To appear in Biosystems, 2001

Modelling Relative Recency Discrimination Tasks using a Stochastic Working Memory Model.

Guido Bugmann and Raju S. Bapi*

Centre for Neural and Adaptive Systems, University of Plymouth,
Plymouth, PL4 8AA, United Kingdom
http://www.tech.plym.ac.uk/soc/staff/GuidBugm/BUGMANN.HTML

* Dept of Computer & Information Science, University of Hyderabad
Gachibowli, Hyderabad - 500 046, Andhra Pradesh, India
bapics@uohyd.ernet.in

Abstract

Relative recency discrimination task is typically used to assess the temporal organization function of the prefrontal cortex (PFC). Subjects look at a series of cards (with words or drawings on them) and on seeing a test card determine which of the two items was seen more recently. Results show that patients with damage to the prefrontal cortex are severely impaired on this task. We propose a memory trace-priming mechanism, based on automatic time-marking process hypothesis (Schacter, 1987), to offer a computational account of the results. In this model, successive words seen by subjects leave decaying memory traces in PFC, which subsequently prime the representations in higher sensory areas such as inferior temporal Cortex (IT) during discrimination judgements. The paper focuses on the evaluation of a probabilistic pre-frontal trace mechanism using a pool of clusters of neurons with self-sustained firing that ends at a random time. The results show that the probabilistic behavior of subjects can be accounted for by the stochasticity of the trace model. A good fit to experimental data is obtained with a PFC memory persistence probability with a decay time constant of approximately τ=30 seconds. The model allows for a distributed representation in IT and PFC, but the best fit suggests a sparse representation. It is concluded that further data are needed on representations in IT and PFC, on the connectivity between these two areas, and on the statistical and dynamic properties of memory neurons in PFC

Keywords: Working memory, short-term memory, recency discrimination, prefrontal cortex, temporal trace, retention time.

1. Introduction

Prefrontal cortex is thought to participate in temporal organization of behavior (Fuster, 1989). Accordingly, many tests designed to assess the integrity of prefrontal cortex measure performance on temporal order discrimination. For instance, a family of tasks involves recency discrimination wherein subjects have to determine which of two test objects was seen later in a sequential presentation. Many of the deficits observed were attributed to the fact that prefrontal damage interferes with the ability to structure and segregate events in memory. In the absence of contextual cues, prefrontal patients are seen to have difficulties assigning temporal salience to items that appear successively (see Milner, Corsi and Leonard, 1991 for a summary). In the literature, two major hypotheses have been proposed for this basic deficit of assigning time-stamp to events in working memory (WM): (i) there is an *effortless* automatic time-marking process taking place at the time of encoding of events in WM which is disrupted due to damage to the prefrontal cortex and (ii) there are *effortful* encoding and retrieval processes that use active attention, rehearsal, and organization strategies that are disrupted in recency judgement tasks and that some of these processes depend on intact prefrontal cortex (see summary in Schwartz et al., 1991). A

clean resolution of the above hypotheses has not yet been made in an experimental paradigm. In the work reported here, we show that the effortless automatic time marking hypothesis can explain the results of two of the recency judgement experiments, using word items and representational drawings.

The main hypothesis of the model – that needs to be tested experimentally - is that working memory units exhibit a stochastic retention pattern. It will be shown that simple hypotheses on the dynamics and connectivity of clusters of prefrontal neurons can explain behavioural data of an apparently complex task such as relative recency discrimination. This result gives support to the hypothesis of *effortless* automatic time-marking process (Schacter, 1987).

This paper is organised as follows. Relative recency discrimination experiments and their results are described in section 2. Modelling behavioural data requires a model of how behaviour is generated. Our model is an extension of the model for top-down selective attention described by (Usher and Niebur, 1996). Details on this extended model are given in section 3. The parameters of the model that produce the best fit to the experimental data are given and discussed in section 4. The conclusion follows in section 5.

