April 1984

Rcport No. STAN-CS-84-1004

A Computational Theory of Higher Brain Function

by

Lcslic M. Goldschlagcr

Department of Computer Science

Stanford University
Stanford, CA 94305

LELAND STANFORD JUNIOR UNIVERSITY
ORGANIZED 1891

A COMPUTATIONAL THEORY OF HIGHER BRAIN FUNCTION[†]

Leslie M. Goldschlager

Department of Computer Science
Stanford University, Stanford CA 94305

(currently on Sabbatical from
Basser Department of Computer Science
University of Sydney
N.S.W. 2006 Australia)

[†] This work is partially supported by funds from AFOSR 80-0212.

A Computational Theory of Higher Brain Function

Leslie M. Goldschlager

Computer Science Department
Stanford University
CA 94305 U.S.A.

(currently on Sabbatical from
Basser Department of Computer Science
University of Sydney
N.S.W. 2006 Australia)

April 1984

Abstract

The higher functions of the brain are believed to occur in the cortex. This region of the brain is modelled as a memory surface which performs both storage and computation. Concepts are modelled as patterns *of activity on the memory surface, and the model explains how these patterns interact with one another to give the computations which the brain performs. The method of interaction can explain the formation of abstract concepts, association of ideas and train of thought. It is shown that creativity, self, consciousness and free will are explainable within the same framework. A theory of sleep is presented which is consistent with the model.

1. Introduction

This paper brings together many of the facts and concepts from the literature on animal and human brains into a unified, coherent theory of brain function. Such a theory must necessarily be a gross oversimplification of reality, since a complete description of just one animal's brain, even if the total knowledge was obtainable by today's techniques, would fill many libraries. It is hoped that the vast number of details ignored by the theory are not central, and that the fundamental computational properties of the brain are adequately captured by the model presented here.

Of all known animals, the human appears to have the most interesting brain, partly because the human brain seems to exhibit the most complex properties, such as creativity, self-awareness and free will, and partly because we are human and are interested in understanding ourselves. However, the human brain, like every other animal, has unique structures, some left over from earlier evolutionary days and some specially constructed to deal with the particular details of the environment in which the brain finds itself. This environment includes the world external to the body (e.g. gravity), the world internal to the body but external to the nervous systems (e.g. the heart), and perhaps evolutionary older portions of the nervous system which need to be regulated and “tamed”. These unique structures will be ignored by the theory, which will instead concentrate on an abstract type of brain which may perhaps be considered to be a number of steps further up the evolutionary ladder than the human brain. In this manner, attention will be focussed on the “higher” aspects of human brain function such as those mentioned above.

The literature which has been drawn upon in this study comes from a spectrum of authors: philosophers, psychologists, neurobiologists, organic chemists, etc. In a sense, all of these authors can be thought of as scholars of the human brain, the main difference between them being the level at which they are interested in describing the brain. The discipline of computer science can be placed into the same spectrum, somewhere between the descriptive levels of the psychologist and the neurobiologist. It is this new descriptive level which allows a useful synthesis to be made between these diverse branches of study.

More specifically, the current paper attempts to extract the main concepts which philosophers have developed and found useful in describing the human brain (the most important being the concept of association) and to produce a computational model which embodies or can explain these concepts, taking into account the actual hardware and structures available for the task, as discovered by neurobiology (the most important being the operation of neurons and the anatomy of the cortex). The lowest, neuronal level of the model is discussed in section 2, and the associative level is covered in section 3. Creativity and the role of sleep are the topics of section 4, the self and consciousness are discussed in section 5, and free will is investigated in section 6.

2. *Neuronal* level

The computational model of the brain described herein is best viewed as a series of levels. Progressively higher levels deal with progressively more abstract concepts. (*Concepts* means things like “red”, “chair” and “number” .) At the lowest level, we will find the most fundamental concepts [Condillac 1754] which will be represented by the *simple patterns* introduced in this section. Associations between any two such simple patterns will represent the next level of concepts discussed in section 3, and higher and higher levels of abstraction of concepts can then be formed by further associations of concepts from lower levels of the hierarchy.

