A hard wired model of coupled frontal working memories for various tasks

Oury Monchi a,c,*, John G. Taylor a,b

a Centre for Neural Networks, King's College London, Strand, London WC2R 2LS, UK
b Institute for Medicine, KFA-Jülich, 52425 Jülich, Germany
c McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, 3801 University St., Montreal, H3A 2B4, QC, Canada

Received 3 March 1997; received in revised form 3 February 1998; accepted 23 April 1998

Abstract

The majority of previous models of working memory (WM), rule production and strategy decisions have not taken into account the known anatomy of the human brain. Here, a simple hard wired neural network is developed which investigates the role of the dorsolateral prefrontal cortex (DLPFC), the ventrolateral prefrontal cortex (VL) and the Anterior Cingulate Gyrus (ACG), together with their associated basal gangliothalamocortical loops, in solving the classical delayed response task (DRS), Delayed visual matching (DRO) and the Wisconsin card sorting task (WCST). This is done by constraining the model by some of the known connectivity within these areas. Lesions are also applied to the model thereby providing possible interpretations for some of the deficits observed in various kinds of patients while performing these tasks. © 1999 Elsevier Science Inc. All rights reserved.

1. Introduction

The psychological concept of working memory (WM) was proposed in [1] as a means of holding an active representation of features and is also meant to be involved in other cognitive functions such as learning, reasoning and planning. Cognitive and attentional frontal tasks such as the delayed response task (DRS), delayed visual matching (DRO) and Wisconsin card

* Corresponding author. Fax: +1-514 398 8948; e-mail: oury@bic.mni.mcgill.ca

0020-0255/99/$19.00 © 1999 Elsevier Science Inc. All rights reserved.
PI: S0020-0255(98)00051-8
sorting task (WCST) rely on it. It is known from experimental studies on animals (e.g. [2]) and brain imaging studies in humans (e.g. [3]) that the prefrontal cortex (PFC) is involved in WM. However, most tasks involving WM activate other structures besides the PFC such as the basal ganglia and the thalamus, as has been shown by positron emission tomography (PET) studies [4]. The involvement of basal ganglia in tasks involving WM has been supported by anatomical evidence such as the ones presented in [5]. It is further substantiated by the cognitive deficits observed in patients with Parkinson’s disease (PD) [6,7], which resemble those seen after frontal-lobe lesions. More specifically PD patients at the early stages of the disease have shown deficits in tasks involving WM, planning and set-shifting. It is thought that the neurological abnormality in these patients is confined to the striatum [8].

At the physiological level, WM relies on the sustained activity from 2 to 60 s, observed in the memory cells of the PFC [9]. There are two possible recurrent circuits which could provide for the long lasting activity in the memory cells of the PFC. One of these circuits involves the cortico-cortical reciprocal connections which occur either locally within PFC or from the inferior temporal cortex and the posterior parietal cortex to the PFC [10,11]. The other involves the reciprocal excitatory connections between PFC and mediodorsal thalamus (MD) [5] together with the connections from the PFC to the basal ganglia and the latter to MD. The system described in the last sentence is sometimes referred as a basal ganglia-thalamocortical loop. The model presented in this paper is based on the hypothesis that the basal ganglia-thalamocortical loops play an important part in the cognitive functions that rely on WM.

1.1. The anatomical and physiological substrates of WM

The frontal lobe (FC) of the human and non-human primates include the premotor cortex, the motor cortex, the cingulate cortex and the PFC. The PFC is the most anterior part of these structures, the motor cortex being the most posterior. There appear to be a hierarchy with respect to the complexity of information processing within the FL with the more posterior parts dealing with lower order information and the anterior parts dealing with higher order concepts. There exists some discrepancies regarding the subdivisions of the PFC and their functions. It is traditionally subdivided into two regions the dorsolateral and the orbitofrontal. Some researchers refer to three subdivisions of the PFC (e.g. [12]), the dorsolateral, the medial and orbitofrontal, and the ventrolateral cortex (VL). We will adopt here this latter classification.

Experimental anatomical studies have shown that the cortico-cortical pathways from posterior parietal cortex to DLPFC are responsible for sending
processed visual information with respect to location and, similarly, pathways from the inferior temporal cortex to VL with respect to object features [13]. These pathways would thereby excite the relevant areas in PFC and provide for the activity observed in the visual tonic cells of the PFC while a feature is being cued. There is not much disagreement regarding this last point in the experimental literature. However, there is no consensus as to the neural mechanisms providing for the sustained activity observed in memory cells of the PFC nor of other cognitive aspects of WM. Some authors [10] have proposed that the cortico-cortical recurrent projections within the PFC or from posterior parietal cortex and inferior temporal cortex to PFC play the most important role in the building up of WM. It has also been proposed [14] that it is the thalamocortical projections from PFC to MD which play that role. However none of those possibilities can simply account for the WM deficits observed in PD and Huntington’s Chorea as these diseases are thought to be constrained to basal ganglia deficits [6, 7].

