A computational approach to prefrontal cortex, cognitive control and schizophrenia: recent developments and current challenges

JONATHAN D. COHEN[1,2], TODD S. BRAVER[1] AND RANDALL C. O'REILLY[1]

[1] *Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.*
[2] *Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.*

SUMMARY

In this chapter we consider the mechanisms involved in cognitive control – from both a computational and a neurobiological perspective – and how these might be impaired in schizophrenia. By 'control', we mean the ability of the cognitive system to flexibly adapt its behaviour to the demands of particular tasks, favouring the processing of task-relevant information over other sources of competing information, and mediating task-relevant behaviour over habitual, or otherwise prepotent responses. There is a large body of evidence to suggest that the prefrontal cortex (PFC) plays a critical role in cognitive control. In previous work, we have used a computational framework to understand and develop explicit models of this function of PFC, and its impairment in schizophrenia. This work has lead to the hypothesis that PFC houses a mechanism for representing and maintaining context information. We have demonstrated that this mechanism can account for the behavioural inhibition and active memory functions commonly ascribed to PFC, and for human performance in simple attention, language and memory tasks that draw upon these functions for cognitive control. Furthermore, we have used our models to simulate detailed patterns of cognitive deficit observed in schizophrenia, an illness associated with marked disturbances in cognitive control, and well established deficits of PFC.

Here, we review results of recent empirical studies that test predictions made by our models regarding schizophrenic performance in tasks designed specifically to probe the processing of context. These results showed selective schizophrenic deficits in tasks conditions that placed the greatest demands on memory and inhibition, both of which we have argued rely on the processing of context. Furthermore, we observed predicted patterns of deterioration in first episode vs multi-episode patients. We also discuss recent developments in our computational work, that have led to refinements of the models that allow us to simulate more detailed aspects of task performance, such as reaction time data and manipulations of task parameters such as interstimulus delay. These refined models make several provocative new predictions, including conditions in which schizophrenics and control subjects are expected to show similar reaction time performance, and we provide preliminary data in support of these predictions. These successes notwithstanding, our theory of PFC function and its impairment in schizophrenia is still in an early stage of development. We conclude by presenting some of the challenges to the theory in its current form, and new directions that we have begun to take to meet these challenges. In particular, we focus on refinements concerning the mechanisms underlying active maintenance of representations within PFC, and the characteristics of these representations that allow them to support the flexibility of cognitive control exhibited by normal human behaviour.

Taken *in toto*, we believe that this work illustrates the value of a computational approach for understanding the mechanisms responsible for cognitive control, at both the neural and psychological levels, and the specific manner in which they break down in schizophrenia.

1. INTRODUCTION

(a) *Prefrontal cortex and cognitive control*

One of the central concepts within modern cognitive psychology is a distinction between controlled and automatic processing. Controlled processing is classically defined as relying on a limited capacity attentional system, while automatic processing is assumed to occur independently of attentional resources. Most contemporary theories posit that there is actually a continuum between controlled and automatic processing. Nevertheless, virtually all theorists acknowledge the need for some mechanism, or set of mechanisms responsible for the coordination of processing in a flexible fashion – particularly in novel tasks – and their relationship to attentional control. Perhaps the most explicit of these is Baddeley's theory of working memory (Baddeley 1986), which postulates a specific subcomponent responsible for executive control. The postulation of a cognitive system involved in executive control closely parallels theories concerning frontal lobe involvement in executive functions (Bianchi 1922; Damasio 1985; Luria 1969), and the long-standing clinical observation that patients with frontal lesions

exhibit a 'dysexecutive syndrome.' However, traditional theories have not specified the mechanisms by which the executive operates, either at the psychological or neurobiological levels. Recent advances have begun to address this need for more explicit theorizing.

Shallice (1982) has proposed the 'Supervisory Attentional System' (SAS) as a mechanism by which PFC coordinates complex cognitive processes, and Gathercole (1994) has suggested that this may provide a mechanism by which the executive controller operates in Baddeley's theory of working memory. Shallice's theory is described in terms of a production system architecture. This has appeal, as it relates the executive functions of frontal cortex to the well characterized mechanisms of other production system theories, which include the active maintenance of goal states to coordinate the sequences of production firings involved in complex behaviours (Anderson, 1983). One feature of goal representations is that they are actively maintained throughout the course of a sequence of behaviours, which coincides with the observation that PFC appears to be specialized for the active maintenance of task-relevant information (discussed below). Kimberg and Farah (1993) have also proposed a model of frontal function using a production system architecture, that they have used to simulate performance in a variety of tasks considered to rely on frontal lobe function. However, because both this and Shallice's SAS theory are cast in production system terms, they lack a transparent mapping onto specific neurobiological mechanisms. While neurobiological plausibility is not a requirement, *per se*, of a theory that seeks to explain the cognitive functions of PFC, this does become important if our goal is to understand the cognitive manifestations of a complex neuropsychiatric disorder, such as schizophrenia, in terms of its underlying pathophysiological processes. This goal has been one of the important motivations in our effort to understand how cognitive control may arise from the neurobiological mechanisms housed within PFC.

(*b*) *A connectionist approach*

Toward this goal, we have begun to develop a theory of PFC function using a computational architecture closer to the one used by the brain. We have done this within the connectionist, or parallel distributed processing framework (Rumelhart & McClelland 1986). The principles of this framework capture central features of computation as it is carried out in the brain (e.g. multiple simple processing units, graded flow of activation, modifiable connection weights, etc.), and allow us to explore their influence on behaviour by simulating performance in specific cognitive tasks (Cohen & Servan-Schreiber 1992). The central hypothesis that has emerged from this work is that PFC is used to represent context information, which we define as information necessary to mediate an appropriate behavioural response. This can be a set of task instructions, a specific prior stimulus, or the result of processing a sequence of prior stimuli (e.g. the interpretation resulting from processing a sequence of words in a sentence). We consider context representations to be closely related to goal representations within production system architectures. Thus, the actions associated with a particular goal may, in other contexts, be relatively infrequent or 'weak' behaviours. Such actions require the maintenance of an internal representation of the goal, or of goal-related knowledge to favour their execution, to suppress competing, possibly more compelling behaviours, and to coordinate their execution over temporally extended periods.

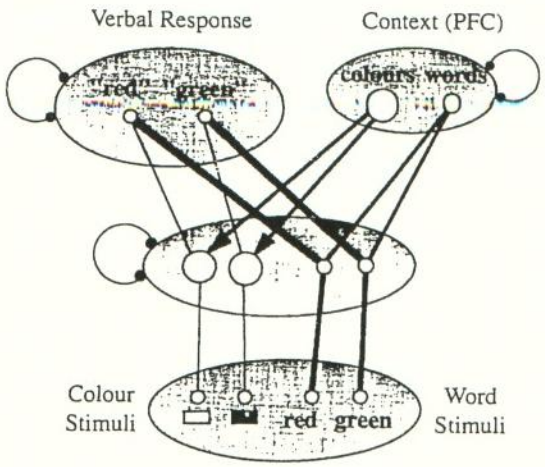


Figure 1. Simulation model of the Stroop task. Heavier lines in the word reading pathway indicate stronger connections in this pathway. Larger circles indicate active (solid) or potentiated (open) units. Arrows indicate flow of activity. Loops with dots indicate lateral inhibitory connections among units within the module.

We have begun to implement this theory in models of performance in specific cognitive tasks that tap simple forms of cognitive control. Our initial work focused on performance in the Stroop task (Cohen *et al.* 1990), in which subjects must selectively attend to one dimension of information (e.g. the colour in which a word is displayed) and ignore information in a competing, but prepotent dimension (e.g. the word itself). Our model of this task illustrates the close relationship between goals, context representations, and attentional effects, all of which we argue reflect the functioning of PFC. In the model, a task (i.e. goal) is specified by the appropriate pattern of activation over a set of units (context layer) that represents each of the two dimensions over which stimuli can vary. Activation of the appropriate units modulates the flow of activity along the pathway from input to response (attention), favouring processing in the task-relevant pathway over the stronger competing one. What provides the context layer with the capacity for control of processing is that it has representations of the stimulus dimensions that are relevant to the task, and that activating these biases processing in favour of one pathway over the other. This observation is fundamental to our hypothesis of how PFC is involved in cognitive control: Representations in PFC bias processing in posterior neocortex in favour of task-relevant pathways.