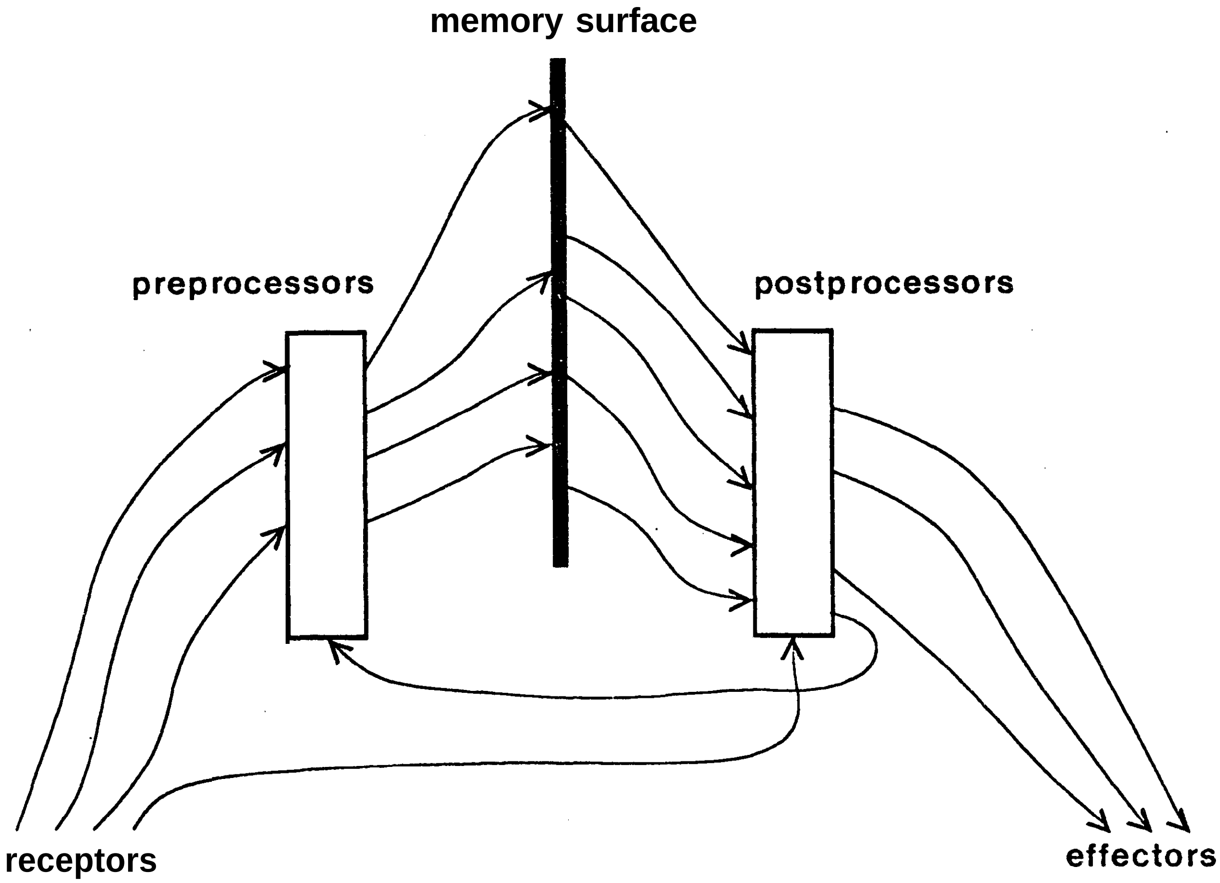

Figure 1. The brain as an input/output transducer.

In order to understand what are the most fundamental concepts which the brain can handle and how these are represented by simple patterns, it is helpful to see the purpose of the brain from a particular viewpoint. This viewpoint is the ***Input/Output transducer***, shown in figure 1. This diagram depicts the brain as a machine whose purpose is to interact with its environment. As mentioned in the introduction, the term environment is taken to relate to everything outside the brain, including the world outside the body and the organs within the body. Thus ***receptors*** include the senses (eyes, skin, etc.) and internal sensors (blood sugar level, muscle contracted, etc.). Similarly ***effectors*** cause interaction with the outside world (move tongue, raise arm etc.) and control of internal organs (release hormones, contract a blood vessel, etc.).