The basal ganglia is divided into two pathways respectively called the direct and the indirect pathway. Briefly, the internal segment of the globus pallidus (GPI) [15] is generally regarded as the output nuclei of the basal ganglia and it sends inhibitory projections to MD. The direct pathway arises from inhibitory striatal efferents to the GPI. The indirect pathway passes first through the external segment of the globus pallidus (GPe) which receives inhibitory striatal projections and sends inhibitory efferents to the subthalamic nucleus. The latter sends excitatory projections to the GPI. The two pathways, thus appear to have opposite effects, the direct one having a disinhibitory effect onto MD while the indirect one has an inhibitory one.

Many theories of the basal ganglia, based on electro-physiological studies [16, 17], argue that the disinhibitory effect of the direct pathway of the basal ganglia enhances the reverberating activity in PFC occurring from the excitatory reciprocal thalamo-cortical connections. It has also been proposed [18] that the PFC projects itself topographically onto the striatum. According to this theory, the circuits formed by the direct pathway of the basal ganglia and the thalamo-cortical projections serve to organise and select cortical features and sustain their activity. In this paper these circuits will be referred as BG-thalamo-cortical loops. It should also be noted that other theories of the basal ganglia have been proposed, one of the more popular being that a combination of topographic projections and of converging projections occur from the PFC onto the basal ganglia [19] with the topographic projections being more substantial.

Alexander et al. [20] suggested the existence of at least five parallel BG-thalamo-cortical loops which take their functionality from the portion of FL they connect to. They have proposed four distinct functional roles for these loops which they have nominated “oculomotor”, “motor”, “limbic” and “prefrontal”. According to their classification the cortical part of the
The oculomotor circuit comprises the supplementary eye fields and the frontal eye fields, the motor circuit includes the motor cortex, the supplementary areas and the premotor cortex, the limbic circuit includes orbitofrontal cortex and the anterior cingulate gyrus (ACG) each forming a loop of their own and the PFC circuit includes DLPFC and VL also forming a loop each. The totality of these basal ganglia circuits and the interaction between them seem required in cognitive processes such as access to contextual memory, selective attention and decision making for action.

1.2. Anatomical and physiological motivation of the model

In the following model only the direct pathway of the basal ganglia is considered. Each of these circuits [15] involves a FL area sending excitatory glutamatergic projections to its associated part of the caudate (CD). The latter sends inhibitory projections involving GABA and substance P to a corresponding area of the GPi which sends GABA inhibitory projections to MD hence resulting in disinhibition. Reciprocal excitatory projections occur between MD and FL involving glutamate. The reverberating function of the reduced BG-thalamocortical circuit described above was proposed in [17] as a mechanism of WM for eye movement. Of importance also are the strong long-range inhibitory connections occurring within CD and GPi as a result of the principal spiny neurons [21], the inhibition occurring from inter-neurons within MD, the inhibition occurring from inter-neurons within PFC and the role of dopamine in FL and basal ganglia. It is the effect of each of these loops and the interaction between them which has been proposed in the ACTION network theory in [22] as an explanation for context-dependent memory-driven action and behaviour. The models presented here are an implementation of this theory.

The more specific hypothesis is that each of these loops codes for specific features to be activated as WM when required. The model only considers the ACG of the limbic circuit, the DLPFC and VL loops of the prefrontal circuit. We will take the somewhat classical view that the DLPFC is associated with spatial features while VL is associated with coding for object features [13]. This view is clearer in monkeys than in humans. Recent theories of the PFC (e.g. [2]) have argued that the DLPFC is more involved in the monitoring of WM while the orbitofrontal cortex is involved in aspects such as the reward/penalty values associated with features. It should be noted that, in our model, it is the ACG which also deals with those aspects and can be seen as representing the combined action of the ACG and the orbitofrontal cortex of the "limbic" circuit. It has been shown, on the basis of neuropsychological experiments [2], that VL is not an integral part of the limbic circuit. It is therefore considered as distinct from the orbitofrontal cortex.
In the model described below we consider the DLPFC loop, the VL loop and the ACG loop together with connections from DLPFC and VL to the primary motor areas (PMA). The PMA is taken as the final response of our network. We therefore do not consider the whole motor loop. We examine here the role of the three loops mentioned above with respect to the DRS and DRO as they are a good test-bed to probe the sustained activity of WM. We also examine these loops while performing the WCST as this is an attentional tasks which involves higher level processes, such as set shifting. PET studies [23] have shown the importance of the ACG with respect to selective and sustained attention. Furthermore, on the basis of PET studies Pardo et al. proposed [24] that "the anterior cingulate is involved in the selection process between competing processing alternatives on the basis of some pre-existing internal conscious plan". Our interpretation of this statement is that the ACG is effectively coding for strategies.