We have used these mechanisms to simulate detailed aspects of normal human performance in the Stroop task (Cohen *et al.* 1990), and a variety of other tasks

involving attentional control (Cohen *et al.* 1994; Cohen & Servan-Schreiber 1992; Cohen *et al.* 1992). We have also used them to help understand the mechanisms underlying other cognitive functions that have typically been associated with frontal function, and that are of relevance to schizophrenia: active memory and behavioural inhibition. We have argued that both of these functions reflect the operation of the context layer under different task conditions. Under conditions of response competition, when a strong response tendency must be overcome for appropriate behaviour, the context module can be seen to play an inhibitory role, by supporting the processing of task relevant information, which can then compete more effectively with irrelevant information. This is exemplified by our model of the Stroop task (see figure 1). In contrast, when there is a delay between information relevant to a response and the execution of that response, then the context module can be seen to play a role in memory, by actively maintaining that information over time, so that it can be used later to mediate the appropriate response. This is exemplified by our model of the AX version of the continuous performance task, in which the correct response to one stimulus depends on the identity of a previous one (described in detail below; see figure 3). This theory has allowed us to account for a wide, and seemingly disparate array of cognitive deficits that arise in schizophrenia, in terms of a disturbance of frontal function. Frontal dysfunction has long been considered to be a central pathophysiological feature of schizophrenia (Kolb & Whishaw 1983; Kraeplin 1950; Levin 1984), and our simulation models of frontal function have begun to provide an integrated and explicit account of how frontal impairment could give rise to specific patterns of cognitive deficit observed in this illness.

2. APPLICATIONS TO SCHIZOPHRENIA

(a) *Simulation studies*

One of the most overt features of schizophrenia is a failure of cognitive control. This is manifest in clinical symptoms such as distractibility, loosening of associations, and disorganized or socially inappropriate behaviour. In the laboratory, these disturbances have been observed as deficits of attention (Zubin 1975; Kornetsky & Orzack 1978; Wynne *et al.* 1978; Nuechterlein 1991; Cornblatt & Keilpk 1994;), working memory (Weinberger *et al.* 1986; Goldman-Rakic 1991; Park & Holzman 1992), and behavioural inhibition (Wapner & Krus 1960; Chapman *et al.* 1964; Storms & Broen 1969; Abramczyk *et al.* 1983; Wysocki & Sweet 1985; Manschreck *et al.* 1988; Carter *et al.* 1993; Cohen & Carter 1996). We have proposed that these deficits can be understood in terms of a disturbance in the processing of context, arising from a disturbance in frontal function. We have made this explicit, by committing performance in cognitive tasks that tap these cognitive functions to explicit simulation models, and showing that when damage is introduced to the context layer, the models are able to simulate schizophrenic performance in these tasks. For example, in Cohen & Servan-Schreiber 1992, we reported computer simulation models of three tasks that have commonly been considered to tap different cognitive processes: the Stroop task (selective attention and behavioural inhibition), the identical pairs version of the CPT (vigilance and signal detection ability), and a lexical disambiguation task (language processing). Models of all three tasks used a similar architecture. We showed that, in simulations of these tasks, a single disturbance to the layer used to represent context produced changes in performance that both quantitatively and qualitatively matched those observed for schizophrenics.

As noted above, our models suggest that both the active memory and behavioural inhibition functions commonly ascribed to PFC may reflect the operation of the context layer, under different behavioural conditions. This hypothesis has begun to receive empirical support in studies of normal cognitive function (Roberts *et al.* 1994), however it also has direct relevance to schizophrenic cognitive disturbances. Deficits of both active memory and inhibition have been reported in schizophrenia. However, to our knowledge, these have never been studied together under controlled task conditions. According to our theory, the ability to process context should be most important when there is both a need to maintain information over time and use that information after a delay to inhibit a habitual response. Below, we will discuss recent empirical results demonstrating that schizophrenics show selective and specific deficits under such conditions.

Our simulation studies also suggest that memory and inhibition effects are differentially sensitive to disturbances of the context layer. With mild to moderate disturbances, a deficit appears only when there is a delay between the context and a response. This is because, with partial degradation of context information, a sufficient amount may remain at short or no delays to mediate a contextually appropriate response (i.e. inhibit the irrelevant response). At longer delays, however, degraded representations succumb to the cumulative effects of noise, and a failure to inhibit the habitual (but incorrect) response may be observed. Thus, the models predict that mild to moderate disturbances in patients with schizophrenia should manifest primarily at long delays, which might be viewed as a memory deficit. With sufficiently severe disturbances of the context mechanism, however, deficits should begin to emerge without any delay, which might be viewed as a deficit of inhibition. Assuming a Kraeplinian view of schizophrenia, this may map onto the course of the illness, with a 'memory' deficit early in the course and, as the illness progresses, a gradual reduction in the delay that can be tolerated. Eventually, late in the illness, a deficit would be observed even without a delay and may be perceived clinically as a 'failure of inhibition.' In both cases, the same mechanism is impaired, though to a different degree. Thus, our models make an interesting, and counterintuitive prediction: that 'memory' deficits should emerge earlier than 'inhibitory' deficits. This is counterintuitive insofar as inhibitory deficits require suppression of strong competing responses, something

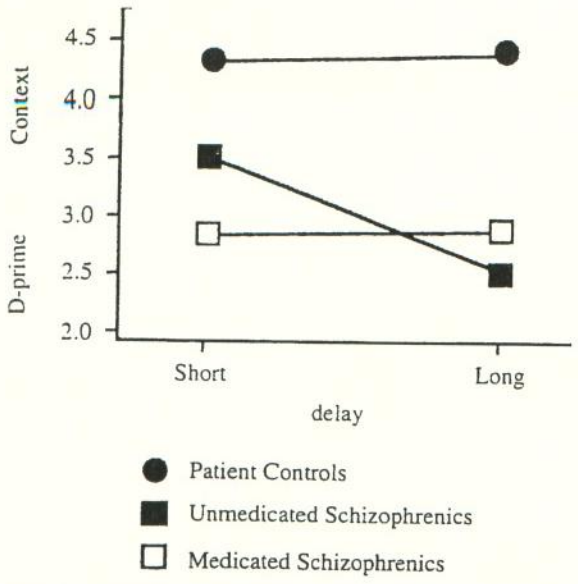


Figure 2. Context sensitivity in the AX-CPT task. 'D-prime context' refers to d-prime calculated for AX hits vs. BX false alarms (see text for explanation).

that might otherwise be considered to be more demanding and therefore subject to earlier failure, than simply maintaining information over a brief delay.

(*b*) *Empirical studies*

Recently, we have begun to use a variety of cognitive tasks to test predictions made by our models, and the theory upon which they are based. As an example of this work, we will focus on studies using the continuous performance test (CPT; Rosvold *et al.* 1956). This task was chosen to make contact between our computational models and an extensive literature concerning schizophrenic deficits (for reviews see Nuechterlein 1991; Cornblatt & Keilpk 1994). The task is simple enough to be able to simulate, yet generates a rich set of empirical findings that can be used to test detailed predictions made by the models. In this task, subjects observe sequences of letters presented one at a time, and must respond to a designated target. A version of this task in common use is the AX-CPT, which requires subjects to respond to a probe (e.g. the letter X), but only when it follows a designated cue (e.g. A). Performance in this task depends on the representation and maintenance of context information, insofar as the correct response to the probe depends on knowledge of the previous cue (A or not A). In standard versions of this task, target sequences (e.g. A-X) are typically low frequency (20%), and the delay between these is short (1 s). We modified this standard procedure, in order to probe both the inhibitory and memory functions associated with the processing of context. First, to assess 'inhibition,' we introduced a strong response bias by increasing the frequency of target sequences to 80%, with remaining trials evenly divided between two types of distractor sequences B-X and A-Y, where 'b' corresponds to any non-A stimulus, and 'y' to any non-X. This required subjects to respond to X in 7 of every 8 trials that it appeared, thus producing a strong bias to respond to X. This, in turn, required the use of context (non-A cue) to override this response. A schizophrenic failure to process context would predict an increase in false alarms on BX trials. This runs counter to the usual finding for schizophrenics: an increase in misses, but no increase in false alarms. It also circumvents concerns about a general lack of motivation or tendency to respond as a basis for schizophrenic failures, since we predict that they will make a greater number of responses in this condition (i.e. BX false alarms) as compared to controls. Finally, we manipulated the delay between the cue and the probe (1 vs 5 s), predicting that schizophrenic failures to actively maintain context would produce a selective increase in BX errors at the long vs short delay condition. Conversely, we predicted that control subjects would perform equally or better at the long delay, given the slower pace of the task (Parasuraman 1979).