Signals arriving from the receptors are first *preprocessed* by special purpose brain components which are highly tuned to the needs of each particular animal. For example, in the case of a frog, some of the signals from the receptors may represent the geometrical pattern of light falling on the retina of the frog's eye, whereas some of the signals coming out of the preprocessing units may represent "small dark object moving rapidly across field of vision" i.e. a fly [Arbib 1972]. In general, the receptors produce raw signals representing the direct impact of the environment on particular sensors, whereas the signals leaving the preprocessing components represent higher level features of use to the animal. The more primitive the animal, the more specific will be the preprocessed signals and therefore the less flexible the animal will be in its response to a changing environment. By way of contrast in the case of humans, the preprocessed signals from the retina are probably no more specific than detecting features like stationary and moving edges in the field of vision. Therefore, for the purpose of the model presented here, preprocessing of sensory information will not be considered in detail. Instead, attention will be directed to the signals arriving on the "memory surface"[De Bono 1969], bearing in mind that these signals may have originated directly at sensors, or may represent "features" detected by the preprocessors.

Similarly, signals produced by the memory surface are usually channelled through components of the brain which are again specifically tuned to the needs of the individual animal. For example, a signal produced by the memory surface may represent the command "raise the right arm" whereas the postprocessed signals which are sent to the muscles will contain the detailed information needed to control the exact sequence of muscle contractions and relaxations to produce the desired effect. In general, he fast acting feedback loops needed to control complex muscle movements will be considered part of the postprocessing units. Again, the details of the animal-specific postprocessors will be ignored for the purposes of this model, aud the signals emanating from the memory surface will be considered as the outputs of major interest.

Unfortunately, in paying less attention to the pre- and postprocessing components of the brain, much of the very interesting and important information which has been gleaned by neurobiology will not be directly relevant to the model, even though it is highly relevant to understanding the pre- and postprocessors. As it happens, the pre- and postprocessing components are the parts of the brain most directly connected with the environment and therefore most amenable to study. However, some study has been directed at the higher centers of the brain and this information is critical to the model. Specifically, it is believed [Albus 1981] that the human cortex is responsible for the higher level processing in the brain; that it can be spread out into a flat sheet about two square meters in area, and that it consists of many millions of more or less identical columns. About 70% of the human brain is cortex, and this fraction decreases as one looks at progressively less "intelligent" animals (monkeys, dolphins, rats etc.). It is also believed that on the evolutionary scale, the size of the human cortex has been expanding at a dramatic rate. This rapid expansion tends to indicate that the structure of the cortex scales in a reasonable manner [Ullman 1984]. For example, scaling considerations rule out models in which each pair of neurons are directly connected because doubling the number of neurons would quadruple the number of connections, resulting in connection difficulties as the cortex expands.

Thus the higher levels of the human brain will be **modelled** as a ***memory surface***, which is a two-dimensional sheet comprised of millions of points, which will be called ***columns.*** Each column is connected to other columns nearby and no long distance connections are required, so that the structure scales reasonably (Figure 2). As well as communicating with nearby neighbors, columns may have connections to specific receptors (possibly via preprocessing components), and they may have connections to specific **effectors** (possibly via postprocessing components). Some columns may be connected to both receptors and **effectors,** some to none, and some to either one or the other.

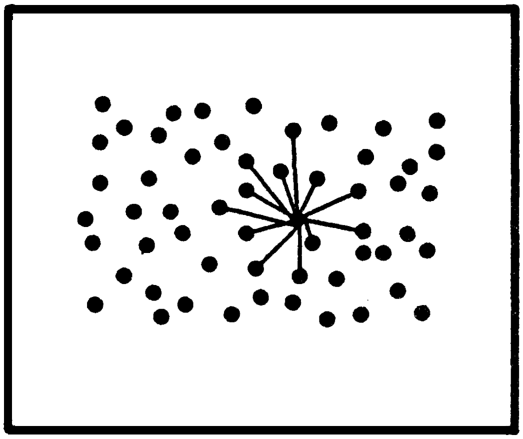

Figure 2: A memory surface consisting of many columns (shown as dots). Only the connections for one specific column have been shown.

