, D., & Cohen, J. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749.

Cavada, C., Compañy, T., Tejedor, J., Cruz-Rizzolo, R., & Reinoso-Suárez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cerebral Cortex*, *10*(3), 220–242.

Cavada, C., & Goldman-Rakic, P. (1991). Topographic segregation of corticostriatal projections from posterior parietal subdivisions in the macaque monkey. *Neuroscience*, *42*(3), 683–696.

Chambers, J., & Gurney, K. (2008, November). *A computational model of 'inaction selection' in multiple domains of basal ganglia.* Poster session presented at the Society For Neuroscience Annual Meeting, Washington D.C., USA.

Chevalier, G., & Deniau, J. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*, *13*(7), 277–280.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1585–1599.

Cohen, M., & Frank, M. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural Brain Research*, *199*(1), 141–156.

Cooper, R., Schwartz, M., Yule, P., & Shallice, T. (2005). The simulation of action disorganisation in complex activities of daily living. *Cognitive Neuropsychology*, *22*(8), 959–1004.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297–338.

Cooper, R., & Shallice, T. (2006a). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887-916.

Cooper, R., & Shallice, T. (2006b). Structured representations in the control of behavior cannot be so easily dismissed: A reply to Botvinick and Plaut (2006). *Psychological Review*, *113*(4), 929–931.

Courtney, S. (2004). Attention and cognitive control as emergent properties of information representation in working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 501–516.

Cromwell, H., & Berridge, K. (1996). Implementation of action sequences by a neostriatal site: a lesion mapping study of grooming syntax. *The Journal of Neuroscience*, *16*(10), 3444–3458.

Cromwell, H., & Schultz, W. (2003). Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *Journal of Neurophysiology*, *89*(5), 2823–2838.

Crutcher, M., & DeLong, M. (1984). Single cell studies of the primate putamen. II. Relations to direction of movements and patterns of muscular activity. *Experimental Brain Research*, *53*(2), 244–258.

Curtis, C., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423.

Curtis, C., & Lee, D. (2010). Beyond working memory: the role of persistent activity in decision making. *Trends in Cognitive Sciences*, *14*(5), 216–222.

Dalley, J., Cardinal, R., & Robbins, T. (2004). Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neuroscience & Biobehavioral Reviews*, *28*(7), 771–784.

Daw, N., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

DeLong, M. (1990). Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, *13*(7), 281–285.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 761–772.

Dezfouli, A., & Balleine, B. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036–1051.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *308*(1135), 67-78.

Doll, B., Jacobs, W., Sanfey, A., & Frank, M. (2009). Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94.

Dominey, P. (1995). Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics*, *73*(3), 265–274.

Dominey, P., & Arbib, M. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, *2*(2), 153–175.

Dominey, P., Arbib, M., & Joseph, J. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *Journal of Cognitive Neuroscience*, *7*(3), 311–336.

Draganski, B., Kherif, F., Klöppel, S., Cook, P., Alexander, D., Parker, G., ... Frackowiak, R. (2008). Evidence for segregated and integrative connectivity patterns in the human basal ganglia. *The Journal of Neuroscience*, *28*(28), 7143-7152.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829.

Eysenck, M., & Keane, M. (2005). *Cognitive psychology: A student's handbook* (4th ed.). Hove & New York: Psychology Press.

Fagg, A., & Arbib, M. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, *11*(7-8), 1277–1303.

Felleman, D., & Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, *1*(1), 1-47.

Fogassi, L., & Luppino, G. (2005). Motor functions of the parietal lobe. *Current Opinion in Neurobiology*, *15*(6), 626–631.

Forde, E., & Humphreys, G. (2000). The role of semantic knowledge and working memory in everyday tasks. *Brain and Cognition*, *44*(2), 214–252.

Forde, E., & Humphreys, G. (2002). Dissociations in routine behaviour across patients and everyday tasks. *Neurocase*, *8*(1-2), 151–167.

Forde, E., Humphreys, G., & Remoundou, M. (2004). Disordered knowledge of action order in action disorganisation syndrome. *Neurocase*, *10*(1), 19–28.

Frank, M. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism.

Journal of Cognitive Neuroscience, *17*(1), 51–72.

Frank, M., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral Cortex*, *22*(3), 509–526.

Frank, M., Loughry, B., & O'Reilly, R. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160.

Frank, M., Scheres, A., & Sherman, S. (2007). Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1641–1654.

Fujii, N., & Graybiel, A. (2003). Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science*, *301*(5637), 1246–1249.