We have completed two studies using this paradigm. In the first study (Servan-Schreiber *et al.* 1996) we tested medicated and unmedicated schizophrenics and patient controls. Both overall performance, and sensitivity to context (d-prime computed for AX hits and BX false alarms) were measured for all three groups. Both groups of schizophrenics showed overall worse performance than control subjects. However, there was a significant interaction between group and delay, with unmedicated schizophrenics exhibiting a predicted significant increase in AX misses and BX errors at the long delay, but not at the short delay (see figure 2). Groups did not differ in any other conditions of the task. Furthermore, the lack of a delay effect in either the control subjects, or the medicated schizophrenics who performed as poorly overall as the unmedicated schizophrenics, satisfies the criteria proposed by Chapman & Chapman (1978) for establishing psychometric equivalence across task conditions, and demonstrating a true differential deficit. Thus, unmedicated schizophrenics exhibited a selective and specific deficit, as predicted by our model of PFC function, its role in the processing of context, and its disturbance in schizophrenia.

We have continued to study subjects in this task, including patients who present with a first episode of psychosis, prior to treatment with neuroleptic medication. Diagnosis is established at 6 month follow-up, which yields a population of subjects with a confirmed diagnosis of schizophrenia, and another with non-schizophrenia-spectrum diagnoses. The latter represent a valuable control population, as they are closely matched in demographic characteristics to the first-episode schizophrenics and, most importantly, present in a form clinically indistinguishable from the schizophrenics. Therefore, any laboratory measures that can distinguish between schizophrenics and these subjects are likely to reflect fundamental pathophysiological features of this illness, and hold promise as valuable clinical instruments. Analysis of the performance of these subjects in the AX-CPT indicate that the neuroleptic-naive first-episode schizophrenics display the predicted pattern of deficits in this task, whereas the

non-schizophrenic subjects do not. This distinction in performance between schizophrenic and non-schizophrenic subjects, at a time when they are clinically indistinguishable, is potentially of great significance, strongly suggesting the specificity of our findings to schizophrenia, the sensitivity of our cognitive tasks, and the value of the theoretical approach that guided their design.

In another study (Cohen *et al.* 1996), we compared performance in the AX-CPT with two other cognitive tasks designed to probe processing of context, as a test of our hypothesis that a disturbance in the processing of context can provide a unified account of schizophrenic performance deficits across a variety of tasks. Performance in these tasks showed a modest but significant correlation in context-sensitive conditions (average $r = 0.38$, $P < 0.05$), but not control conditions (average $r = -0.02$, $P > 0.15$) matched for psychometric properties. These cross-task correlations are an important result, and contrast with historical failures to find such correlations among tasks that individually elicit schizophrenic performance deficits (Kopfstein & Neal 1972; Asarnow & MacCrimmon 1978). In previous studies, the detection of schizophrenic deficits, but failure to observe correlations, may have resulted from the use of tasks that included context-sensitive conditions (thus eliciting schizophrenic deficits), but measures of performance that were not sufficiently specific to these conditions. Aided by our computational models, our theory allowed us to decompose a set of disparate tasks, identify the conditions maximally sensitive to the processing of context, and demonstrate a predicted pattern of schizophrenic deficits across these tasks.

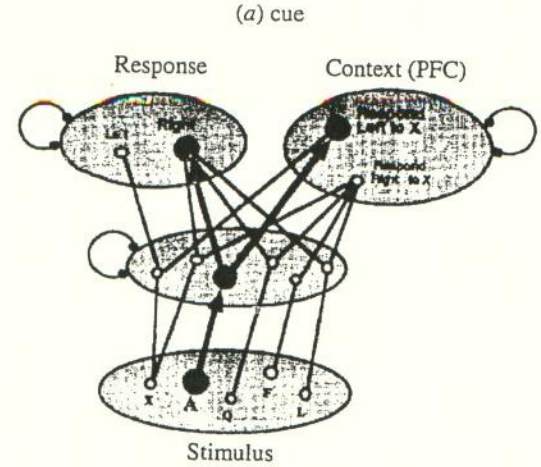
(*c*) *Recent advances in simulation models*

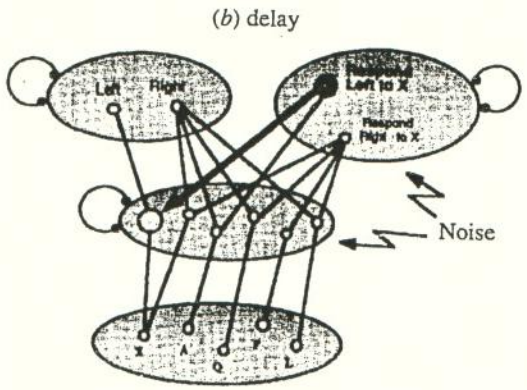
Our initial efforts involved models designed independently to address performance in a variety of different tasks. Recently, we have begun to adapt these models to conform to a more tightly constrained, and biologically plausible set of processing principles. These principles include: a) continuous (vs discrete) processing over time; b) interactivity (i.e. bi-directional connections) between processing layers; and c) the restriction of excitatory influences to between-module connections (information flow) and inhibitory influences to within-module connections (competition). Simulation work within this new, more highly integrated framework has addressed performance in a wide variety of tasks, including ones involving response competition, classical conditioning, covert spatial attention, and eye movements (Cohen *et al.* 1992; Cohen *et al.* 1994*b*; Armony *et al.* 1995; Forman & Cohen 1995). Most importantly, refinement of our earlier models has allowed us to address new, more detailed empirical data from tasks such as the AX-CPT.

For example, an important limitation of our original model of the CPT (Cohen & Servan-Schreiber 1992) was that it treated time in a discrete fashion (activation updates occurred only at whole-trial intervals). Therefore, while it could simulate accuracy data, it could not simulate the delay manipulations in our experiments, nor could it address the dynamics of performance (e.g. RT data). Our revision of the model overcomes these limitations, allowing us to simulate performance in continuous time (see Braver *et al.* 1995*a* for details). Thus, with the new model we can simulate reaction time as well as accuracy. To take full advantage of this feature of the new model, we have modified our version of the AX-CPT, so that responses are generated on every trial (one button for target, another for non-targets), thus providing RT data for correct as well as incorrect trials in every condition. We also added another trial condition to the task ('BY') that provides an important control measure by which to compare performance of the model against empirical data. The architectural addition of recurrent connectivity in the model allows us to directly examine the effects of delay on performance. Recurrent connections among units within the context layer allow it to actively maintain representations in the absence of external input. This makes it possible to interpose a delay between the cue and the probe (see figure 3), and then examine performance under conditions of different delay durations. We have tested the model with delays corresponding to both the short and long delays used in our empirical studies, and compared the model's performance with that of human subjects. As shown in figure 4, the model (solid line) fits a complex pattern of data concerning normal performance in this task (solid circles).

We have also examined the effects of degrading processing in the context layer of the model, by reducing the gain parameter of units in this layer (see Appendix 1). This manipulation produced a dramatic change in the profile of accuracy and RT data generated by the model across task conditions (dashed line in figure 4). These, in turn, make a number of new, and non-intuitive predictions concerning the accuracy and RT of schizophrenic vs. control subjects (Braver *et al.* 1995*b*). Among these is the prediction that, while schizophrenic RT s will be slower in most conditions, they will be comparable for correct responses in the AY condition (see figure 4). This prediction is especially interesting given the almost universal finding of slower RT s for schizophrenics in laboratory tasks. The model makes this prediction because under normal conditions, there is a relative slowing of correct (i.e. non-target) responses in AY trials. This a result of the fact that the context set up by the occurrence of the 'A' cue is maintained over the delay and serves to prime the target response, which is actually the incorrect response to make in the AY condition. This produces interference with the correct response, slowing RT. Under conditions of reduced gain, however, the context information is less reliably maintained. This produces less priming of the target response, less interference with the correct response, and thus less slowing of RT. We have recently collected data from six schizophrenic subjects in this version of the task (open circles in figure 4), which provide preliminary corroboration of these predictions.

