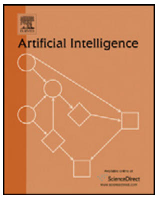
Artificial Intelligence 174 (2010) 205–214

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



# Cortical hierarchies, sleep, and the extraction of knowledge from memory

Bruce L. McNaughton

*Canadian Centre for Behavioural Neuroscience, Department of Neuroscience, University of Lethbridge, Lethbridge AB, Canada*

ARTICLE INFO

*Article history:*
Received 13 October 2009
Received in revised form 24 October 2009
Accepted 24 October 2009
Available online 17 November 2009

ABSTRACT

*Odin the Allfather had in his service two great ravens. These ravens' names were Hugin (Thought) and Munin (Memory) and every morning at dawn they would fly off over Midgard (the world) in search of news and information to learn more about humans and their activities. At sundown, they would return to Odin where they would perch one on each of Odin's shoulders, and whisper into his ears all that they had seen and heard.*

Experience, stored in the brain as memory, is the raw material for intelligence and thought. It has been suggested that *at sundown* (i.e., during sleep) the brain adjusts its own synaptic matrix to enable adaptive responses to future events by a process of gradient descent optimization, involving repeated reactivations of recent and older memories and gradual adjustment of the synaptic weights. Memory retrieval, thought, and the generation of adaptive behavioral responses involve globally coordinated trajectories through the neuronal state-space, mediated by appropriate synaptic linkages. Artificial neural networks designed to implement even the most rudimentary forms of memory and knowledge extraction and adaptive behavior incorporate massively and symmetrically interconnected nodes; yet, in the cerebral cortex, the probability of a synaptic connection between any two arbitrarily chosen cells is on the order of $10^{-6}$, i.e., so close to zero that a naive modeler might neglect this parameter altogether. The probability of a symmetric connection is even smaller ($10^{-12}$). How then, are thought and memory even possible? The solution appears to have been in the evolution of a modular, hierarchical cortical architecture, in which the modules are internally highly connected but only weakly interconnected with other modules. Appropriate inter-modular linkages are mediated indirectly via common linkages with higher level modules collectively known as association cortex. The hippocampal formation in the temporal lobe is the highest level of association cortex. It generates sequentially coupled patterns unique to the location and content of experience, but which do not contain the actual stored data. Rather, the patterns serve as pointers or 'links' to the data. Spontaneous reactivation of these linking patterns during sleep may enable the retrieval of recent sequences of experience stored in the lower levels of the cortex and the gradual extraction of knowledge from them. In this essay I explore these ideas, their implications, and the neuroscientific evidence for them.

**Contents**

---

*E-mail address:* bruce.mcnaughton@uleth.ca.

0004-3702/$ – see front matter
doi:10.1016/j.artint.2009.11.013

## 1. Introduction

The idea that the brain makes use of 'off-line' periods, such as sleep, to 'sort-out' and consolidate memories has a long history, probably as long as humans have wondered about the meaning of their dreams. Computationally principled suggestions for why such post-experience reprocessing should be necessary, however, have a much more modern history [1, 15,16], and were at least partly inspired by the phenomenon of temporally limited retrograde amnesia, the so-called Ribot gradient, following damage to the hippocampus and surrounding cortex. Such damage leaves humans cognitively normal except for a loss of recently acquired memories and a profound inability to form new ones. What follows is a review of the theoretical considerations for why the brain might require an active reprocessing of memories during periods when it is relatively 'disconnected' from external input, and a summary of current understanding of the phenomenon based on neurophysiological investigations in animals.

The fundamental necessity for off-line reprocessing can be understood from the most basic model for associative memory: a network of neurons containing a primary (e.g., sensory) input that determines the output pattern, and an association input pathway that is exhaustively (i.e., all-to-all) connected via Hebbian, associatively modifiable synapses. In the simplest scheme, the modifiable synapses have binary weights that are initially 0 and convert permanently to 1 following the principle of association outlined by Hebb [6]: correlated pre- and post-synaptic activity. Input vectors (also binary) are of a fixed length, and retrieval of the paired-associate of a given pattern (or pattern fragment) on the association input is accomplished by summing the net synaptic current to each cell (forming a dot product of the input with the synaptic matrix) and performing integer division by the number of active input axons. The latter operation, which captures the all-or-none nature of neuronal impulse communication, ensures that only those output neurons fire which contain a maximum proportion of already potentiated synapses in the current input pattern. Marr proposed that the division is accomplished by inhibitory interneurons which shunt the membrane resistance of the spike generating region, thus approximating a division according to Ohm's law. The physiological and anatomical properties of at least one known class of inhibitory interneurons are surprisingly consistent with this hypothesis [20]. Similar fundamental principles apply to the recurrent networks which can implement auto-association and pattern completion, as well as a simple form of sequence encoding.

The three primary factors that determine the storage capacity (in terms of the number of arbitrary patterns) of such networks are well understood [16,5,28]: connectivity density, coding sparsity, and 'orthogonality'. Optimizing for the latter two constraints involves redundancy reduction and the generation of efficient feature detectors. It is the first of these constraints that provides a basis for understanding both the modular and hierarchical organization of cortical association areas, and the fundamental necessity for a memory consolidation phase involving off-line reprocessing. To be clear about the basis of this constraint, however, it is necessary to make a theoretical distinction between two types of modifiable synapse: physical synapses and virtual synapses. Physical synapses can be identified at the ultrastructural level by a characteristic, morphological connection between two neurons, consisting of distinct pre- and post-synaptic elements. The weight ($\mathbf{w}_{ij}$) of a physical synapse from neuron **j** onto neuron **i** may be zero or non-zero (up to some maximum value). In cerebral cortex, the weights of synapses involved in storage of information are almost universally positive valued ("excitatory"); this is not the case everywhere in the brain. Non-zero physical synapses almost universally propagate information in one direction. A 'virtual synapse' has a weight of zero and no physically identifiable contact between pre- and post-synaptic elements. It can be converted to a physical synapse with non-zero weight with some probability that depends on complex factors, of which physical proximity and correlated activity of the pre- and post-synaptic elements are of main concern for a theory of association. Because the pre- and post-synaptic elements are themselves dynamic structures, exhibiting possible growth and retraction, the probability of conversion from virtual to physical is also dynamic: the elements have to be at the right place at the right time. Physical synapses can also be converted to virtual synapses, possibly as a result of anti-correlated activity between the two neurons, or by correlated activity in the wrong temporal order (i.e., post-synaptic activity *preceeding* pre-synaptic). The weights of physical synapses can be positively or negatively modified on a rapid time-scale (seconds to minutes), whereas the modification of a virtual synapse is stochastic. The pre- and post-synaptic elements need to be in close proximity at the time of a conjunction of activity of the two neurons. Hence, rearrangement of physical synaptic connectivity is slow on average (hours to days), and typically requires multiple conjunction events before one happens to occur at a time when the physical processes of the neurons involved happen to be appropriately aligned.