Fukai, T. (1999). Sequence generation in arbitrary temporal patterns from theta-nested gamma oscillations: a model of the basal ganglia-thalamo-cortical loops. *Neural Networks*, *12*(7-8), 975–987.

Funahashi, S., Inoue, M., & Kubota, K. (1993). Delay-related activity in the primate prefrontal cortex during sequential reaching tasks with delay. *Neuroscience Research*, *18*(2), 171–175.

Funahashi, S., Inoue, M., & Kubota, K. (1997). Delay-period activity in the primate prefrontal cortex encoding multiple spatial positions and their order of presentation. *Behavioural Brain Research*, *84*(1-2), 203–223.

Fuster, J. (2001). The prefrontal cortex-an update: time is of the essence. *Neuron*, *30*(2), 319–333.

Fuster, J., & Alexander, G. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654.

Georgopoulos, A., Kalaska, J., Caminiti, R., & Massey, J. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *The Journal of Neuroscience*, *2*(11), 1527-1537.

Gerfen, C., Engber, T., Mahan, L., Susel, Z., Chase, T., Monsma, F., & Sibley, D. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science*, *250*(4986), 1429-1432.

Gerfen, C., & Wilson, C. (1996). Chapter II the basal ganglia. *Handbook of Chemical*

Neuroanatomy, *12*, 371–468.

Geyer, S., Matelli, M., Luppino, G., & Zilles, K. (2000). Functional neuroanatomy of the primate isocortical motor system. *Anatomy and Embryology*, *202*(6), 443–474.

Gibson, J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Gillies, A., & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, *15*(5), 762–770.

Girard, B., Filliat, D., Meyer, J., Berthoz, A., & Guillot, A. (2005). Integration of navigation and action selection functionalities in a computational model of cortico-basal-ganglia-thalamo-cortical loops. *Adaptive Behavior*, *13*(2), 115-130.

Glover, S. (2002). Visual illusions affect planning but not control. *Trends in Cognitive Sciences*, *6*(7), 288–292.

Gnadt, J., & Mays, L. (1995). Neurons in monkey parietal area LIP are tuned for eye-movement parameters in three-dimensional space. *Journal of Neurophysiology*, *73*(1), 280–297.

Goldman, P., & Nauta, W. (1977). An intricately patterned prefronto-caudate projection in the rhesus monkey. *The Journal of Comparative Neurology*, *171*(3), 369–385.

Goodale, M., & Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Graybiel, A. (2000). The basal ganglia. *Current Biology*, *10*(14), R509–R511.

Graybiel, A. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, *31*, 359–387.

Graziano, M., Aflalo, T., & Cooke, D. (2005). Arm movements evoked by electrical stimulation in the motor cortex of monkeys. *Journal of Neurophysiology*, *94*(6), 4209–4223.

Groenewegen, H., Mulder, A., Beijer, A., Wright, C., Lopes Da Silva, F., & Pennartz, C. (1999). Hippocampal and amygdaloid interactions in the nucleus accumbens. *Psychobiology*, *27*(2), 149–164.

Gross, C. (2002). Genealogy of the “Grandmother Cell”. *The Neuroscientist*, *8*(5), 512–518.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, *11*(1), 23–63.

Gruber, A., Dayan, P., Gutkin, B., & Solla, S. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of Computational Neuroscience*, *20*(2), 153–166.

Gurney, K., Prescott, T., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*(6), 401–410.

Gurney, K., Prescott, T., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, *84*(6), 411–423.

Haber, S. (2003). The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, *26*(4), 317–330.

Haber, S., & Calzavara, R. (2009). The cortico-basal ganglia integrative network: the role of the thalamus. *Brain Research Bulletin*, *78*(2-3), 69–74.

Haber, S., Fudge, J., & McFarland, N. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *The Journal of Neuroscience*, *20*(6), 2369–2382.

Haber, S., Kim, K., Mailly, P., & Calzavara, R. (2006). Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *The Journal of Neuroscience*, *26*(32), 8368–8376.

Haber, S., & Mcfarland, N. (2001). The place of the thalamus in frontal cortical-basal ganglia circuits. *The Neuroscientist*, *7*(4), 315–324.

Hanakawa, T. (2011). Rostral premotor cortex as a gateway between motor and cognitive networks. *Neuroscience Research*, *70*(2), 144–154.

Hanes, D., & Schall, J. (1996). Neural control of voluntary movement initiation. *Science*, *274*(5286), 427-430.