We have also used this model to examine the effects that different degrees of PFC disturbance may have on task performance. Panel *a* of figure 5 shows empirical





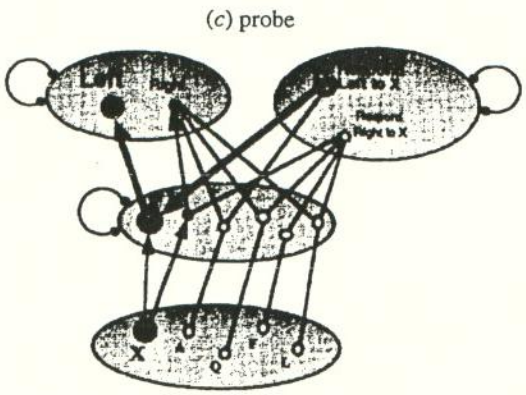


Figure 3. Revised model of the AX-CPT task. Graphic conventions are the same as figure 1. Recurrent self-excitatory connections for units in the context module are not shown (see Braver *et al.* 1995*a*). Panels show sequence of network states during a target sequence (A-X) trial. Panel (*a*) shows the state of the network during presentation of the cue, Panel (*b*) its state during the delay between the cue and the probe, and Panel (*c*) its state during presentation of the probe. Note that the input unit for the letter X has connections to both of the response units, and thus requires additional input from the context layer in order to elicit the correct response.

data from our original study, with the performance of unmedicated subjects shown separately for first episode and multi-episode patients. Panel *b* shows the

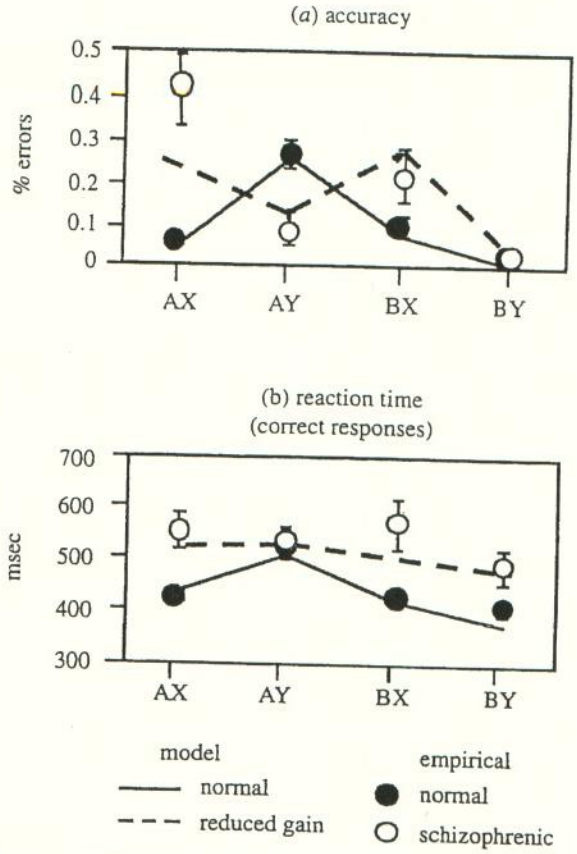


Figure 4. Simulated and empirical performance in the AX-CPT task (long delay condition Lines designate simulation results (solid: intact network; dashed: reduced gain in context layer)), circles designate empirical results (solid: normal subjects; open: pilot data from schizophrenic subjects).

patterns predicted by the model for normal performance and for two levels of impairment of the context layer. If we assume that disturbances of PFC function are progressive with course of illness, then the pattern of empirical results closely matches the predictions of the model. In the model, a moderate reduction of gain produces a pattern of results similar to that observed for first episode patients: performance is impaired at the long but not short delay, suggesting the presence of a 'memory' deficit (see appendix 2). A more severe disturbance in the model (i.e. a greater reduction of gain) produces further degradation of performance, similar to that observed for the multi-episode patients: performance is now impaired at the short as well as the long delay, suggesting the emergence of an additional 'inhibitory' deficit. Clinically, these disturbances in 'memory' and 'inhibition' might have been viewed as separate deficits, arising at different stages of illness. However, the model illustrates that, in fact, both can be explained in terms of a single, progressive disturbance, that varies only in severity. Of course, the results of these preliminary studies are by no means definitive (given that the empirical findings are confounded by a number of factors, including the potential chronic effects of neuroleptic administration that may endure even with temporary discontinuation). Nevertheless, they serve to highlight the potential value of computational

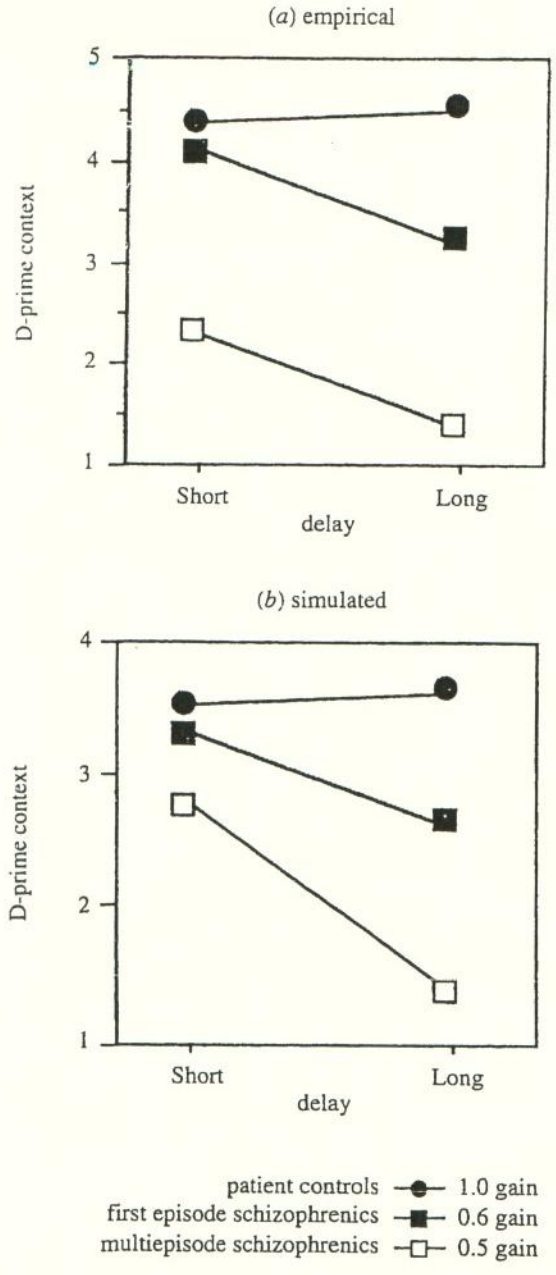


Figure 5. Effects of severity of disturbance on performance in the AX-CPT task. Panel (*a*) compares performance of normal subjects to schizophrenic patients suffering from their first psychotic episode and those at a later stage of illness. Panel (*b*) shows performance of the model under different degrees of gain reduction.

models, both in providing mechanistic and explicit accounts of empirical phenomena that may elude less formal accounts, and in making subtle, and sometimes even counterintuitive predictions regarding new phenomena.

3. CURRENT CHALLENGES AND NEW DIRECTIONS

We believe that we have made significant headway in using computational models to develop an explicit account of the mechanisms underlying the role of PFC in cognitive control, using these to account for cognitive disturbances in schizophrenia, and acquiring empirical support for the predictions made by our account. At the same time, the mechanisms we have implemented are still highly simplified, and lack several critical components necessary for a more general account of cognitive control. First, although we assume that a characteristic of PFC function is its ability to actively maintain context representations, our models have used relatively limited mechanisms for active maintenance, and have focused on tasks that do not include intervening stimuli that could interfere with maintained representations. Our models do not incorporate mechanisms for managing such interference, a function typically ascribed to PFC. Second, we have not specified how task-relevant information is identified, gains access to PFC, and is updated at appropriate junctures in processing. Finally, we have argued that context information can be used to support task-relevant processing in a variety of ways – by representing relatively abstract information (e.g. the relevant stimulus dimension in the Stroop task), or more detailed, specific information (e.g. a particular previous stimulus in the AX-CPT). This requires that the representational scheme within PFC be flexible enough to support a wide range of information. We have not yet specified what this scheme is, nor the principles that might characterize it.