The Nucleus Accumbens (Nacc) or ventral striatum has been shown experimentally to be involved in reward/penalty learning via the use of its strong concentration of dopamine [25]. PET studies of the excited and inhibited regions of a planning task [26] have shown the involvement of the basolateral amygdala (bLAMG), the Nacc and the orbitofrontal cortex in those processes. There has been further evidence for it using functional magnetic resonance imaging (fMRI) studies on the evaluation of the emotional valence of visual stimuli [27]. The amygdala is usually associated with fear detection, firing when a negative event has occurred, while it has been suggested [28] that error detection occurs in the hippocampal formation. In our model the error detection comes from an external signal which causes the bLAMG to fire. We suggest that the reason the Nacc and the orbitofrontal cortex are activated in the studies mentioned above is that they must be involved in the change of strategy following a negative feedback. This is implemented in the following manner in our simulations. The ACG, which can also account for some of the orbitofrontal cortex functions in our model, codes for possible strategies. Upon the detection of an error, the bLAMG sends a penalty signal by firing so as to change the pattern of activity in Nacc which will result in a change of strategy choice occurring in ACG via the effect of the whole ACG loop. The bLAMG acts as a switch onto the Nacc.

Eraser cells, which are linked to turning off a memory cell, have been observed in the PFC. There are three possible hypotheses regarding their mode of action. They could be exciting inhibitory inter-neurons in MD so as to have a global inhibitory effect on MD [29]. They could be sending lateral inhibitory connections to the memory PFC cells or could be exciting the principle spiny neurons of the CD thereby resulting in striatal inhibition (Tanji, J. personal communications). We have chosen to implement the first hypothesis via the inhibition of MD. It should be noted however that our simulation can be made to work under any of these three hypotheses.
1.3. Information processing functionality of the model

The model described below shows how artificial neurons coding for features can exhibit reverberating activity to be used as context. This is achieved via the effect of loops, coding for categories of features such as location and object features, which contain positive recurrent feedback as well as a disinhibitory pathway to achieve the sustained activity of the relevant feature. A higher level loop, namely the ACG loop, allows for the selection of the relevant feature loops and the shifting from one another. The ACG itself is a winner-take-all-layer which act as a control in our system. The decision-making process of the system occurs from an external penalty signal implemented by bLAMG which results in a switch of activity (or strategy) in the ACG. There are many computational ways of achieving this functionality; it is our aim to investigate how the human brain performs it by constraining the architecture of the model to some known anatomical connectivity.

The crucial aspect of the ACTION network is that it allows for the creation of moveable basins of attraction of recurrent neural networks. It is usual in a recurrent net, such as the standard Hopfield net [30], to create only static basins of attractors. This is done by means of suitable feedback from the recurrence; once an attractor has been created, say by Hebbian learning of a pattern, then the activity of the selected neurons is persistent.

Transitions between different attractors in a recurrent net can be achieved by learning the overlap between the corresponding neural activity patterns [31]. The energy surface is time-varying, with the attractors varying from one to the next as time develops. The resulting sequence of attractors is, however, relatively immune to external activity and inflexible in its use for guiding behaviour. The idea behind the ACTION network is to extract from the neuroanatomy and neurophysiology of the FLs the basic principles behind the development and use of WM as based on moveable attractor basins. These basins can be directly influenced from other cortical regions: they have been constructed to allow a specific path to be taken through the space of attractors of a given neural network architecture. The set of ACTION networks we have delineated above, based on the Alexander et al. BG-thalamocortical loops, are composed of several interacting recurrent nets in which movement is flexibly achieved in the multi-dimensional space of numerous attractors.

The flexibility in attractor space is achieved by a given ACTION network (or loop) in the presence of the cortical-basal ganglia-MD-cortex pathway. Influences flowing onto the basal ganglia (in particular the striatum) from other cortical regions cause one of two things: the activation of a previously learnt attractor by posterior input or the change of one activated attractor to another by prefrontal input (including the ACG). The learning involved in the construction of such attractors and their transitions is described elsewhere [32]. Let us consider briefly here the transition process.
When a WM is active, with a given attractor basin active, the transition to another within the same loop is achieved by a new input being sent to the FL part of the loop which will result in the inhibition of the original attractor via the lateral inhibitory connections in the basal ganglia. In the models presented here the annihilation of the initial attractor is helped by an external inhibitory signal onto the MD. This results in the reduction of the overall activity of the loop, with however the initial attractor still having the strongest activity. Upon the presentation of a new input or cue, the new competing population reverberating via the loop will now have stronger activity and will suppress the activity of the first attractor. This leads to the activation of a second attractor within the loop.