Hartman-von Monakow, K., Akert, K., & Künzle, H. (1978). Projections of the precentral motor cortex and other cortical areas of the frontal lobe to the subthalamic nucleus in the monkey. *Experimental Brain Research*, *33*(3-4), 395–403.

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(8), 1242–1254.

Hasegawa, R., Blitz, A., & Goldberg, M. (2004). Neurons in monkey prefrontal cortex whose activity tracks the progress of a three-step self-ordered task. *Journal of Neurophysiology*, *92*(3), 1524–1535.

Hazy, T., Frank, M., & O'Reilly, R. (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.

Henson, R. (1996). Unchained memory: error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *49*(1), 80–115.

Henson, R. (1998). Short-term memory for serial order: the start-end model. *Cognitive Psychology*, *36*(2), 73–137.

Hikosaka, O., Takikawa, Y., & Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Reviews*, *80*(3), 953–978.

Hinton, G., McClelland, J., & Rumelhart, D. (1986). Distributed representations. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 77–109). MIT Press: Cambridge, MA.

Hoover, J., & Strick, P. (1993). Multiple output channels in the basal ganglia. *Science*, *259*(5096), 819–821.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554-2558.

Horvitz, J. (2002). Dopamine gating of glutamatergic sensorimotor and incentive motivational input signals to the striatum. *Behavioural Brain Research*, *137*(1-2), 65–74.

Hoshi, E., Shima, K., & Tanji, J. (1998). Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *80*(6), 3392–3397.

Houghton, G. (1990). The problem of serial order: a neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current research in natural language generation* (pp. 287–319). Academic Press: London.

Houk, J., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P., ... Simo, L. (2007). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1573–1583.

Houk, J., & Wise, S. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cerebral Cortex*, *5*(2), 95–110.

Humphreys, G., & Forde, E. (1998). Disordered action schema and action disorganisation syndrome. *Cognitive Neuropsychology*, *15*(6-8), 771–811.

Humphreys, G., Forde, E., & Francis, D. (2000). The organization of sequential actions. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (p. 427-442). MIT Press: Cambridge MA.

Humphries, M., & Gurney, K. (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network: Computation in Neural Systems*, *13*(1), 131–156.

Humphries, M., & Prescott, T. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in Neurobiology*, *90*(4), 385–417.

Humphries, M., Stewart, R., & Gurney, K. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *The Journal of Neuroscience*, *26*(50), 12921–12942.

Ikemoto, S. (2007). Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens–olfactory tubercle complex. *Brain Research Reviews*, *56*(1), 27–78.

Inase, M., Sakai, S., & Tanji, J. (1996). Overlapping corticostriatal projections from the supplementary motor area and the primary motor cortex in the macaque monkey: an anterograde double labeling study. *The Journal of Comparative Neurology*, *373*(2), 283–296.

Jeannerod, M., Arbib, M., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences*, *18*(7), 314–320.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*(4-6), 535–547.

Joel, D., & Weiner, I. (1994). The organization of the basal ganglia-thalamocortical circuits: open interconnected rather than closed segregated. *Neuroscience*, *63*(2), 363–379.

Joel, D., & Weiner, I. (1997). The connections of the primate subthalamic nucleus: indirect pathways and the open-interconnected scheme of basal ganglia-thalamocortical circuitry. *Brain Research Reviews*, *23*(1-2), 62–78.

Joel, D., & Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, *96*(3), 451–474.

Jordan, M. (1986). *Serial order: a parallel distributed processing approach.* (Tech. Rep.). ICI Report 8604, Institute for Cognitive Science, University of California, San Diego.

Jurado, M., & Rosselli, M. (2007). The elusive nature of executive functions: a review of our current understanding. *Neuropsychology Review*, *17*(3), 213–233.

Kantak, S., Sullivan, K., Fisher, B., Knowlton, B., & Winstein, C. (2010). Neural substrates of motor memory consolidation depend on practice structure. *Nature Neuroscience*, *13*(8), 923–925.

Kastner, S., & Ungerleider, L. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*, 315–341.

Kawato, M., & Samejima, K. (2007). Efficient reinforcement learning: computational theories, neuroscience and robotics. *Current Opinion in Neurobiology*, *17*(2), 205–212.

Kermadi, I., & Joseph, J. (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *Journal of Neurophysiology*, *74*(3), 911–933.

Khamassi, M., Girard, B., Berthoz, A., & Guillot, A. (2004). Comparing three critic models of reinforcement learning in the basal ganglia connected to a detailed actor in a S-R task. In F. Groen, N. Amato, A. Bonarini, E. Yoshida, & B. Krse (Eds.), *Proceedings of the 8th International Conference on Intelligent Autonomous Systems IAS-8.* (pp. 430–437). IOS Press.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, *302*(5648), 1181–1185.