To address these issues, we have begun to construct a new, more detailed theory, that builds on our previous work using neurobiologically plausible mechanisms to understand the role of PFC in cognitive control. The basic structure of this theory is as follows: In order for a system to be able to maintain information in the face of intervening, and potentially interfering stimuli, a mechanism is required to stabilize these representations. We propose that dopaminergic neuromodulation of PFC implements a gating function, that governs which representations gain access to PFC and when this occurs, thus protecting them from interference. We assume that this gating mechanism is driven by top-down projections from cortical areas and is able to learn when to update PFC based on general principles of reward-based learning, thus circumventing a potential regress in the locus of control. Furthermore, we argue that the need to maintain representations in PFC has direct implications for the nature of these representations. Specifically, we hypothesize that this leads to attractor-based representations that are both categorical and combinatorial, and that this provides a flexible scheme by which PFC can effectively bias processing in other parts of the system. Finally, we assume that, in order to perform controlled processing in novel domains, the system must be able to rapidly learn new associations, that can be used to guide behaviour. We hypothesize that this adaptability arises through interactions between PFC and the episodic storage capabilities of the hippocampal memory system.

Simulation studies are currently under way to establish the computational plausibility of this theory, and its ability to account for human performance in tasks that rely on cognitive control, that are more

complex than those we have studied to date. In the remainder of this chapter, we briefly review the components of this theory that address the mechanisms of active maintenance within PFC, and the nature of the representations involved. Our theory concerning the interactions between PFC and hippocampus is presented elsewhere (Cohen & O'Reilly 1996), and thus is omitted from this discussion.

(a) *Mechanisms of active maintenance within PFC*

A large body of converging evidence suggests that PFC plays a crucial role in maintaining task-relevant information in an active state. It is well established that populations of neurons in monkey PFC exhibit sustained, stimulus-specific activity during delays between a stimulus and a contingent response (Barone & Joseph 1989; Fuster & Alexander 1971; Goldman-Rakic 1987; Kubota & Niki 1971). Neuropsychological studies of frontally-damaged patients also provide strong support for the idea that PFC plays a role in active maintenance (Damasio 1985; Petrides & Milner 1982; Stuss *et al.* 1994), and a growing number of neuroimaging studies have strengthened this view (Cohen *et al.* 1994*a*; Grasby *et al.* 1993; Jonides *et al.* 1993; Petrides *et al.* 1993). However, the specific mechanisms by which neural activity in PFC is sustained has not been clearly elucidated.

(i) *Attractor systems*

A number of different computational approaches have been reported to deal with tasks requiring short-term storage in neurobiologically plausible terms. One of the most common is the use of recurrent connections between neuron-like processing units (Hopfield 1982). With sufficiently strong recurrent connections, networks of these units will develop 'attractors', defined as stable states in which a particular pattern of activity is maintained. Thus, attractors can be used to actively store information. Indeed, a number of computational models have demonstrated that both physiological and behavioural data regarding PFC function can be captured in simple tasks using an attractor-based scheme (Braver *et al.* 1995*a*; Dehaene & Changeux 1989; Zipser *et al.* 1993). However, in simple attractor systems, the state is strongly determined by its inputs. Thus, presentation of a new input will interfere with the state of the system and drive it into a new attractor state. Although attractor networks can be configured to display resistance to disruption from input (i.e. hysteresis), this impairs their ability to be updated in a precise and flexible manner. One way in which attractor networks can overcome these difficulties is through the addition of a gating mechanism. Such systems only respond to inputs, and change their attractor state, when the 'gate' is opened. Below, we examine the possibility that the DA system provides a gating signal in PFC which acts to maintain context representations in an active, stable, and flexible manner.

(ii) *Modulatory effects of DA.*

A number of lines of evidence suggest that DA acts in a modulatory fashion in PFC, which is consistent with a role for DA in gating access to active memory. DA agonists have been found to improve memory performance in humans (Luciana *et al.* 1992), while in primates DA antagonists interfere with performance in delayed-response tasks (Sawaguchi & Goldman-Rakic 1991) and directly affect PFC neuronal activity (Williams & Goldman-Rakic 1995). Electron microscopy studies of the local connectivity patterns of DA in PFC have revealed that DA typically makes triadic contacts with prefrontal pyramidal cells and excitatory afferents, and also with inhibitory interneurons (Lewis *et al.* 1992; Williams & Goldman-Rakic 1993). The triadic synaptic complexes formed in PFC suggest that DA can modulate both afferent input and local inhibition. Electrophysiological data support this view, indicating that DA potentiates both afferent excitatory and local inhibitory signals (Chiodo & Berger 1986; Penit-Soria *et al.* 1987).

(iii) *DA as a gating signal.*

In our previous work, we have simulated the modulatory effects of DA as a change in the gain (slope) of the activation function of processing units in PFC (Cohen & Servan-Schreiber 1992; Cohen & Servan-Schreiber 1993; Servan-Schreiber *et al.* 1990), thus influencing the active maintenance of information. In this work, we assumed that DA effects were prolonged (tonic), consistent with the widely-held assumption that neuromodulatory systems (such as DA, NE, 5-HT) are slow-acting, diffuse, and non-specific in informational content. However, recent findings suggest a revision of this view. Schultz and colleagues (Schultz 1986, 1992) observed in behaving primates that during learning of a spatial delayed response task, stimuli that failed to activate ventral tegmental area DA neurons on initial presentation, came to elicit transient activity when the animal learned that they were significant for the task. Specifically, Schultz's group observed transient, stimulus-locked activity in response to stimuli that were themselves unpredictable, but predicted later meaningful events. This is precisely the timing required for this information to be gated into and maintained in active memory. These findings, together with the modulatory effects of DA on PFC afferents, suggest that DA may serve as a gating signal for attractor networks housed within PFC.

(iv) *DA as a learning signal.*

At the same time, DA is widely thought to be involved in reward learning (Wise & Rompre 1989). Indeed, in Schultz's experiments, DA activity was found for cues that were predictive of a reward. Montague *et al.* (1996) have built on these findings, and recently reported a computational model that treats DA as a widely distributed error signal, driving the learning of temporal predictors of reinforcement. Intriguingly, the parameter they used to simulate the learning effects of DA is formally equivalent to the gain

parameter used in our model to simulate its modulatory effects on active maintenance in PFC (Thimm *et al.* 1996). Furthermore, it is well established that DA neurons in VTA receive strong cortical projections, particularly from PFC (Oades & Halliday 1987). These observations have lead us to the following refinements of our original theory (Cohen & Servan-Schreiber 1992): a) DA implements gating and learning functions, both of which rely on the same mechanism; b) the gating function is used to regulate the access of information to active memory within PFC, and protect it from interference; c) the learning function produces training signals in the cortex; d) both of these functions can be elicited by descending cortical signals; e) the coincidence of the gating and learning signals produces cortical associations between the information being gated, and a triggering of the gating signal in the future. The power of this theory, if confirmed, is that it will describe a system that has both the capacity to control its behaviour (by gating and stabilizing representations within PFC) and, critically, the ability to learn how and when to do so on its own, thus avoiding the perennial problem of a homunculus in most theories of executive control.

In recent simulation studies, we have implemented a gating mechanism, and shown that it can successfully account for the phenomena addressed by our earlier models involving active maintenance. However, the assumption that DA has phasic, in addition to tonic effects in PFC changes the dynamics of processing in our simulations, and makes a number of new predictions regarding both normal and schizophrenic performance in delay tasks. It also allows us to simulate performance in tasks not possible with our previous models (e.g. ones involving intervening distractors). Our work has currently turned to the empirical validation of these simulation studies, and an integration of the gating mechanism with one for reward learning.