2. Recency Discrimination (RD) Experiments

In recency discrimination (Milner et al, 1991), subject is given a pack of cards each of which contains two items and is allowed to inspect each card for a few seconds and told to turn to the next one. Whenever a test card bearing a question mark appears in addition to the two items, subject has to point to the item seen more recently. Usually both items have been shown before (say, one of them 8 items ago versus another one 32). In some cases one of the objects has never been shown, hence the task becomes one of recognition memory. Three such tests that use items from one of the three modalities (involving concrete words, representational drawings, or abstract drawings) are usually administered.

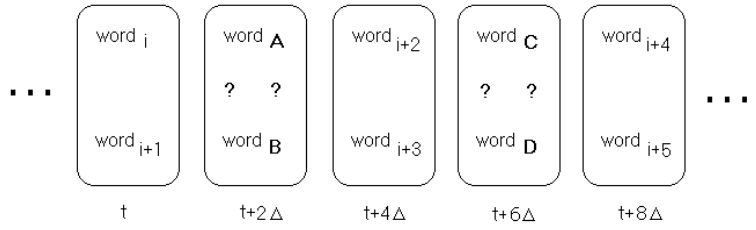

Figure 1. Example of card sequence manipulated by subjects. Δ is the average time separating the inspection of successive items.

In Milner et al (1991), RD tests were administered on 137 subjects --117 patients with unilateral cortical removals (of either the anterior temporal or the frontal lobe) and 20 normal control subjects. RD experiments were conducted in the following order: i) verbal practice task, ii) concrete words, iii) representational drawings, and iv) abstract drawings. In tests ii) and iii), 184 cards are used out of which 95 are white inspection cards with items to be remembered and 89 are yellow response cards with items for discrimination task. 29 of the response cards are used for recognition judgement, that is an item would be presented on the response card for the first time on the test and the subject has to recognise that the item has never been shown before and hence it is the most recent item. In contrast, 60 of the response cards that contain two previously seen items are arranged in 5 recency levels -- 4, 8, 16, 32, and 128. A recency level of 4 implies that the item was seen 4 items ago before appearing on the response card as one of the test items. All ten possible groupings of recency levels such as -- 4 vs. 8, 4 vs. 16, 4 vs. 32, 4 vs. 128, etc are used (see table 1). The test is self-paced and subjects scanned 184 word cards in approximately 15 minutes, resulting in an average of 5 seconds for observing and manipulating each card, i.e. 2.5 seconds per word. With representational drawings the observation time was limited to three

seconds per card (two images). Allowing for some time for manipulation, this corresponds probably to 2 seconds per drawing. Only recency judgement task for words and representational drawings is being modeled in the work presented here. Recognition tasks do probably involve the hippocampus and not pre-frontal cortex (Milner et al (1991). Milner et al (1991) did not provide performance versus interval ratio data for abstract drawings. Among the 137 subjects, only those 33 with left prefrontal lesions showed a significantly diminished average performance (59-62% correct responses) for concrete word items. The average performance for all the other subjects ranged between 64% and 72%. For the test with representational drawings, only 13 subjects with right frontal lesions showed a significant impairment (59% average correct response) compared to performances between 63% to 71% for the other subjects. Results of mean percent correct responses over all the 137 subjects (including frontal lesions!) and over all the word recency levels corresponding to an interval ratio from Milner et al (1991) are given in table 1.

Ratio Group	Recency level pairs			Average Performance Words	Drawings
½	4-8	8-16	16-32	0.6128	0.6027
¼	4-16	8-32	32-128	0.6905	0.6702
1/8	4-32	16-128		0.7311	0.7209
1/16	8-128			0.7615	0.7716
1/32	4-128			0.8425	0.8966

Table 1: Pairs of item (words or drawings) composing a ratio group. The last two columns indicate the average performance on word recency and on drawing recency. These values have been measured on figure 11 in (Milner et al, 1991).

The results in table 1, show that small interval ratios (i.e., large difference between the presentation times of the two test items) favored correct discrimination. We explore here the hypothesis that the characteristic inverse relationship observed between performance and interval ratio could be due to an effect of memory decay on discrimination decisions rather than due to a more complex mechanism for directly encoding intervals, temporal order or sequences.