Figure 3 illustrates the functional characteristics of a single column. For each direction d, the column has connections to and from nearby columns in that direction. To simplify the description, directions will always be measured relative to the column under consideration. These connections are schematically shown as a single arrow coming into the column and a single arrow leaving the column at direction d. The column may also have a connection from a receptor (shown as an arrow head coming into the center of the column from the third dimension) and a connection to an **effector** (shown as an arrow tail going out of the center of the column into the third dimension). The function of each single column can be described in terms of its communication characteristics and its memory characteristics.

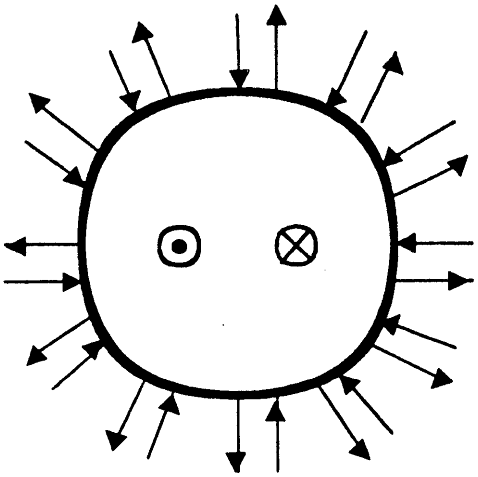

Figure 3: The functional characteristics of a single column.

All communication will be in the form of trains of pulses running in the direction of an arrow. This models trains of nerve impulses which travel along the axon of a neuron away from the cell body [Stevens 1979]. Such impulses are caused when the net voltage at the input end of the neuron (i.e. the "dendrites") exceeds a certain threshold value. After a fixed "minimum refractory period", the next impulse will be fired down the axon just as soon as the net voltage on the dendrites again reaches threshold. The nerve impulses always have the same shape and amplitude, but their frequency can vary with time. The frequency at one instant of time is called the *instantaneous frequency* of the pulse train and it represents the net voltage arriving at the dendrites over a short period of time immediately preceding that instant.

Now the communication characteristics of a single column can be specified. Firstly, whenever a pulse arrives at the column from direction d, a new pulse will be produced and sent out of the column in direction $d+$ 180 degrees. That is, each incoming pulse will behave as if it travelled in a straight line through the column and out the other side. Secondly, the number of pulses arriving at the column from all directions will be summed over a short time period and a pulse train will be produced whose instantaneous frequency f, called the instantaneous frequency of the column or the *activity* of the column, is proportional to the value of the sum. This sum will slowly decay over time so that incoming pulses do not have a lasting effect. If the column happens to have a connection from a receptor, the pulses arriving along that connection will also be added into the overall sum of arriving pulses, but with a higher weighting factor. If the column happens to have a connection to an effector, then the pulse train of instantaneous frequency f will be transmitted to that effector. Finally, for every direction d, there is a number m_d (explained in the next paragraph), so that a pulse train of instantaneous frequency fm_d (the product of f and m_d) will be transmitted in direction d.

It remains to describe the memory characteristics of each column. The term ***memory*** refers to any change in the column caused by the behavior of the memory surface which can, in turn, **affect** the behavior of the memory surface at a later time. Firstly, the mechanism which sums the incoming pulses to the column will exhibit short term ***habituation.*** That is, if the sum is large for a long time thus giving rise to a large value of f, then the mechanism will tire and will begin reducing the value of f. Conversely, after a period when the sum is low, the habituation or tiredness will slowly wear off. Secondly, for any direction d, whenever a pulse arrives from that direction, m_d will be incremented slightly, provided that f exceeds a certain value at the time the pulse arrives. The values m_d for all directions d represent the memory stored in that column. The increased values of the m_d's will persist for a long time, perhaps weeks, months or even (in the case of very large m_d's) years but they too will slowly decay over time. By contrast, the habituation may only last a fraction of a second.