(*b*) *Nature of representations within PFC*

In addition to maintaining representations in active memory, our theory requires that PFC be able to use these to control behaviour, by biasing processing in other parts of the system. This was illustrated in the Stroop model, where activation of a unit representing the colour dimension biased processing in favour of colour naming over word reading. Here, the relevant representation was a stimulus dimension (colour vs word form). However, as we have noted, this might be any bit of information that specifies the dimensions of information relevant to the task. Important questions remain about how this information is actually represented, and how representations can emerge that provide the flexibility characteristic of human behaviour, without requiring unlimited capacity. It would be unreasonable to expect, for example, that a separate, dedicated unit (or set of units) exists to bias processing for every task that people are able to perform. In this section, we propose a representational scheme that we believe can address these issues. We consider two functional requirements that constrain the nature of representations within PFC: a) the need to be self-maintaining; and b) the ability to flexibly bias processing in the rest of the neocortex. Our hypothesis is that these dual constraints shape the nature of PFC representations in a synergistic fashion: The constraint imposed by the need to be self-maintaining leads to representations that can be used in a combinatorial, and therefore highly flexible way.

i) *Independent*

The strong recurrent excitation required for self-maintenance imposes the constraint that if two representations are to be reliably distinguishable, they must also be independent of one another. This can be illustrated by the simple case shown in figure 6, where three processing units are used to store three different representations. Panel (*a*) shows how this can be done using distributed representations of features (e.g. keyboard, monitor and speaker) that efficiently encode the shared structure among the items (terminal, synthesizer, and television). This supports spreading-activation like computations (pattern completion, content-addressable memory, etc.) that are considered to be characteristic of posterior neocortex (e.g. McClelland *et al.* 1995; Plaut *et al.* 1996). However, problems arise when recurrent connections are made strong enough to maintain activity in the absence of external input. This can be seen in the example, by noting that no value of excitatory weights between these features would allow one pair of units to remain active without activating the third. This can be solved by making each feature independent of the others, with its own recurrent self-excitation (Panel (*b*)). However, this eliminates the representation of the items (e.g. terminal) as a related (connected) set of features (keyboard and monitor). Items can still be stored by activating the appropriate combination of features, but without the benefits of spreading activation among them. Thus, this scheme sacrifices the representation of the internal structure of each item, and its relationship to others. An alternative is to dedicate a separate unit to each item (e.g. a new unit for terminal), since semantic structure is already lost. This would add units to the system, but might only occur for frequently encountered or important items. Either way, the critical point is that there is a trade-off between the capacity for self-maintenance and the ability to represent structured (semantic) relationships among features and/or items.

(ii) *Flexible*

Representations in PFC must be flexible enough to account for the variety of tasks of which humans are capable. We propose that this flexibility is provided by the independence of self-maintaining representations. Precisely because they are independent of one another, there are no constraints on which representations can be active at the same time, and so they can be used in arbitrary combinations. The ability to rapidly adapt behaviour to novel situations (e.g. problem solving) requires just this capacity for the arbitrary re-combination of existing knowledge, and the idea that this capacity resides within PFC is consistent with the neuropsychological literature suggesting that damage

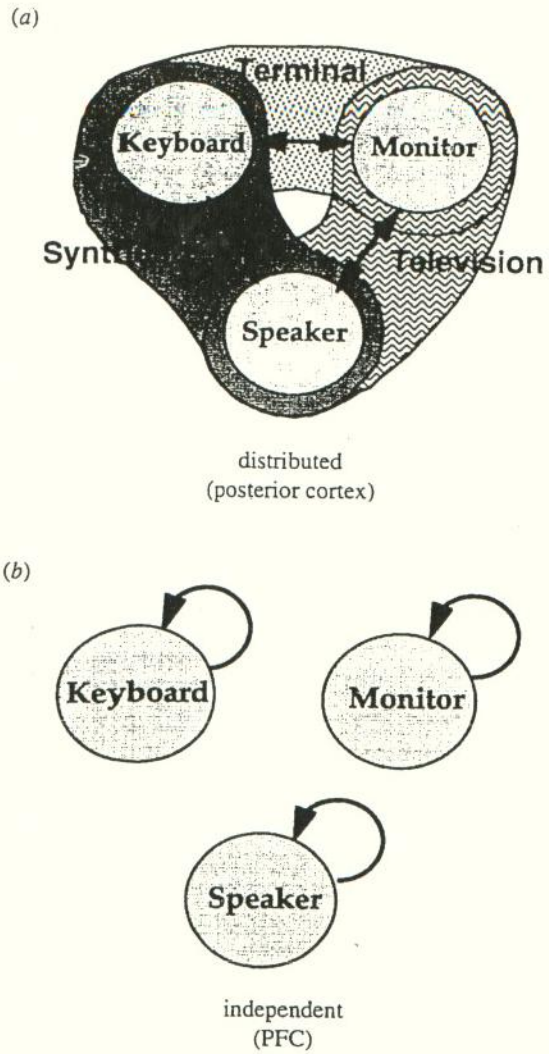


Figure 6. An example of distributed versus independent representations. Panel (*a*) shows the associative connections between features used in the distributed representation of three different concepts. Panel (*b*) shows these units as independent features, which can be used to support self maintenance and combinatorial representations, but at the expense of capturing the associative relationships between the concepts being represented.

to PFC impairs this flexibility, and problem solving abilities (Shallice & Burgess 1991*b*). However, while combinatoriality is a necessary requirement for flexibility, it is not sufficient. The representations themselves must correspond in appropriate ways to information stored in the rest of system, so that their activation can bias processing in behaviourally useful ways.

(iii) *Categorical*

We assume that the content of PFC representations is constrained by the functional characteristics we have just described – self-maintaining and independent – in interaction with posterior neocortex, under conditions of Hebbian learning. The independence of PFC representations frees them to become associated with arbitrary dimensions of information in the neocortex. We suggest that these effects will generate representations within PFC that are categorical in character. That is, PFC representations will become associated with the central tendency of similar representations within the neocortex. We use the term 'categorical' here guardedly, as we do not want to suggest that all representations within PFC correspond directly to 'basic level', or even verbalizable categories. For example, we imagine that PFC representations will exist for very general categories of information, such as stimulus dimensions (e.g. colour, shape, size, etc.) or natural and/or functional kinds (e.g. animals, tools), but also much more specific ones (e.g. a particular colour, word, or person which, though more specific, also correspond to invariants in the world) and non-obvious or abstract ones (i.e. ones that do not correspond directly to recognizable concepts). Finally, we assume that a primary determinant of these representations is their behavioural relevance – that is, their ability to successfully bias the interpretation of stimuli and the selection of responses to produce task-appropriate behaviour. Thus, dimensions that are frequent, or particularly relevant to behaviour are most likely to be represented.

(iv) *Properties of PFC vs posterior neocortex*

The preceding discussion highlights features of representations and processing within PFC that distinguish it from other parts of association neocortex. We assume that posterior neocortex uses distributed representations, and attractor states that are not constrained to be self-maintaining. Thus, we view PFC and posterior neocortex as having complementary properties. Each is specialized along particular dimensions of representation and processing, that involve tradeoffs in functionality (e.g. self-maintaining and flexible vs representation of structured knowledge). We propose that the benefit of such a scheme is fully realized only when these systems interact with one another. Semantically rich, distributed representations in posterior neocortex are not only required to encode the statistical structure of the environment, and support the processing pathways responsible for task performance, but also serve as the base for developing the appropriate set of representations in PFC. These, in turn, provide representations that enable the flexible biasing of representations in posterior neocortex, and maintenance over intervals required to guide task performance in a temporally extended manner.

4. SUMMARY AND CONCLUSIONS

The prefrontal cortex is the region of the brain most significantly expanded in humans compared to other species, and appears to be at the heart of those faculties that we consider to be uniquely human, such as execution of the complex behaviours involved in planning and problem solving. Consistent with these observations, schizophrenia, which is thought to involve frontal dysfunction, is an illness that is unique to humans. At best, the effort to understand the function of PFC, and its role in schizophrenia promises to be a complex endeavor, and will require the most powerful conceptual tools we have available. We believe that computational modelling represents one

such tool. We began this article by reviewing initial progress in the use of computational models to understand the role of PFC in cognitive control, and its impairment in schizophrenia. We illustrated how such models can provide new insights into the mechanisms underlying cognitive control, can make new predictions about the behavioural performance of both normal and schizophrenic subjects, and provided examples of empirical support for such predictions. We then identified some of the limitations of our current approach, and outlined a set of new principles that we believe can be used to define a more complete theory of PFC function and cognitive control. According to this theory, PFC governs processing in a top-down manner, while at the same time remaining responsive to bottom-up input from other parts of the system. We believe that this interplay of bottom-up and top-down processing, that we hypothesize is mediated by the DA system, can give rise to a system of 'regulated interactivity' that accounts for the full flexibility of control exhibited in human behaviour. These are bold claims. However, as our previous work has shown, by pursuing such ideas within a computational framework, we are able to make them explicit in simulation models. This not only provides a check on their conceptual validity, but also allows us to explore, in detail, their implications for behaviour. Success in this effort would not only provide insights into the mechanisms underlying some of our highest and uniquely human faculties, but might also allow to us to contend better with their failings, which appear to lie at the heart of our deepest vulnerabilities.

We thank David Servan-Schreiber, Marius Usher, Jay McClelland, and David Plaut, for the many hours of productive discussion that we have shared with them, and for their comments and insights which helped substantially to shape the ideas presented in this chapter. We also thank Deanna Barch, Cameron Carter, Grace Nah and Vijoy Abraham, without whom the empirical studies described may never have been completed.