Lashley, K. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior* (p. 112-136). New York: Wiley.

Lévesque, M., Charara, A., Gagnon, S., Parent, A., & Deschênes, M. (1996). Corticostriatal projections from layer V cells in rat are collaterals of long-range corticofugal axons. *Brain Research*, *709*(2), 311–315.

Lewis, S., Dove, A., Robbins, T., Barker, R., & Owen, A. (2004). Striatal contributions to working memory: a functional magnetic resonance imaging study in humans. *European Journal of Neuroscience*, *19*(3), 755–760.

L'hermitte, F. (1983). 'Utilization Behaviour' and its relation to lesions of the frontal lobes. *Brain*, *106*(2), 237–255.

Luppino, G., Rozzi, S., Calzavara, R., & Matelli, M. (2003). Prefrontal and agranular cingulate projections to the dorsal premotor areas F2 and F7 in the macaque monkey. *European Journal of Neuroscience*, *17*(3), 559–578.

Luria, A. (1965). Two kinds of motor perseveration in massive injury of the frontal lobes. *Brain*, *88*(1), 1–10.

Matsumoto, R., Nair, D., LaPresto, E., Bingaman, W., Shibasaki, H., & Luders, H. (2007). Functional connectivity in human cortical motor system: a cortico-cortical evoked potential study. *Brain*, *130*(1), 181-197.

McFarland, N., & Haber, S. (2002). Thalamic relay nuclei of the basal ganglia form both reciprocal and nonreciprocal cortical connections, linking multiple frontal cortical areas. *The Journal of Neuroscience*, *22*(18), 8117-8132.

McHaffie, J., Stanford, T., Stein, B., Coizet, V., & Redgrave, P. (2005). Subcortical loops through the basal ganglia. *Trends in Neurosciences*, *28*(8), 401–407.

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(1), 103–107.

Middleton, F., & Strick, P. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Reviews*, *31*(2-3), 236–250.

Miller, E. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, *1*(1), 59–65.

Miller, E., & Cohen, J. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, P., Brody, C., Romo, R., & Wang, X. (2003). A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. *Cerebral Cortex*, *13*(11), 1208–1218.

Mink, J. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425.

Mink, J., & Thach, W. (1993). Basal ganglia intrinsic circuits and their role in behavior. *Current Opinion in Neurobiology*, *3*(6), 950–957.

Monchi, O., & Taylor, J. (1999). A hard wired model of coupled frontal working memories for various tasks. *Information Sciences*, *113*(3-4), 221–243.

Morady, K., & Humphreys, G. (2009). Comparing action disorganization syndrome and dual-task load on normal performance in everyday action tasks. *Neurocase*, *15*(1), 1–12.

Muakkassa, K., & Strick, P. (1979). Frontal lobe inputs to primate motor cortex: evidence for four somatotopically organized 'premotor' areas. *Brain Research*, *177*(1), 176-182.

Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *Journal of Cognitive Neuroscience*, *13*(5), 626–647.

Nakayama, Y., Yamagata, T., Tanji, J., & Hoshi, E. (2008). Transformation of a virtual action plan into a motor plan in the premotor cortex. *The Journal of Neuroscience*, *28*(41), 10287–10297.

Nambu, A., Kaneda, K., Tokuno, H., & Takada, M. (2002). Organization of corticostriatal motor inputs in monkey putamen. *Journal of Neurophysiology*, *88*(4), 1830–1842.

Nambu, A., Tokuno, H., & Takada, M. (2002). Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience Research*, *43*(2), 111–117.

Norman, D. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1-15.

Norman, D., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. Davidson, G. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, p. 1-18). New York: Plenum.

Norman, D., & Shallice, T. (2000). Attention to action: Willed and automatic control of behavior. In M. Gazzaniga (Ed.), *Cognitive neuroscience: a reader* (p. 376-390).

Wiley-Blackwell:Oxford.

O'Reilly, R. (2006). Biologically based computational models of high-level cognition. *Science*, *314*(5796), 91-94.

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

O'Reilly, R., Herd, S., & Pauli, W. (2010). Computational models of cognitive control. *Current Opinion in Neurobiology*, *20*(2), 257–261.

Ostlund, S., Winterbauer, N., & Balleine, B. (2009). Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *The Journal of Neuroscience*, *29*(25), 8280–8287.