There is also a decision-making aspect to the models. A higher order loop, namely the ACG, codes for strategies in the form of moveable attractor basins once again. Each node in the ACG inhibits a set of other loops, leaving the possibility of only one other loop being active at a time. The shift in strategy is achieved by the activation of a new attractor in the ACG loop which erases the previously activated one by lateral inhibition. The transition is achieved via an external signal onto the Nacc which switches the activity within this layer upon the occurrence of an error. Once again the transition is helped via an external signal onto MD which reduces the global activity of the ACG loop. The above mentioned processes are at the basis of our model of WM.

1.4. Contents of the paper

Section 2 describes the DRS version of the model. The tasks considered here are described in Section 2.1. The remainder of Section 2 is devoted to discussion of the architecture of the DRS network and to the account of its simulation.

Similarly, Section 3 deals with the WCST version of the model. Section 3.1 describes how the task is usually performed on humans. Section 3.2 explains the architecture of the WCST network and gives an account of its simulation.

In Section 4 the results of lesion studies of the networks are presented. These include predictions regarding the origins of the WM deficits observed in patients with PD and schizophrenia.

Section 5 provides a discussion where a comparison with other computational functional and network models is given. The relevance of the present framework to relating anatomical data to psychological studies and possible future improvements are also discussed.

2. The delayed response task model

The model was simulated in two forms, in its DRSs version and in its WCST version. The latter is a simple extension of the former. In this paper we present
the minimal versions of the models. We do not examine here the learning process but only the way the human brain could use various working memories to solve the tasks once these have been learned. We have tackled the learning process elsewhere [32] using distributed coding and reinforcement learning of the Hebbian type occurring between PFC and basal ganglia and from PFC to thalamus. In its current version the model can be considered as a decision-making one. Each unit in the network represents a population of brain cells with the same physiological behaviour.

2.1. The delayed response tasks

DRSs have been extensively used to examine FL deficits in human and monkey, see for example [2,33]. The basic procedure is the same for both the DRS and the DRO. A subject is cued for an object, followed by a variable delay period, usually between 2 and 60 s. The subject is then presented with two objects situated at different locations and must choose one of them. In DRS the rule is to choose the object situated at the same location as the one cued before the delay. This is the classical delayed response task.

In the DRO task two different objects are shown after the delay and the subject has to choose the object that was cued before the delay irrespective of its location. The DRO is a delayed matching task. There are two kinds of delayed matching tasks. One includes recurring stimuli where the same two objects are used at each trial. The object being cued and the location of the two objects are changed at each trial. The other kind includes trial unique stimuli, which means that a novel object is used alongside the cued object at each trial. The DRO here is a delayed matching task with recurring stimuli.

2.2. The delayed response tasks architecture

The architecture, the name of the grids and the connectivity of the delayed response tasks network are shown in Fig. 1. The input layer contains four units which represent four populations of cells lying in posterior visual cortex. We let A denote object A, B denote object B, L mean left location and R mean right location. The first unit in the input layer from the left in Fig. 1 codes for object A being at the left location (LA), and similarly the second unit codes for LB, the third for RA and the fourth RB. The grid marked DlPFC in Fig. 1 has two units, one coding for L the other coding for R. It is meant to represent two populations of cells within the DLPFC. Similarly the grid marked VL in Fig. 1 is meant to represent cells lying in the VL with one unit coding for A and the other coding for B. LA in the input grid sends excitatory connections to L in DLPFC and A in VL and similarly for LB, RA and RB. The value of the lateral inhibitory weights in dlCD and ldmGPI are set to be twice as large as those in DLPFC and pMD, since in the first two grids they represent the effect
of the principal spiny neurons in the basal ganglia, whereas in the case of the other two grids they represent the effect of inhibitory inter-neurons. The connections within each of the DLPFC, VL and ACG loops in the model are exactly the same.

As mentioned before, it has been postulated in [29] that resetting of WM's could occur from cortical eraser cells exciting inhibitory inter-neurons in MD. We have explored this possibility further: the grids SIGD, SIGV, SIGA represent three populations of cortical cells which would excite the inhibitory
inter neurons of the appropriate part of MD, so as to inhibit MD during the reset of WM. We have implemented this last phenomenon in the following manner: three units in the models (SIGD, SIGV and SIGA in Fig. 1) send inhibitory connections to both units in their corresponding MD. The O unit in ACG codes for "object strategy" and sends inhibitory projections to DLPFC. Similarly the S unit in ACG stands for spatial strategy and inhibits VL.