This work was supported by a NIMH Physician Scientist Award (MH00673), a NIMH FIRST Award (MH47073), and a research grant from the Scottish Rite Schizophrenia Research Program, N.M.J., U.S.A. to the first author, as well by a NIMH Program Project (MH47566) and a NIMH Center (MH45156) in which the first author is a participant.

REFERENCES

Abramczyk, R. R., Jordan, D. E. & Hegel, M. 1983 'Reverse' Stroop effect in the performance of schizophrenics. *Perceptual and Motor Skills* **56**, 99–106.

Anderson, J. R. 1983 *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Armony, J. L., Servan-Schreiber, D., Cohen, J. D. & LeDoux, J. E. 1995 An anatomically constrained neural network model of fear conditioning. *Behavioural Neuroscience* 109(2), 1–12.

Asarnow, R. F. & MacCrimmon, X. 1978 Residual performance deficit in clinically remitted schizophrenics: a marker of schizophrenia? *J. Abnormal Psychology* 87, 597–608.

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

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

Bianchi, L. 1922 *The mechanism of the brain and the function of the frontal lobes*. Edinburgh: Livingstone.

Braver, T. S., Cohen, J. D. & Servan-Schreiber, D. 1995*a* A computational model of prefrontal cortex function. In *Advances in neural information processing systems* (ed. D. S. Touretzky, G. Tesauro & T. K. Leen), vol. 7, pp. 141–148. Cambridge, MA: MIT Press.

Braver, T. S., Cohen, J. D. & Servan-Schreiber, D. 1995*b* Neural network simulations of schizophrenic performance in a variant of the CPT-AX: a predicted double dissociation. *Schizophrenia Res.* 15(1–2), 110.

Carter, C. S., Robertson, L. C., Nordahl, T. E., O'Shora-Celaya, L. J. & Chaderjian, M. C. 1993 Abnormal processing of irrelevant information in schizophrenia: the role of illness subtype. *Psychiatry Res.* **4**, 178–26.

Chapman, L. J. & Chapman, J. P. 1978 The measurement of differential deficit. Journal of *Psychiatric Res.* **14**, 303–311.

Chapman, L. J., Chapman, J. P. & Miller, G. A. 1964 A theory of verbal behaviour in schizophrenia. In *Progress in experimental personality research* (ed. B. A. Maher), vol. 4, pp. 135–167. New York: Academic Press.

Chiodo, L. & Berger, T. 1986 Interactions between dopamine and amino-acid induced excitation and inhibition in the striatum. *Brain Res.* **375**, 198–203.

Cohen, J. D. & Carter, C. S. 1996 Scizophrenic performance in the Stroop task: empirical findings and a theoretical model of differences between faciliation and interference effects. (In preparation.)

Cohen, J. D. & O'Reilly, R. C. 1996 A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In *Prospective memory: theory and applications* (ed. M. Brandimonte, G. Einstein & M. McDaniel). Hillsdale, NJ: Erlbaum.

Cohen, J. D. & Servan-Schreiber, D. 1992 Context, cortex and dopamine: a connectionist approach to behaviour and biology in schizophrenia. *Psychological Review* **99**, 45–77.

Cohen, J. D. & Servan-Schreiber, D. 1993 A theory of dopamine function and cognitive deficits in schizophrenia. *Schizophrenia Bull.* **19**(1), 85–104.

Cohen, J. D., Dunbar, K. & McClelland, J. L. 1990 On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Rev.* **97**(3), 332–361.

Cohen, J. D., Servan-Schreiber, D. & McClelland, J. L. 1992 A parallel distributed processing approach to automaticity. *Am. J. Psychology* **105**, 239–269.

Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B. J., Servan-Schreiber, D. & Noll, D. C. 1994*a* Activation of prefrontal cortex in a nonspatial working memory task with functional MRI. *Human Brain Mapping* **1**, 293–304.

Cohen, J. D., Romero, R. D., Farah, M. J. & Servan-Schreiber, D. 1994*b* Mechanisms of spatial attention: the relation of macrostructure to microstructure in parietal neglect. *J. Cognitive Neuroscience* 6(4), 377–387.

Cohen, J. D., Barch, D. M., Carter, C. S. & Servan-Schreiber, D. 1996 Schizophrenic deficits in the processing of context: converging evidence from three theoretically motivated cognitive tasks. (Submitted.)

Cornblatt, B. A. & Keilpk, J. G. 1994 Impaired attention, genetics, and the pathophysiology of schizophrenia. *Schizophrenia Bull.* 20(1), 31–62.

Damasio, A. R. 1985 The frontal lobes. In *Clinical neuropsychology* (ed. K. M. Heilman & E. Valenstein), pp. 339–375. New York: Oxford University Press.

Dehaene, S. & Changeux, J. P. 1989 A simple model of

prefrontal cortex function in delayed-response tasks. *J. Cognitive Neurosci.* **1**, 244–261.

Forman, S. D. & Cohen, J. D. 1995 Modelling saccadic eye movements in schizophrenia: insights into memory mechanisms. *Schizophrenia Res.* **15**(1–2), 175.

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

Gathercole, S. E. 1994 Neuropsychology and working memory: a review. *Neuropsychology* **8**(4), 494–505.

Geraud, G., Arne-Bes, M. C., Guell, A. & Bes, A. 1987 Reversibility of hemodynamic hypofrontality in schizophrenia. *J. Cerebral Blood Flow Metabolism* **7**, 9–12.

Goldman-Rakic, P. S. 1987 Circuitry of primate prefrontal cortex and regulation of behaviour by representational memory. In *Handbook of physiology – the nervous system* (ed. F. Plum & V. Mountcastle), vol. 5, pp. 373–417. Bethesda, MD: American Physiological Society.

Goldman-Rakic, P. S. 1991 Prefrontal cortical dysfunction in schizophrenia: the relevance of working memory. *Psychopathology and the Brain* 1–23.

Grasby, P. M., Frith, C. D., Friston, K. J., Bench, C., Frackowiak, R. S. J. & Dolan, R. J. 1993 Functional mapping of brain areas implicated in auditory-verbal memory function. *Brain* **116**, 1–20.

Hopfield, J. J. 1982 Neural networks and physical systems with emergent collective computational abilities. *Proc. Natn. Acad. Sci.* **79**, 2554–2558.

Jonides, J., Smith, E. E., Koeppe, R. A., Awh, E., Minoshima, S. & Mintun, M. A. 1993 Spatial working memory in humans as revealed by PET. *Nature* **363**, 623–625.

Karoum, F., Karson, C. N., Bigelow, L. B., Lawson, W. B. & Wyatt, R. J. 1987 Preliminary evidence of reduced combined output of dopamine and its metabolites in chronic schizophrenia. *Arch. General Psychiatry* **44**(July), 604–607.

Kimberg, D. Y. & Farah, M. J. 1993 A unified account of cognitive impairments following frontal lobe damage: the role of working memory in complex, organized behaviour. *J. Experimental Pscyhology: General* **122**(4), 411–428.

Kopfstein, J. H. & Neal, J. M. 1972 A multivariate study of attention dysfunction in schizophrenia. *J. Abnormal Psychology* **3**, 294–298.

Kolb, B. & Whishaw, I. Q. 1983 Performance of schizophrenia patients on tests sensitive to left or right frontal, temporal. *J. Nervous Mental Disease* **171**(7), 435–443.

Kornetsky, C. & Orzack, M. H. 1978 Physiological and behavioural correlates of attention dysfunction in schizophrenic patients. *J. Psychiatric Res.* **14**, 69–79.

Kraeplin, E. 1950 *Dementia praecox and paraphrenia* (transl. J. Zinkin). New York: International Universities Press, Inc.

Kubota, K. & Niki, H. 1971 Prefrontal cortical unit activity and delayed alternation performance in monkeys. *J. Neurophysiology* **34**, 337–347.

Levin, S. 1984 Frontal lobe dysfunctions in schizophrenia. II. Impairments of psychological and brain functions. *J. Psychological Res.* **18**, 57–82.

Lewis, D. A., Hayes, T. L., Lund, J. S. & Oeth, K. M. 1992 Dopamine and the neural circuitry of primate prefrontal cortex: implications for schizophrenia research. *Neuropsychopharmacology* **6**(2), 127–134.

Luciana, M., Depue, R. A., Arbisi, P. & Leon, A. 1992 Facilitation of working memory in humans by a D_2 dopamine receptor agonist. *J. Cognitive Neurosci.* **4** (1), 58–68.