The function of each column, as described above, can be implemented in numerous ways using the known properties of neurons. The exact way in which it is implemented in a particular animal cannot be determined without further neuroanatomical evidence, nor is it essential for what follows. Indeed, it is quite conceivable to implement the above structure using different building blocks altogether, such as transistors on a silicon chip. However, for the sake of plausibility, one possible implementation of a column using neurons is shown in figure 4.

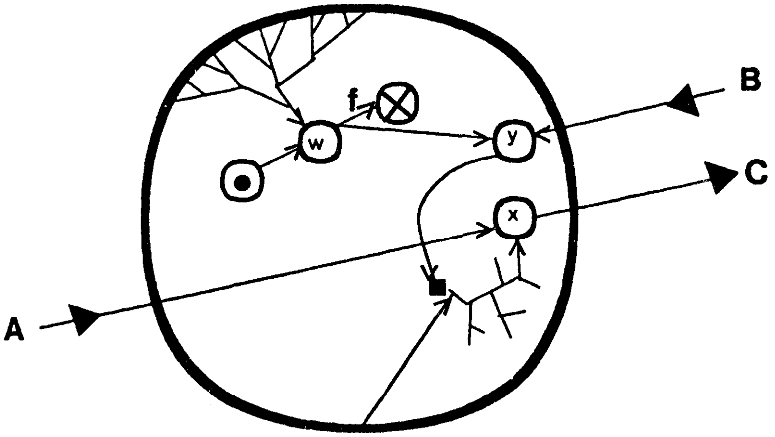

Figure** 4 - A **possible implementation of a column (showing only the neurons relating to one direction).

In this implementation, the neuron w sums all the incoming pulses, producing a pulse train with instantaneous frequency f. This neuron undergoes habituation, which is a form of memory described in [**Kandel 1979**]. The second form of memory is mediated by neuron y. Provided f is at a sufficiently large value, the axon of y will produce a pulse each time a pulse arrives from B. Each of these pulses further sensitizes the presynaptic terminals of x's dendrites [Kandel 1979]. The point at which sensitization takes place is shown as a solid square in figure 4. The degree of sensitization of these terminals is thus proportional to the number of pulses which arrive from B while f exceeds a certain value. Let m_d be the amount of sensitization. Now the dendrites of x are forming a sum similar to that formed by w, but weighted by m_d. The effect is that the instantaneous frequency of the pulse train output by x towards C is fm_d as required. The neuron x also relays pulse trains arriving from A through to C without any change.

With this structure of the memory surface as defined, it is now possible to answer the question posed at the beginning of this section, namely what are the most fundamental concepts which the brain can handle and how are they represented on the memory surface.

Define a *pattern* to be a set of columns (in an arbitrary geometric arrangement on the memory surface, and not necessarily adjacent), together with the relative activities of the columns in the set. More precisely, a pattern P can be represented by a list of non-negative numbers, one for each column on the memory surface. Many of these numbers will be zero, representing the fact that the corresponding column does not participate in this particular pattern. The columns corresponding to non-zero numbers do form part of the pattern, and the value of the number gives the activity (i.e., instantaneous frequency) of the column relative to the activities of the other columns. The convention will be adopted that every pattern P is normalized i.e., the sum of all the numbers in P is exactly one. This emphasizes the fact that it is the relative values of the activities of the columns which define a pattern, rather than the absolute values. The pattern P can be imagined as a vector or as a matrix, but it is perhaps most helpful to imagine P as a picture of the memory surface with a number for each column giving that column's activity relative to the other columns. If s is any non-negative number, and if each number in P is multiplied by s (the result being represented by sP), then it is possible that at some instant of time, the columns of the memory surface are active in exact correspondence to sP. That is, each column in the memory surface can have an instantaneous frequency given by its corresponding number in sP. In such a case, it may be said that the pattern P is currently *active with strength* s on the memory surface. In practice, patterns will usually consist of many thousands of active columns. Each pattern will represent a concept which the brain can handle. These concepts may be less abstract such as "dog", or more abstract such as "ownership".