Parent, A. (1990). Extrinsic connections of the basal ganglia. *Trends in Neurosciences*, *13*(7), 254–258.

Parent, A., & Hazrati, L. (1995a). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews*, *20*(1), 91–127.

Parent, A., & Hazrati, L. (1995b). Functional anatomy of the basal ganglia. II. The place of subthalamic nucleus and external pallidium in basal ganglia circuitry. *Brain Research Reviews*, *20*(1), 128–154.

Partiot, A., Grafman, J., Sadato, N., Flitman, S., & Wild, K. (1996). Brain activation during script event processing. *Neuroreport*, *7*(3), 761-766.

Pennartz, C., Berke, J., Graybiel, A., Ito, R., Lansink, C., van der Meer, M., ... Voorn, P. (2009). Corticostriatal interactions during learning, memory processing, and decision making. *The Journal of Neuroscience*, *29*(41), 12831–12838.

Percheron, G., & Filion, M. (1991). Parallel processing in the basal ganglia: up to a point. *Trends in Neurosciences*, *14*(2), 55-56.

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, *20*(3), 249–262.

Phillips, J., Humphreys, G., Noppeney, U., & Price, C. (2002). The neural substrates of action retrieval: an examination of semantic and visual routes to action. *Visual Cognition*, *9*(4-5), 662–685.

Ponzi, A. (2008). Dynamical model of salience gated working memory, action selection and reinforcement based on basal ganglia and dopamine feedback. *Neural Networks*, *21*(2-3), 322–330.

Prescott, T., Montes González, F., Gurney, K., Humphries, M., & Redgrave, P. (2006). A robot model of the basal ganglia: behavior and intrinsic processing. *Neural Networks*, *19*(1), 31–61.

Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107.

Rao, S., Rainer, G., & Miller, E. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, *276*(5313), 821–824.

Reason, J. (1979). Actions not as planned: The price of automatization. In G. Underwood & R. Stevens (Eds.), *Aspects of consciousness, vol 1: Psychological issues* (p. 67-89). London: Wiley.

Reason, J. (1984). Lapses of attention in everyday life. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (p. 515-549). New York: Academic Press.

Reason, J. (1990). *Human error*. New York: Cambridge University Press.

Redgrave, P., Prescott, T., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*(4), 1009–1023.

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

Reep, R., Cheatwood, J., & Corwin, J. (2003). The associative striatum: organization of cortical projections to the dorsocentral striatum in rats. *The Journal of Comparative Neurology*, *467*(3), 271–292.

Reynolds, J., & O'Reilly, R. (2009). Developing PFC representations using reinforcement learning. *Cognition*, *113*(3), 281–292.

Rhodes, B., Bullock, D., Verwey, W., Averbeck, B., & Page, M. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science*, *23*(5), 699–746.

Riddoch, M., Humphreys, G., & Edwards, M. (2000). Visual affordances and object selection. In S. Monsell & J. Driver (Eds.), (pp. 603–625). MIT Press: Cambridge, MA.

Riddoch, M., Humphreys, G., & Price, C. (1989). Routes to action: evidence from apraxia. *Cognitive Neuropsychology*, *6*(5), 437–454.

Rockel, A., Hiorns, R., & Powell, T. (1980). The basic uniformity in structure of the neocortex. *Brain*, *103*(2), 221–244.

Rolls, E., & Treves, A. (1990). The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network: Computation in Neural Systems*, *1*(4), 407–421.

Rougier, N., Noelle, D., Braver, T., Cohen, J., & O'Reilly, R. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7338-7343.

Rougier, N., & O'Reilly, R. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*(4), 503–520.

Rowe, J., Toni, I., Josephs, O., Frackowiak, R., & Passingham, R. (2000). The prefrontal cortex: response selection or maintenance within working memory? *Science*, *288*(5471), 1656–1660.

Rumelhart, D., & Norman, D. (1982). Simulating a skilled typist: a study of skilled cognitive-motor performance. *Cognitive Science*, *6*(1), 1–36.

Rumelhart, D., Smolensky, P., McClelland, J., & Hinton, G. (1986). Schemata and sequential thought processes in PDP models. In D. Rumelhart, J. McClelland, & and the PDP Research Group (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (Vol. 2, pp. 7–57).

Rumiati, R., & Humphreys, G. (1998). Recognition by action: dissociating visual and semantic routes to action in normal observers. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 631-647.

Rutishauser, U., & Douglas, R. (2009). State-dependent computation using coupled recurrent networks. *Neural Computation*, *21*(2), 478–509.