The proposed model has to meet three targets: (i) it needs to replicate the profile of average performance versus interval ratios (table 1), (ii) it needs to reproduce the average normal performance of unimpaired subjects (words --70%, drawings -- 68%), and (iii) by appropriately damaging the memory process, the model needs to show impaired performance as seen in patients (words -- 61%, drawings -- 59%). Lesion modeling will be discussed elsewhere.

3. Model

3.1 Overview

Response movement generation: The computational model of the temporal trace is imbedded in a model of voluntary movement generation that realizes the task of the subject (figure 2A). For simplicity it is assumed that the firing of object selective cells in IT is mirrored by cells in posterior parietal cortex (PPC) coding for response actions directed towards the *position* of their respective objects (figure 2B). The biological circuit that realizes the transfer of activity from object selective IT cells to position-specific PPC cells is unknown. We will assume that object-specific neurons in IT with the strongest firing do eventually initiate a response towards the location of the corresponding object.

Trace in prefrontal cortex: There are neither physiological nor anatomical data specifying the neural circuits underlying the sustained activity in PFC nor systematic investigations of the time course of

the sustained activity at the neuronal level. It is known that (i) most PFC neurons show a constant firing rate during the delay period (Miller et al., 1996) and (ii) sustained firing does sometimes end prematurely (Funahashi et al., 1989). Beyond that, a number of assumptions need to be made, first in terms of modelling approach. In the standard "ensemble averaging" approach, a decaying memory trace would be modelled by a smoothly decaying function of the time. A problem with this approach is the need to convert the analogue trace activity into a response selection probability in a biological plausible way. Therefore, in our approach, rather than attempting to hide the unreliability of neurones by ensemble averaging, the unreliability is exploited to explain behavioural response errors. Sustained firing is assumed here to arise within a cluster of neurones from recurrent connections with probabilistic synapses. These introduce noise in terms of the duration of the sustained firing, but firing stays at a constant level until it ends (see details in section 3.2). The variable that decays with time is the *probablility* to find a neurone in an active state, not the activity itself.