The combination of projections from ACG to DLPFC and VL is more complicated than just simple inhibition in the real biological system. It includes a combination of inhibition and excitation. However this simplification is used here as this is not central to the theory of our model. Simple inhibition is sufficient to deal with all the tasks that this paper addresses. It seems that for other tasks, learning is needed on those connections. This point is dealt with in more depth in the discussion.

The O unit in ACG sends excitatory connections to the O unit in Nacc and similarly the S unit in ACG feeds to its companion in Nacc. We do not suggest here that the Nacc codes for strategies, but that the competing populations in ACG coding for strategies send projections to competing populations of the Nacc. The units of the Nacc have been denoted O and S for ease of description. bLAMG in the model plays the role of a switch, one neuron sending a positive projection to the left unit of the Nacc and a negative one to the right while the other does the reverse. PMA in the model stands for primary motor area and represents population of cells in the supplementary motor area. MC which stands for motor cortex in the model represents a population responding to the fact that a movement is occurring and thus fires into PMA during a response period. The threshold of the neurons in PMA, set to three is made so high that any neuron in PMA will not fire strongly with no input from MC.

All the neurons in the model except for the units in input, SIGD, SIGV, SIGA, MC and bLAMG are Leaky Integrator Neurons (LIN). Leaky integrator models were originally used in [34] to model the transmission of action potentials along a squid axon. The behaviour of LIN's were more recently studied for a restricted class of architectures [30]. Their equations are the following

\[
\tau \frac{dv_i}{dt} = -v_i + \sum_j w_{ij} u_j \quad \text{with} \quad u_j = f(v_j - \theta)
\]

and

\[
f(x) = 1 / (1 + e^{-\beta x})
\]

where \(v_i\) is the potential of neuron \(i\), \(u_i\) its output, \(\theta\) its threshold, \(u_j\) the output of each unit \(j\) which feeds to unit \(i\) and \(w_{ij}\) are the weights connecting every neuron \(j\) to unit \(i\). \(\tau\) is the time constant which is set to 0.11 seconds for every LIN in the network. \(\theta\), the threshold, is set to 1 for every LIN in the network except for the ones in PMA which are set to three. The remaining units in the
grids mentioned above are input type neurons with their activity set to one when required, otherwise set to zero. All the LIN’s are updated synchronously every 10 milliseconds. DRS and DRO are simulated consecutively.

2.3. The delayed response tasks simulations

Before the simulation starts, the output of the S unit in Nacc is set to 1, the other set to 0. After a few updates this causes the S neuron in ACG to have a strong output and therefore inhibits VL. The cue period lasts for 0.75 s, with 1 neuron in the input layer set to 1. This is followed by a delay period where all the input neurons are set to 0 which lasts 2.60 s. During the delay period sustained activity is observed in DLPFC for the cued location (L or R). This sustained activity occurs from the effect of the DLPFC loop. The part of the loop which passes through the basal ganglia results in disinhibition and there are excitatory reciprocal projection between DLPFC and pcMD, all of this providing for the sustained activity. The basal ganglia also allows for the organisation of features in WM in our model as will be shown in Section 4. The potential of the unit L in DLPFC is shown for a successful trial of DRS in Fig. 2. The inter-trial period is not shown in full. A slight decrease of potential can be observed during the delay. It should be noted however that the resulting

![Successful trial graph]

Fig. 2. The potential of unit L in DLPFC during a trial of DRS with L cued.
potential is much greater than the threshold which is set to 1. The L unit in DLPFC not only represents context cells in DLPFC but all cells that are selective to the left direction in DLPFC. While the context cells reported in [9] fire strongly during the delay, the fixation cells and the visual tonic cells fire more strongly during the cue period and the response period. During the response period the network is presented with two units switched on in the input layer for a duration of 2.5 s. The output in PMA at the end of this period is taken as the final output upon which performance is judged. There is an intertrial period of 35 s during which SIGD, SIGV and MC are set to 1 and the WM's are reset. At this point the network classifies according to location correctly and solves the DRS. When the first trial of DRO occurs, the network still classifies according to location and an error occurs. This causes the neuron bIAMG which inhibits S and excites O in Nac and SIGA to be switched on during the inter-trial period which results in O being activated in ACG. Thus, classification is sought by object from the second trial onwards. In this case, sustained activity is observed for the cued object (A or B) in VL. The DRO can also be solved when the objects are reversed during the delay. If this is followed by further trials of DRS then an error occurs as the network classifies according to object. This causes the other neuron in bIAMG to fire during the inter-trial period switching the activity in Nac to S again.