Ryou, J., & Wilson, F. (2004). Making your next move: dorsolateral prefrontal cortex and planning a sequence of actions in freely moving monkeys. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 430–443.

Salinas, E. (2009). Rank-order-selective neurons form a temporal basis set for the generation of motor sequences. *The Journal of Neuroscience*, *29*(14), 4369–4380.

Sandson, J., & Albert, M. (1984). Varieties of perseveration. *Neuropsychologia*, *22*(6), 715–732.

Schmidt, R. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, *82*(4), 225-260.

Schroll, H., Vitay, J., & Hamker, F. (2012). Working memory and response selection: a computational account of interactions among cortico-basalganglio-thalamic loops. *Neural Networks*, *26*, 59–74.

Schwartz, M. (1995). Re-examining the role of executive functions in routine action production. *Annals of the New York Academy of Sciences*, *769*, 321–336.

Schwartz, M., & Buxbaum, L. (1997). Naturalistic action. In L. Rothi & K. Heilman (Eds.), *Apraxia: the neuropsychology of action* (p. 269-289). Hove, UK: Psychology Press.

Schwartz, M., Buxbaum, L., Montgomery, M., Fitzpatrick-DeSalme, E., Hart, T., Ferraro, M., … Coslett, H. (1999). Naturalistic action production following right hemisphere stroke. *Neuropsychologia*, *37*(1), 51–66.

Schwartz, M., Montgomery, M., Buxbaum, L., Lee, S., Carew, T., Coslett, H., … Mayer, N. (1998). Naturalistic action impairment in closed head injury. *Neuropsychology*, *12*(1), 13-28.

Schwartz, M., Montgomery, M., Fitzpatrick-DeSalme, E., Ochipa, C., Coslett, H., & Mayer, N. (1995). Analysis of a disorder of everyday action. *Cognitive Neuropsychology*, *12*(8), 863–892.

Schwartz, M., Reed, E., Montgomery, M., Palmer, C., & Mayer, N. (1991). The quantitative description of action disorganisation after brain damage: a case study. *Cognitive Neuropsychology*, *8*(5), 381–414.

Searle, J. (1980). The intentionality of intention and action. *Cognitive Science*, *4*(1), 47–70.

Seger, C., & Spiering, B. (2011). A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, *5*, Article 66.

Selemon, L., & Goldman-Rakic, P. (1985). Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. *The Journal of Neuroscience*, *5*(3), 776–794.

Selemon, L., & Goldman-Rakic, P. (1991). Parallel processing in the basal ganglia: up to a point [reply]. *Trends in Neuroscience*, *14*(2), 58-59.

Şengör, N., Karabacak, O., & Steinmetz, U. (2008). A computational model of cortico-striato-thalamic circuits in goal-directed behaviour. In *Proceedings of the 18th international conference on artificial neural networks part II* (pp. 328–337). Springer.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *298*(1089), 199–209.

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.

Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, *445*(7125), 315–318.

Sidibé, M., Bevan, M., Bolam, J., & Smith, Y. (1997). Efferent connections of the internal globus pallidus in the squirrel monkey: I. Topography and synaptic organization of the pallidothalamic projection. *The Journal of Comparative Neurology*, *382*(3), 323–347.

Sirigu, A., Zalla, T., Pillon, B., Grafman, J., Agid, Y., & Dubois, B. (1996). Encoding of sequence and boundaries of scripts following prefrontal lesions. *Cortex*, *32*(2), 297-310.

Smith, Y., Bevan, M., Shink, E., & Bolam, J. (1998). Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience*, *86*(2), 353-387.

Stafford, T., & Gurney, K. (2007). Biologically constrained action selection improves cognitive control in a model of the stroop task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1671–1684.

Strick, P., Dum, R., & Picard, N. (1995). Macro-organization of the circuits connecting the basal ganglia with the cortical motor areas. In J. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 117–130). Cambridge, MA: MIT Press.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tachibana, Y., Nambu, A., Hatanaka, N., Miyachi, S., & Takada, M. (2004). Input-output organization of the rostral part of the dorsal premotor cortex, with special reference to its corticostriatal projection. *Neuroscience Research*, *48*(1), 45–57.

Takada, M., Nambu, A., Hatanaka, N., Tachibana, Y., Miyachi, S., Taira, M., & Inase, M. (2004). Organization of prefrontal outflow toward frontal motor-related areas in macaque monkeys. *European Journal of Neuroscience*, *19*(12), 3328–3342.