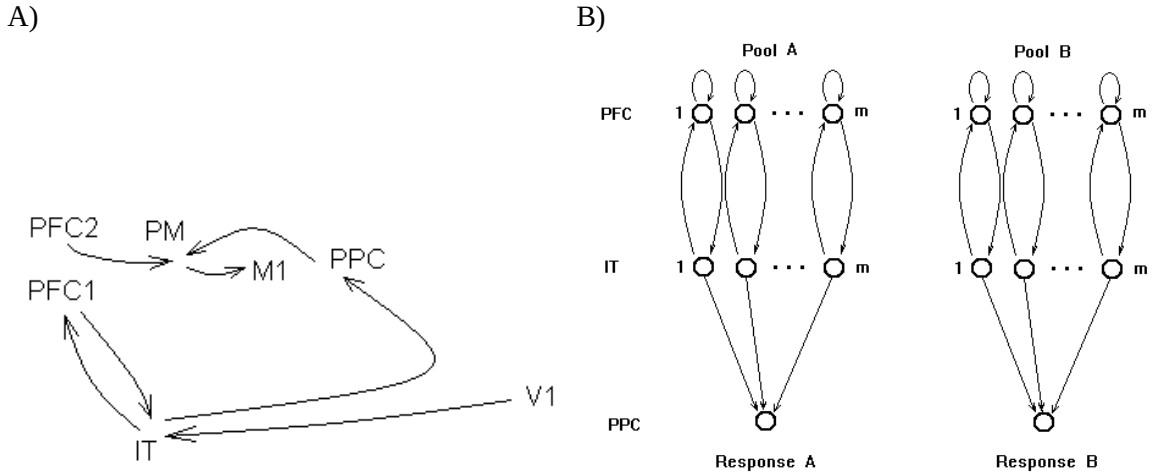

Figure 2: A). R*esponse generating circuit.* The link from PFC2 to PM carries a control signal, which enables responses during testing. The path from V1 to IT to PFC1 enables storing viewed patterns as temporary sustained firing. The path from PFC1 to IT to PPC via early common visual areas enables defining the position in space of the selected pattern, then selecting the appropriate response. The link from PPC to PM activates pre-motor neurones if enabled by a control signal from PFC2. Other control links are not shown. IT represents VA2 (Pandya & Barnes, 1987). PFC1 and PFC2 represent two prefrontal areas, possibly 46/9v and 10. PM represents supplementary and pre-motor areas such as FEF, SMA and PMd. M1 is the motor cortex (area 4). PPC represents a number cortical areas such as 5, 7, LIP known to contain action maps. **B)** *Model for priming by a pool of probabilistic all-or-none memory neurons*. Memory nodes in PFC are maintained in an active state by probabilistic self-feedback connections, then decay at random times. These nodes are initially activated by neurons in IT representing the observed items. At test time, active memory nodes reinforce the firing of IT neurons. The PFC pool with the largest number of remaining active neurons determines the item selected for the response.

Representation: While the subject observes the two patterns on his card, focusing probably his attention on the one, then on the other[1], the fixated pattern is represented by a pattern of neural activity in areas along the ventral visual stream VA1 (18/19), VA2 (20/21), VA3 (20,21). These areas represent increasingly complex features of the visual stimulus. Theses areas project to frontal areas, 45, 46/9v and 10/11/12 respectively (Pandya & Barnes, 1987). While area 45 is essentially a pre-motor area, working-memory related sustained activation has been observed in areas 46/9v and 10/11/12. Neurones in area 10/11/12 neurones show sustained and phasic firing that is much more stimulus-specific than that in

[1] We assume serial acquisition. The article by (Milner et al., 1991) does not describe eye movements of the subjects.

46/9v, responding only to complex objects such as faces (Courtney et al., 1997). This parallels the increasing complexity of visual features represented along the ventral stream. Area 10 has also been shown to be involved in branching tasks where visual stimuli prompt the switching between subtasks (Koechlin et al., 1999). Lesion studies suggest a crucial role for area 46/9v in representing the temporal order of stimuli (Petrides, 1991; Milner et al., 1991). This area is fed by VA2 in which objects are represented by the activation of multiple disjoint groups of cells representing partial features of the object (Wang et al., 1998). We explore here the possibility that the memory of an object is represented by similar groups of WM cells in area 46\9v (fig 2B). The number of such groups will be one of the free parameters in the data fitting procedure (Section 4).

3.2 Numerical model.

The patterns to be memorized are assumed to be coded in IT by a number *m* of sensory cells. Each of these cells activates its dedicated cluster of memory cells in PFC (figure 2B). These clusters have a probabilistic all-or-node behaviour, i.e., fire at a constant rate for a random time, then become silent (figure 3B). Such a firing pattern can theoretically be produced by a cluster of neurones with self-sustained firing mediated by probabilistic self-feedback connections (Bugmann and Taylor, 1993). These connections can sustain firing at a high rate until an abrupt stop, which occurs when, by virtue of the laws of probability, the feedback signal fails to be transmitted. In practice, simulations of clusters of Leaky Integrate-and-Fire (LIF) neurones have shown that such systems are very parameter-critical and tend either to decay in less than a second, or stay active indefinitely (Bugmann, 1997). However, failure of PFC memory cells after several seconds of activity has been observed, hence the behaviour of the theoretical model may nevertheless be appropriate.

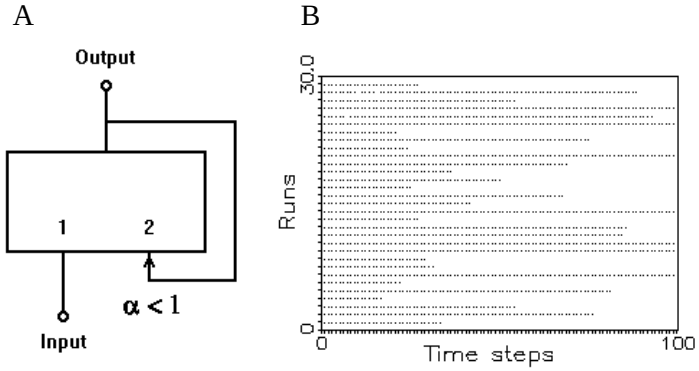

Figure 3: A) Abstract model of a cluster of working memory neurons with self-sustaining feedback connections. A single connection is assumed here with a probability α to reproduce the activity at each time step. **B)** Sample of spike trains produced by such a model with α = 0.98, for 30 different runs.