3. The WCST model

3.1. Description of the task

The WCST (e.g. [35]) is a task which has been used to investigate frontal related deficits in humans. Typically a subject is shown four cards each inscribed with a set of objects. The cards vary in the number of objects they contain, their colours and their shapes. At each trial the subject is given a card to match with one of the four reference cards without being told what the rule of classification is. The experimenter chooses a feature (shape, colour or number) with respect to which the matching is to be done only telling the subject whether the classification is right or wrong after each trial. After a fixed number of correct classifications the experimenter changes the rule without telling the subject and the same procedure continues.

In the network described in the next section only three reference cards are used. This was done for ease of simulation. The only main difference with the standard paradigm is that this paradigm does not allow for a non-matching-classification. A non-matching-classification means that the subject thinks that the criteria for classification is that the presented card should not match any of the features within a single reference card.
3.2. The WCST network architecture

The architecture and the principles of the WCST network are very similar to the delayed response network. The architecture of the WCST network is shown in Fig. 3. The network now contains one prefrontal loop for each attribute together with the ACG loop. The same areas as in the delayed response version are represented, with the subscript C indicating that an area is part of the colour loop, similarly for the subscript S with respect to shape and N with respect for number. They are arguments for FLC, FLS and FLN being part of both DLPFC and VL. Since they code for object features one could argue that...
they are situated in VL. However PET studies [35] show DLPFC as having the strongest activation while performing WCST. It should also be noted that some authors include VL as part of DLPFC. This is why we just expose here FLC, FLS and FLN as being part of the PFC. The connectivity within each loop are the same as for the delayed response network.

There are now three input grids, in which inputC codes for green, red and blue, inputS square, circle and triangle and inputN for 1, 2 and 3. They are 1 to 1 excitatory projections from inputC to FLC, from inputS to FLS and from inputN to FLN. PMA has three units which represent the reference cards. The first one represent a card with 1 blue cross, the second a card with two green squares and the third one with three red circles. Each node in FLC sends an excitatory connection to the node in PMA it matches with respect to colour, for example the node coding for red sends excitatory projections to the third node of PMA which represents the card with three red circles. Similarly each node in FLS sends excitatory projections to the node in PMA it matches with respect to shape and each node in FLN sends excitatory projections to the node it matches with respect to number. The C node in ACG sends inhibitory connections to FLS and FLN, the S node in ACG sends inhibitory connections to FLC and FLN and the N node in ACG sends inhibitory projections to FLC and FLS. Each node in bIAMG sends excitatory projections to a single node in Nacc and inhibitory ones to the other 2. They are now three competing nodes C, S, N in Nacc representing three competing populations in Nacc. As before we do not suggest that these populations code for the strategies but that the strategies population of ACG send topographic projections to them. The neurons used in the WCST network were LIN’s with the same equation as the one given in Section 2.2 except for the units in the three input grids, the four SIG grids and MC which are input type neurons: their activity is set to 1 when required, otherwise set to 0. A card is presented to the network by turning on one unit in each of the three input grids.

3.3. The WCST simulations

Preceding the simulation, one neuron in Nacc is turned on. This allows for one strategy neuron in ACG to be activated via the effect of the loop. The presentation period is 2.5 s. It is the output of PMA at the end of these 2.5 s which is taken as the response of the network. The classification is judged upon the following criteria: a correct classification occurs if the neuron in PMA which matches the current input card with respect to the current strategy (colour, shape or number) fires strongly and the other 2 have very low activity. An inter-trial of 5 s occurs between each card presentation. Following a correct classification SIGC, SIGS, SIGN and MC output 1 so as to reset the WM’s of the features during the inter-trial period. Upon a wrong classification they also output 1 and so does SIGA and one neuron in bIAMG. This activity in
blAMG causes a new node to be activated in Nacc and thus a new strategy node C, S or N is activated in ACG via the effect of the whole ACG loop and a new strategy is chosen for classification. If following the next presentation another error occurred the node in blAMG that was last activated would have its output set to 1. As such, the task can be solved under any order of card presentation, with the possibility of a maximum of two errors in a row. The strategy for classification was changed after six correct trials.

4. Lesions studies

4.1. Human infants and prefrontal lesions

Human infants of 7.5–9 months have been reported to perform badly on the DRS. Furthermore they are known to make the classical AB error while performing the DRS for delays over 1 s [36]. This error consists of the following: when a first location (e.g. left) is cued the subject correctly reaches for the left well. If after the inter-trial period the subject is cued for right the subject will wrongly reach for the previously cued left well. This error has also been reported in monkeys with lesions to the DLPFC. The deficit is thought to originate in both cases from a lower than normal concentration of dopamine in the DLPFC [36].