Takada, M., Tokuno, H., Nambu, A., & Inase, M. (1998). Corticostriatal projections from the somatic motor areas of the frontal cortex in the macaque monkey: segregation versus overlap of input zones from the primary motor cortex, the supplementary motor area, and the premotor cortex. *Experimental Brain Research*, *120*(1), 114–128.

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review*, *15*(3), 535–542.

Tanji, J. (2001). Sequential organization of multiple movements: involvement of cortical motor areas. *Annual Review of Neuroscience*, *24*, 631–651.

Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological Reviews*, *88*(1), 37-57.

Taylor, J., & Taylor, N. (2000). Analysis of recurrent cortico-basal ganglia-thalamic loops for working memory. *Biological Cybernetics*, *82*(5), 415–432.

Tepper, J., Koós, T., & Wilson, C. (2004). GABAergic microcircuits in the neostriatum. *Trends in Neurosciences*, *27*(11), 662–669.

Tricomi, E., Balleine, B., & O'Doherty, J. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*(11), 2225–2232.

Tyrrell, T. (1992). Defining the action selection problem. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 1152–1157). Hillsdale, NJ: Lawrence Erlbaum.

Vitay, J., & Hamker, F. (2010). A computational model of basal ganglia and its role in memory retrieval in rewarded visual memory tasks. *Frontiers in Computational Neuroscience*, *4*, Article 13.

Voorn, P., Vanderschuren, L., Groenewegen, H., Robbins, T., & Pennartz, C. (2004). Putting a spin on the dorsal–ventral divide of the striatum. *Trends in Neurosciences*, *27*(8), 468–474.

Voytek, B., & Knight, R. (2010). Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences of the United*

States of America, *107*(42), 18167–18172.

Webster, M., Bachevalier, J., & Ungerleider, L. (1994). Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cerebral Cortex*, *4*(5), 470–483.

West, M., Carelli, R., Pomerantz, M., Cohen, S., Gardner, J., Chapin, J., & Woodward, D. (1990). A region in the dorsolateral striatum of the rat exhibiting single-unit correlations with specific locomotor limb movements. *Journal of Neurophysiology*, *64*(4), 1233–1246.

Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*(1), 1-15.

Wickens, J. (1997). Basal ganglia: structure and computations. *Network: Computation in Neural Systems*, *8*(4), R77–R109.

Wood, W., & Neal, D. (2007). A new look at habits and the habit-goal interface. *Psychological Review*, *114*(4), 843-863.

Wu, Y., Hu, J., Wu, W., Zhou, Y., & Du, K. (2012). Storage capacity of the hopfield network associative memory. In *Proceedings of the 5th international conference on intelligent computation technology and automation (ICICTA)* (pp. 330–336). IEEE.

Xiong, J., Parsons, L., Gao, J., & Fox, P. (1999). Interregional connectivity to primary motor cortex revealed using MRI resting state images. *Human Brain Mapping*, *8*(2-3), 151–156.

Yin, H., & Knowlton, B. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476.

Yin, H., Mulcare, S., Hilário, M., Clouse, E., Holloway, T., Davis, M., … Costa, R. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nature Neuroscience*, *12*(3), 333–341.

Zalla, T., Pradat-Diehl, P., & Sirigu, A. (2003). Perception of action boundaries in patients with frontal lobe damage. *Neuropsychologia*, *41*(12), 1619–1627.

Zheng, T., & Wilson, C. (2002). Corticostriatal combinatorics: the implications of corticostriatal axonal arborizations. *Journal of Neurophysiology*, *87*(2), 1007–1017.

Appendix A

Glossary of notation

This appendix contains a glossary of the notation used to describe the model in chapters 2 and 4. Some additional terms are included to clarify the full specification of parameters and weights for each simulation presented in the main text, which may be found in appendices B & C on the accompanying disc.

Basic leaky integrator neuron notation

a	Neuron activation
$\tilde{a}$	Neuron equilibrium
y	Neuron output
u	Neuron input
ϵ	Neuron output threshold
m	Neuron output function gradient

Simulation parameters

T	Simulated time
τ_m	Membrane time constant
dt	Simulation Timestep size
ζ	Salience input to PFC (chapter 2)
ϕ	Goal signal (chapter 4)
θ	Cortical selection threshold

Associative loop notation

X_{PFC} Set of all PFC nodes

X_i i^{th} PFC subset

x_i i^{th} PFC node

V Ventral anterior nucleus of thalamus

S Caudate; D1 expressing

C Caudate; D2 expressing

D Ventromedial subthalamic nucleus

G Dorsomedial globus pallidus (external)