A working memory cluster in PFC is modelled here as an abstract unit that is activated by input from IT and stays active by virtue of a self-feedback connection (figure 3A). The connection has a probability α to sustain the activity. In the simplest case, such a unit represents a spiking neurone with an autapse. Then α represents the probability that one spike of that neurone generates another spike in the same neurone. Assuming that each feedback-and-fire cycles takes a time Δt, the probability to find the WM unit in active state after a time t = *k* Δt is given by (Bugmann & Taylor, 1993):

$$p_{on}(t) = \alpha^{t/\Delta t} = \exp(-t/\tau) \text{ where } \tau = -\Delta t / ln(\alpha) \qquad (2.1)$$

Then the probability $P_A(n,t)$ that exactly n clusters out of a pool A of m are still active after time t is:

$$P_A(n,t) = C_n^m \, exp^{-\frac{nt}{\tau}} (1 - exp^{-\frac{t}{\tau}})^{m-n} \qquad (2.2)$$

As for selecting the response, it is assumed that, if at time t the pool corresponding to pattern A contains more active memory neurons that the pool corresponding to pattern B, then priming will lead to a response pointing to pattern A. The probability of a correct response is essentially the probability that the most recent pattern has the most active pool. However, in the case where both pools have the same number of active neurons (usually both will have no active neuron left), it is assumed that noise will lead to each response to be chosen by chance with equal probability[2]. The probability $P(a>b)$ that pattern A has more active memories a than pattern B is:

$$P(a > b) = \sum_{n=0}^{m-1} \left[P_B(n,t) \cdot \sum_{i=n+1}^{m} P_A(i,t) \right] \qquad (2.3)$$

The probability that both pools have the same numbers a and b of active memories is:

$$P(a = b) = \sum_{n=0}^{m} P_B(n,t) \cdot P_A(n,t) \qquad (2.4)$$

The probability P_c of a correct response is thus given by:

$$P_c = P(a > b) + 0.5 \cdot P(a = b) \qquad (2.5)$$

In order to reproduce the experimental data in table 1, one needs to average the values of P_c over all possible interval pairs forming a given ratio. For instance, $P_{1/2}$ is calculated by:

$$P_{1/2} = \frac{1}{3} P_c(4,8) + \frac{1}{3} P_c(8,16) + \frac{1}{3} P_c(16,32) \qquad (2.6)$$

Where $P_c(4,8)$ indicates that pattern A is 4 items old and pattern B is 8 items old. The equal weight (1/3) given to all cases reflects the assumption that an equal number of cases of each pair of intervals are tested in the experiments. This is not specified in the paper by (Milner et al., 1991). Until confirmation from the authors can be obtained, this must be taken as an assumption of the model. It has two free parameters, the decay rate τ of the memories and the number m of memories in the pool.

4. Results and Discussion

Recency data fitting: A least-squares error minimization procedure was used to find the best values of τ and m. For each interval ratio, the probability of a correct response is calculated using (*2.5*) and (2.6). For word data in table 1, an excellent fit was obtained with τ=30 seconds and m =1 (sum squared error = 0.00169) (figure 4). With representational drawings data, the best fit gave τ=27 seconds and m =1 (sum squared error = 0.0011)(figure 4). The average performances over all recency pairs were 68% for words

[2] As the priming signal is missing when both pools are inactive, longer integration times are expected in PPC. Such a model may therefore explain the relation between latency and performance observed in (Sawaguchi et al., 1994)

and 69% for drawings. Overall, the good quality of the fit gives credit to the "automatic time-marking" hypothesis for memory of temporal order.

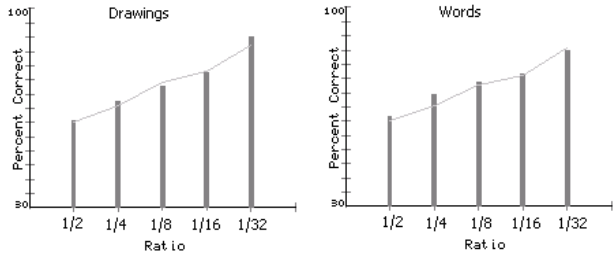

Figure 4. *Results:* The columns represent the experimental data from table 1. The lines joining the columns are best-fit values of the model.