Reducing the gain is a method that was used reviewed in [37] to model dopamine deregulation. When the gains in the DLPFC of the delayed response model are reduced to a low value then an AB type of error occurs with respect to DRS. For instance, when the network is cued for LA, after the delay period, the L neuron correctly fires with a firing rate above 0.9 while the other neurons in PMA would fire below 0.1. Following the inter-trial period, the network is now cued for RA and after the delay, the L neuron fires again above 0.9 and the other neurons fire below 0.1 in PMA instead of the R neuron firing above 0.9. During the delay preceding the error, sustained activity is not observed in the unit coding for the cued location. With the gains reduced in DLPFC, the inter-trial signal occurring from SIG D is not sufficient to reset the working memories properly in DLPFC: the previously cued unit L has too strong activity and wins the competition over the unit R even after R is cued. A similar perseverative type of error occurs with respect to the DRO when the gains in VL are reduced to a low value. The membrane potential of the unit R in DLPFC is shown in Fig. 4 during an error trial where R is cued, when the gains in DLPFC are reduced. It is shown following a successful trial where L was cued in with the gains in DLPFC reduced. It can be seen that the potential for R during the delay is very low in Fig. 4: this occurs because the activity for L is still strong even though the network is now being cued for R. Those results give further support to our model.
The deficits observed in the delayed response tasks in PD were simulated by significantly reducing the weights entering and exiting dlCD and vmCD. The performance on DRS when these weights are significantly reduced in the model is as follows. During the cue period, the correct location is encoded, but the activity in DLPFC during the delay, is low and equal for both locations. Under the condition of reduced weights we have specified, it is impossible for the cued feature to be stored as a WM. At the end of the response period the PMA (which is effectively the output grid) fires at an equal rate for both locations. Similarly, when performing DRO with the above mentioned weights...
significantly reduced, both A and B in VL fire equally during the delay, and the output in PMA at the end of the response period is equal for both A and B. Only a chance level of choice between the responses to A and B is therefore possible. Fig. 5 shows the potential for neuron L and R in DLPFC during a DRS trial when L has been cued and the weights exiting and entering dlCD and vmCD are significantly reduced. The inter-trial period is not shown in full.

Similarly deficits observed in PD while solving for the WCST were simulated by significantly reducing the weights entering and exiting CDC, CDS and CDN. The performance of the WCST under those conditions is as follows. While the feature loop corresponding to the current classification criteria is correctly selected, the distinction between competing features in WM is not achieved properly. A reduction in the cortical activity is observed in the selected loop. Furthermore, the feature appearing in the card currently presented within the selected criteria is responded to as strongly as the competing ones within the same loop. As an example, suppose a card containing three blue squares is shown and the current classification criteria is number. While the

![Fig. 5: The potential of units L and R in DLPFC during a DRS trial with reduced weights. The weights entering and exiting dlCD and vmCD are significantly reduced. The potential of L is represented by a + line and the R potential by a - line. L is being cued in this trial and it can be seen that during the cue period L is properly distinguished while during the delay and response periods the potentials for both L and R are the same.](image)
card is being presented to the network the number loop is correctly selected. However the nodes representing one and three fire equally strongly as the one representing three. This effectively causes complete loss of effectiveness in the matching process and errors occur. We note that our degradations must be considered here, and in other simulations, as the extreme limit of what could be occurring in the patient population.

Our model would suggest that the degradation of striatal neurons observed in PD causes disruption in the WM of features within a loop. This would assign an organising role to the basal ganglia with respect to WM.

4.3. Working memory deficits in schizophrenia

Schizophrenic patients are known to perform poorly on the WCST. In normal subjects regional cerebral blood flow (rCBF) studies have shown increased activation in DLPFC while performing WCST [39]. This result has been more recently confirmed by PET studies [35]. rCBF studies reported in [39] have also shown a level of activity in DLPFC below baseline metabolic levels in schizophrenic patients which was associated with a reduced dopaminergic tone in the DLPFC of these patients.

There are various theory regarding the neural origins of schizophrenia. We will consider here some of the most popular ones. The first one [40] proposes that the symptoms observed in schizophrenia arise from problems in the connections from the limbic system to the basal ganglia. Dopamine deregulation in the Nacc is proposed as one of the main causes for the deficits observed in schizophrenia. Another one [41] suggests that the abnormal connections in schizophrenia lie between the ACG and the PFC. Weinberger [42] sees the primal neural basis of schizophrenia as lying in an underactive mesocortical dopaminergic innervation of the PFC from subcortical dopaminergic deregulation in the mesolimbic system. Our simulations described below would suggest that these three theories are not necessarily contradictory.