H Dorsomedial globus pallidus (internal)

Motor loop notation

M Pre/motor cortex

v Ventrolateral nucleus of thalamus

s Putamen; D1 expressing

c Putamen; D2 expressing

d Dorsolateral subthalamic nucleus

g Ventrolateral globus pallidus (external)

h Ventrolateral globus pallidus (internal)

Additional regions notation

Γ Transition nodes

ξ Environment representation

Θ Object representations

Λ Action affordances

Model dimensions

N Number PFC nodes

N^R Number stable PFC representations

n^A Number associative BG channels

n^M Number motor BG channels

N^Γ Number transition nodes

N^Θ Number object representations

N^ϕ Number goals

Associative loop synaptic weights

W^{ρ}	Lateral inhibitory influence within transition PFC
$W^{\Gamma x}$	Synaptic strength transition layer → PFC
W^{V}	Synaptic strength VA thalamus → PFC
W^{X}	Synaptic strength PFC subset → caudate
W^{XV}	Synaptic strength PFC subset → VA thalamus
W^{XD}	Synaptic strength PFC subset → vmSTN
W^{SH}	Synaptic strength caudate (D1) → dmGPi
W^{CG}	Synaptic strength caudate (D2) → dmGPe
W^{DG}	Synaptic strength vmSTN → dmGPe
W^{DH}	Synaptic strength vmSTN → dmGPi
W^{GD}	Synaptic strength dmGPe → vmSTN
W^{GH}	Synaptic strength dmGPe → dmGPi
W^{HV}	Synaptic strength dmGPi → VA thalamus
λ_A	Dopaminergic modulation of caudate activation

Motor loop synaptic weights

W^{xp}	Synaptic strength PFC (action) → putamen
W^{xd}	Synaptic strength PFC (diffuse) → dlSTN
$W^{\Lambda M}$	Synaptic strength affordances → motor cortex
$W^{\Lambda p}$	Synaptic strength affordances → putamen
W^{M}	Synaptic strength motor cortex → putamen
W^{Mv}	Synaptic strength motor cortex → VL thalamus
W^{vM}	Synaptic strength VL thalamus → motor cortex
W^{Md}	Synaptic strength motor cortex → dlSTN
W^{sh}	Synaptic strength putamen (D1) → vlGPi
W^{cg}	Synaptic strength putamen (D2) → vlGPe
W^{dg}	Synaptic strength dlSTN → vlGPe
W^{dh}	Synaptic strength dlSTN → vlGPi
W^{gd}	Synaptic strength vlGPe → dlSTN
W^{gh}	Synaptic strength vlGPe → vlGPi
W^{hv}	Synaptic strength vlGPi → VL thalamus
λ_M	Dopaminergic modulation of putamen activation

Additional regions scalar synaptic weights

W^{ϕ} Influence of goal on transition nodes

$W^{x\Gamma}$ Synaptic strength PFC → transition layer

W^{ξ} Synaptic strength environment representation → transition layer

W^{Γ} Lateral inhibitory influence within transition layer

$W^{x\Theta}$ Synaptic strength PFC → objects

W^{Θ} Lateral inhibitory influence within object representations

W^{Λ} Synaptic strength objects → affordances

Neuron output parameters: associative loop

ϵ_X Output threshold: PFC

ϵ_V Output threshold: VA thalamus

ϵ_C Output threshold: caudate

ϵ_D Output threshold: vmSTN

ϵ_G Output threshold: dmGPe

ϵ_H Output threshold: dmGPi

m Neuron output gradient (all nuclei)

Neuron output parameters: motor loop

ϵ_M Output threshold: motor cortex

ϵ_v Output threshold: VL thalamus

ϵ_c Output threshold: putamen

ϵ_d Output threshold: dlSTN

ϵ_g Output threshold: vlGPe

ϵ_h Output threshold: vlGPi

m Neuron output gradient (all nuclei)

Neuron output parameters: additional regions

ϵ_{Θ} Output threshold: object representations

ϵ_{Λ} Output threshold: affordances

ϵ_{Γ} Output threshold: transition nodes

m Neuron output gradient (object representations and affordances)

m_{Γ} Transition node output gradient

Appendix B

Full parameter list:

Associative loop model

Appendix B detailing parameter values for all simulations presented in chapter 2 may be found on the accompanying disc.

Appendix C

Full parameter list: Complete model

Appendix C detailing parameter values for all simulations presented in chapter 4 may be found on the accompanying disc.