Representation: A surprising aspect of the results is that the fitting process leads to a model of the task where a single node is used in IT and a single memory node is PFC (m=1). Recordings in these area always show many neurons involved in a given task and imaging data or EEG could not be produced if only a small number of neurons were involved[3]. It should be noted however that the optimal number of units obtained here depends on the decay function used. Initial tests indicate that a decay slower than exponential (with a longer tail), lead to better fits than the simple exponentially decaying function (2.1) and requires then 2-3 WM units. Hence, using this model for discussing the nature of the representation in IT and PFC will require more precise knowledge of the decay time course of biological WN units.

Limitations of the model: The model is designed for forced choice experiments such as the relative recency discrimination task modeled here or delayed matching to sample tasks (e.g. Fuster et al., 1981). In the latter case, WM decay time constants of the order of 50 seconds explain experimental data in monkey (unpublished results). The model is also specific to temporal memory tasks involving PFC. Other retention experiments fall in different categories. For instance, novelty detection tasks also show temporal decay of retention (Rubin et al., 1999) but do not involve PFC (Milner et al, 1991; Tendolkar & Rugg, 1996). An altogether different model seems required to explain the temporal decay in novelty detection tasks.

On the question of biological plausibility, the model is clearly oversimplified in many ways. Firstly, spike trains produced by WM neurons are treated here as a continuous signal that is either active or inactive. This corresponds to the assumption that the timing of individual PFC spikes is irrelevant for priming in IT. Considering individual spikes in the model would probably create more opportunities for errors and allow longer time constants to be produced by the fitting procedure. Not enough is known of biological time constants to determine if this would be an appropriate step. Other elements of the model may require improvements. For instance i) it is doubtful that the complex dynamics of a cluster of neurones can be captured by a simple probabilistic self-feedback loop; ii) further, it is likely that such clusters may exhibit a distribution of decay time constants; iii) possible effects of learning have also been omitted; iv) noise effects in other parts of the circuit have been mostly ignored. For instance noise must affect the Winner-take-All process in IT (Usher and Niebur, 1996) and affects the time to reach the firing threshold in PM neurons (Hanes & Schall, 1996). It is implicit in the present model that these sources of noise are critical only when two response have priming signals of similar strength, in which case random response selection results (section 3.2). The model can be extended to incorporate more of the biological details noted above, however experimental data are missing to give appropriate guidance.

[3] In optical imaging experiments by (Wang et al., 1998) visual stimulation had two effects in TE (border between VA2/VA3): A global activation of the whole area and additional local activation of small clusters of stimulus specific neurons. It may be possible that brain imaging experiments detect only the non-specific global activation.

Memorisation process: A hypothesis of the model presented here is the pre-existence of neurons (or WM-clusters) in PFC that are specific to given stimuli. This is unlikely, when considering the huge number of possible stimuli extractable from any natural sensory input pattern. We therefore propose that some preparation in done by the subject in order to set up units in PFC, with properties relevant for the task at hand. Further, we could also assume that only those stimuli relevant to the task actually initiate memory activity, thus pointing to the existence of some gating mechanism.

Given these hypotheses, the memory process modeled here is neither automatic in the restricted sense of (Zacks et al., 1984), nor *effortless* in the sense of (Schacter, 1987). However, it is passive in that no rehearsal or memory refreshment is needed to explain the performance of the subjects. This hypothesis offers a parsimonious account of human recency judgement performance.

5. Conclusion.

The proposed trace and priming model reproduces many features of the experimental data, such as average performances of unimpaired subjects, and the details of the performance grouped by interval ratios. In this model, recency discrimination errors made by subjects are entirely accounted for by the stochasticity of the retention time in WN. A surprising result is the suggestion that object representation in IT and PFC is not distributed but sparse, using a single cluster of neurons in each area. The model could suggest a larger number if a different, non-exponential, decay time course for WM units were assumed. It is therefore important to gather more detailed neuro-physiological data on the representation in IT and PFC, the connectivity between these two areas and the dynamics of memory neurons in PFC.