In order to investigate the mechanisms responsible for attentional and cognitive deficits in Schizophrenics we reduced the gains of all the units of the Nacc grid in the WCST model. Reducing the gain is a method that was used in [37] to model dopamine deregulation, as we have noted earlier. We considered grids FLC, FLS and FLN to represent populations of cells in the PFC and measured the sum of the outputs of all their neuron while performing a trial of WCST. This total DLPFC activity was significantly reduced when the gains of Nacc were reduced.

Fig. 6 shows the high activity of the DLPFC neurons in the WCST model when all the gains are set to 1 while it shows the very low activity exhibited by these neurons when the gains of the Nacc grid are reduced to a value of 0.3. This results is in accordance with the rCBF findings of [39]. When the gains of
the neurons in Nacc are reduced to a value equal or below 0.3 in the network while performing WCST the network starts to respond equally to any strategy at a given trial, thereby not being able to decide upon a specific response. Also all the nodes in all DLPFC respond equally and at a low level of activity. This causes more errors to occur and makes the blAMG fire more frequently than necessary causing even more disruption in Nacc making it impossible for the ACG to choose a strategy. Our model predicts that the lack of latent inhibition observed in patients with schizophrenia [28] arises from the fact that all stimuli are stored equally poorly in WM and thus a patient could not make use of the fact that a feature has been rewarded or penalised nor could it distinguish between a conditioned and an unconditioned stimulus. It should also be noted that the same pattern of behaviour occurs if the gains of the neurons of both the Nacc and the ACG grids are reduced.

In conclusion to this section, it can be seen that the model predicts that WM deficits observed in PD arise from the problem of encoding a feature as a WM within one loop while those observed in schizophrenia occur from a poor distinction and selection between the different loops.

![Prefrontal activity during WCST](image.png)

**Fig. 6**: Total prefrontal cortex activity with normal and reduced nacc gains. The added activity of grids 11C, FLS and HIN is shown during a trial of the WCST. It is shown when the gains in Nacc are set to 1 (represented by a - line) and when they are reduced to a value of 0.3 (represented by a * line). The inter-trial period is not shown.
5. Discussion and comparison to other models

The model we have exposed models WM as well as some form of decision making. Various other models [37, 43-46] have been proposed to model the mnemonic and executive functions of FL tasks. These were however developed from a functional point of view without relating the architecture to any of the anatomical structure of the frontal system. These models often satisfactorily explain the psychological data and the cognitive processes involved in the frontal tasks, but can not test many hypotheses regarding anatomical mechanisms. Other models have appeared [47-50] which indeed model the same or related anatomical areas as the ones represented in our model. These models propose mechanisms for the sustained activity aspect of our model, or for the generation of motor sequences, but not for the decision-making processes that have been dealt with here. The model in [47], includes time constants which ranges from 0.01 to 2.1 in their neurons. We do not think this is biologically realistic. It is unlikely that the sustained activity observed during the delays arises from large differences in the time constants of the cells membrane; we propose here that it arises from suitable anatomical connectivity (especially through BG-thalamo-cortical loops). Furthermore no lesion studies were done on the models of [47] and [48]. By constraining the model to some anatomical and physiological constraints, we were able to apply different kinds of lesions which allows us to investigate the deficits of more than one kind of patient which is not possible in the purely cognitive models. Our model is a first step towards relating cognitive data to anatomical and physiological data. It seems appropriate to complement brain imaging studies in this manner as the model allows for predictions which can be tested under PET or fMRI [51].

A very important improvement to our model would be the inclusion of further components of the basal ganglia and of more detailed dopamine pathways. This would have to take into account the indirect pathway of the basal ganglia and the effect of various types of dopamine receptors. This has been attempted in a model of one BG-thalamocortical loop [50] which investigates the effect of pallidotomy in PD. It does, however, not deal with WM directly.

We have recently included reinforcement learning of the hebbian type in a distributed coding version of our model. It allows for the learning of the features coded in WM and is discussed elsewhere [32]. Our model does not solve for delayed alternation. In this task, the reward is placed at the opposite location after each trial. It requires the model to use its current output as a cue for the following trial and reverse its response given as the previous one. This would require feedback from PMA and possibly higher order units in ACG. Another possible extension of our model is the inclusion of selection from the ACG of features of more than one loop at the same time. This would allow for even more complex tasks to be solved. This would require learning between
ACG and the other cortical structures and possibly a more complex blAMG-NAcc structure.

References