Luria, A. R. 1969 Frontal lobe syndromes. In *Handbook of clinical neurology* ed. P. J. Vinken & G. W. Bruyn, vol. 2. pp. 725–757. New York: Elsevier.

Manschreck, T., Maher, B. A., Milavetz, J. J., Ames, D., Weisstein, C. C. & Schneyer, M. L. 1988 Semantic priming in thought disordered schizophrenic patients. *Schizophrenic Res.* 61–66.

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. 1995 Why there are complementary learnings systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Rev.* **102**, 419–457.

Montague, P. R., Dayan, P. & Sejnowski, T. J. 1996 A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947.

Nuechterlein, K. H. 1991 Vigilance in schizophrenia and related disorders. *In Handbook of schizophrenia* (ed. S. R. Steinhauer, J. H. Gruzelier & J. Zubin), vol. 5. *Neuropsychology, psychophysiology, and information processing*, pp. 397–433. Amsterdam: Elsevier.

Oades, R. D. & Halliday, G. M. 1987 Ventral tegmental (A10) system: neurobiology. 1. Anatomy and connectivity. *Brain Res. Rev.* **12**, 117–165.

Parasuraman, R. 1979 Memory load and event rate control sensitivity decrements in sustained attention. *Science* **205**, 924–927.

Park, S. & Holzman, P. S. 1992 Schizophrenics show spatial working memory deficits. *Arch. General Psychiatry* **49**, 975–982.

Penit-Soria, J., Audinat, E. & Crepel, F. 1987 Excitation of rat prefrontal cortical neurons by dopamine: an *in vitro* electrophysiological study. *Brain Res.* **425**, 263–274.

Petrides, E. & Milner, B. 1982 Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia* **20**, 249–262.

Petrides, M. E., Alivisatos, B., Evans, A. C. & Meyer, E. 1993 Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. *Proc. Natn. Acad. Sci.* **90**, 873–877.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. 1996 Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Rev.* **103**, 56–115.

Roberts Jr, R. J., Hager, L. D. & Heron, C. 1994 Prefrontal cognitive processes: working memory and inhibition in the antisaccade task. *J. Experimental Psychology: General* **123**(4), 374–393.

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D. & Beck, L. H. 1956 A continuous performance test of brain damage. *J. Consulting Psychology* **20**(5), 343–350.

Rumelhart, D. E. & McClelland, J. L. 1986 *Parallel distributed processing: explorations in the microstructure of cognition*, vols 1 and 2. Cambridge, MA: MIT Press.

Sawaguchi, T. & Goldman-Rakic, P. S. 1991 D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science* **251**, 947–950.

Schultz, W. 1986 Responses of midbrain dopamine neurons to behavioural trigger stimuli in the monkey. *J. Neurophysiology* **56**, 1439–1462.

Schultz, W. 1992 Activity of dopamine neurons in the behaving primate. *Seminars in Neurosciences* **4**, 129–138.

Servan-Schreiber, D., Cohen, J. D. & Steingard, S. 1996 Schizophrenic performance in a variant of the CPT-AX: a test of theoretical predictions concerning the processing of context. *Arch. General Psychiatry.* (In the press.)

Servan-Schreiber, D., Printz, H. & Cohen, J. D. 1990 A network model of catecholamine effects: gain, signal-to-noise ratio, and behaviour. *Science* **249**, 892–895.

Shallice, T. 1982 Specific impairments of planning. *Phil. Trans. R. Soc. Lond.* B298, 199–209.

Shallice, T. & Burgess, P. 1991*a* Higher-order cognitive impairments and frontal lobe lesions in man. In *Frontal lobe function and dysfunction* ed. H. S. Levin, H. M. Eisenberg & A. L. Benton . New York: Oxford University.

Shallice, T. & Burgess, P. W. 1991*b* Deficits in strategy application following frontal lobe damage in man. *Brain* **114**, 727–741.

Storms, L. H. & Broen, W. E. 1969 A theory of schizophrenic behavioural disorganization. *Arch. General Psychiatry* **20** (Feb), 129–144.

Stuss, D. T., Eskes, G. A. & Foster, J. K. 1994 Experimental neurospychological studies of frontal lobe function. In *Handbook of neuropsychology* (ed. F. Boller & J. Grafman), vol. 9. Amsterdam: Elsevier.

Thimm, G., Moerland, P. & Fusler, E. 1996 The interchangeability of learning rate and gain in backpropagation networks. *Neural Computation* **8**, 451–469.

Wapner, S. & Krus, D. M. 1960 Effects of lysergic acid diethylamide, and differences between normals and schizophrenics, on the Stroop colour-word test. *J. Neuropsychiatry* (Nov/Dec), 76–81.

Weinberger, D., Berman, K. & Illowsky, B. 1988 Physiological dysfunction of the dorsolateral prefrontal cortex. III. A new cohort and evidence for a monoaminergic mechanism. *Arch. General Psychiatry* **45**, 609–615.

Weinberger, D. R., Berman, K. F. & Zec, R. F. 1986 Physiological dysfunction of dorsolateral prefrontal cortex in schizophrenia. I. Regional cerebral blood flow evidence. *Arch. General Psychiatry* **43**, 114–125.

Williams, G. V. & Goldman-Rakic, P. S. 1995 Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature* **376**, 572–575.

Williams, M. S. & Goldman-Rakic, P. S. 1993 Characterization of the dopaminergic innervation of the primate frontal cortex using a dopamine-specific antibody. *Cerebral Cortex* **3**(May–June), 199–222.

Wise, R. A. & Rompre, P.-P. 1989 Brain dopamine and reward. *A. Rev. Psychology* **40**, 191–225.

Wynne, L. C., Cromwell, R. L. & Matthysse, S. 1978 *The nature of schizophrenia new approaches to research and treatment.* New York: John Wiley and Sons, Inc.

Wysocki, J. J. & Sweet, J. I. 1985 Identification of brain damaged, schizophrenic, and normal medical patients using a brief neuropsychological screening battery. *Int. J. Clinical Neuropscyhology* **7**(1), 40–44.

Zipser, D., Kehue, B., Littlewort, G. & Fuster, J. 1993 A spiking network model of short-term active memory. *J. Neurosci.* **13**, 3406–3420.

Zubin, J. 1975 Problem of attention in schizophrenia. *In Experimental approaches to psychopathology* (ed. M. I. Kietzman, S. Sutton & J. Zubin), pp. 139–166. New York: Academic Press.

APPENDIX 1

The gain parameter regulates the responsivity of units to their input, so that a reduction of gain degrades the fidelity of representations in this layer. In our previous work, we have used this parameter to approximate the influence that brain catecholamines have on network processing characteristics and behaviour (Servan-Schreiber *et al.* 1990). The relationship of dopamine to gain, and its influence on PFC function is discussed in greater detail in the last section of this paper. In the current model, gain was reduced exactly the same amount as in our previous work (from 1.0 to 0.6), to simulate the effects of reduced dopamine in PFC that we and others have hypothesized to occur in schizophrenia (Cohen & Servan-Schreiber 1992, 1993; Geraud *et al.* 1987; Karoum *et al.* 1987; Levin 1984; Weinberger *et al.* 1988).

APPENDIX 2

The reader will note that, even at the long delay, BX errors still involve a failure of inhibition, in addition to a memory failure. Thus, impairment of first-episode patients in this condition could reflect the dual burden of a requirement for memory + inhibition, rather than a specific sensitivity to the memory load. To address this concern, we have also examined performance in the AY condition. This provides a measure of memory function unconfounded by inhibitory demands: AY trials will induce a tendency to respond only to the extent that the subject has an intact representation of context, which primes a target response; in the absence of such information, there should be no such tendency to respond. This is confirmed in our simulations, which exhibit a reduction of such errors at the long vs short delay, even with a moderate disturbance in the processing of context, and an accentuation of this effect at both delays when the disturbance to the context layer is worsened. These effects have the added virtue of predicting improvements of performance as the severity of disturbance is increased. Our empirical data conform to these predictions: first episode subjects showed a significant number of AY errors at the short delay (16%), comparable to the rate observed in control populations, indicating intact processing of context. At the long delay, however, they showed a reduction of this effect (6% errors), indicating a memory deficit unconfounded by inhibitory demands. The multi-episode subjects showed a general accentuation of this effect (7% and 1% AY errors at the long and short delays, respectively).