Acknowledgements. This work has benefited from discussions with Prof. Mike Denham, Prof. Dan Levine, Prof. John Taylor, Yakov Kazanovich, Chris Hindle and Ouri Monchi. RSB acknowledges support from the UK Engineering and Physical Sciences Research Council (GR/J42151) and the benefit of discussions with Dr. Kenji Doya. We are grateful for comments by anonymous referees.

References

Bugmann, G. & Taylor J.G. (1993) A stochastic short-term memory using a pRAM neuron and its potential applications. Proceedings of British Neural Network Society Meeting (BNNS'93). Can be downloaded from: http://www.tech.plym.ac.uk/soc/staff/GuidBugm/pub/stm1_ps.zip

Bugmann, G. (1997) Biologically Plausible Neural Computation. *Biosystems*, 40, pp.11-19.

Courtney, S. M., Ungerleider, L. G., Keil, K., and Haxby, J.V. (1997) "Transient and sustained activity in a distributed neural system for human working memory", *Nature*, 386, pp.608-611.

Funahashi S., Bruce C.J. & Goldman_Rakic P. (1989) "Mnemonic Coding of Visual Space in the Monkey's Dorsolateral Prefrontal Cortex", J. of Neurophysiology, 61:2, pp. 331-349.

Fuster J.M., Bauer R.H. and Jervey J.P. (1981) "Effects of cooling inferotemporal cortex on performance of visual memory tasks", Experimental Neurology, 71, pp. 398-409.

Fuster, J. M. (1989). *The prefrontal cortex* (2nd Ed.). New York: Raven.

Hanes D.P. and Schall J.D. (1996) Neural Control of voluntary movement initiation. *Science*, 274, pp. 427-430.

Koechlin E., Basso G., Pietrini P., Panzer S. & Grafman J. (1999) "The role of the anterior prefrontal cortex in human cognition", *Nature*, 399:6732, pp. 148-151.

Miller E.K., Erickson C.A. and Desimone R. (1996) "Neural mechanisms of visual working memory in prefrontal cortex of the macaque", *J. of Neuroscience*, 16:16, pp. 5154-5167.

Milner, B., Corsi, P. & Leonard, G. (1991) Frontal-Lobe contribution to recency judgements. *Neuropsychologia*, 29:6, pp.601-618.

Pandya, D. N., & Barnes, C. L. (1987). Architecture and connections of the frontal lobe. In E. Perecman (Ed.), The frontal lobes revisited (pp. 41-72). New York: IRBN Press.

Rubin D.C., Hinton S. and Wenzel A (1999) “The precise time course of retention”, *J. of Experimental Psychology*, 25, pp. 1161-1176.

Sawaguchi T. & Goldman-Rakic P.S. (1994) “The role of D1-Dopamine receptors in working memory: Local injections of dopamine antagonists into the prefrontal cortex of rhesus monkeys performing an occulomotor delayed task.” *J. Neurophysiology*, 71:2, pp. 515-528.

Schacter, D. L. (1987) "Memory, amnesia, and frontal lobe dysfunction: A critique and interpretation." *Psychobiology*, 15, pp.21-36.

Schwartz, B. L., Deutsch, L. H., Cohen, C., Warden D. & Deutsch, S. I. (1991) "Memory for Temporal Order in Schizophrenia." *Biological Psychiatry*, 29, pp.329-339.

Tendolkar I. & Rugg M.D. (1996) “Electrophysiological dissociation of recency and recognition Memory”, *Neuropsychologia*, 36:6, pp.477-490.

Usher M. & Niebur E. (1996) “Modelling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention”, *Journal of Cognitive Neuroscience*, 8:4, pp. 311-327.

Wang, G., Tanifuji, M. & Tanaka, K. (1998) “Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging”, *Neuroscience Research*, 32:1, pp.33-46

Zacks,R. T., Hasher, L. Alba, J. A., Sanft, H. & Rose, K. C. (1984) "Is temporal order encoded automatically?", *Memory & Cognition*, 12, pp.387-394.