The simplest patterns (which correspond to the least abstract or most fundamental concepts which the brain can handle) are just those patterns which result from some combination of receptors firing in response to some actual events occurring in the environment of the brain. Symmetrically, those patterns which cause the effectors to produce some useful impact on the environment arc also among the simplest patterns.

The above paragraphs have described a memory surface as a collection of columns interconnected in a simple manner. A detailed description of the behavior of each column was also given and the idea of pattern was defined. What are now the computational characteristics of the whole memory surface as described? One observation can be made immediately. At any time various columns will be active to various extents (i.e., they will have various values of f, their instantaneous frequencies). So over any period of time, each column will experience some average amount of stimulation in the form of pulses arriving at the column. The column will therefore tend to habituate to this average stimulation which may be thought of as *background noise.* The net effect of each column continuously adjusting its level of habituation to the background noise is to tend to keep the overall level of activity over the entire memory surface within reasonable bounds. Thus if the overall activity is tending to increase, the columns will tend to habituate more, and thereby reduce their activity. Conversely, if the overall activity is tending to decrease, the habituations will decrease and the activity of the columns will increase.

The computational characteristics of the memory surface may be considered in terms of the interaction of patterns. The simplest case is when the pattern P consists of only one column A whose firing is independent of, i.e., uncorrelated with [Von Neumann 1958], the firing of any other column. Averaged over time and for an arbitrary direction d, A will receive the same number of pulses per second from direction d, whether or not A is active, simply because the activity of A is uncorrelated with the activity of the other columns. Therefore, even though m_d is only altered when A is active, the value of m_d will reflect the average background noise which strikes A from direction d. Whenever A becomes active, say with instantaneous frequency f, it will transmit a pulse train of instantaneous frequency fm_d in each direction d. As these pulses arrive at adjacent columns, they will be relayed onwards and will continue in the same direction. Therefore, pulses will arrive at every column of the memory surface. The net effect will simply be to increase the background noise somewhat. No particular direction will be selected by A for extra stimulation because the number of pulses transmitted in each direction will be proportional to the average background noise in that direction. Thus the presence or absence of activity in A will have no particular effect.

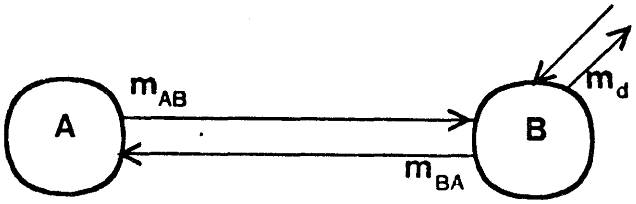

Figure 5 - Two columns whose activity i s *correlated*

The next simplest case to consider is when the pattern P consists of two columns A and B whose firing is somewhat correlated. As before, the values of m_d for all d and for both columns A and B will reflect the average background noise arriving from the remainder of the memory surface at direction d. The only exceptions will be the values of m_{AB} and m_{BA}, where m_{AB} is m_d of column A in the direction of column. B, and m_{BA} is defined similarly (see figure 5). Whenever B is active, it will transmit pulse trains in all directions, and in particular in the direction of A. Since the activities of A and B are correlated, A will tend to be active simultaneously with B more often than if A and B were active at random, and therefore m_{AB} will be increased above the value which represents just background noise. Define the period of correlation t to be the amount of time A and B are simultaneously active, minus the amount of time they could be expected to be simultaneously active if their activities were independent i.e. **uncorrelated.** Then the increase in m_{AB} will be proportional to both the period of correlation and to f_B, the average frequency of B during the period when A and B axe simultaneously active. Similar Comments apply to m_{BA}.

Now whenever A becomes active, say with frequency f, it will transmit a pulse train of instantaneous frequency fm_d in each direction d. For most directions, the effect will be to just increase the background noise, but in the direction of B, there will be a pulse train of frequency fm_{AB} which represents the usual increase in background noise plus an amount proportional to ftf_B. Recall that t is the period of correlation. Thus A sends some extra stimulation in the direction of B. Effectively, the increase in m_{AB} which occurs by virtue of the correlation between A and B serves to help A remember the degree of correlation, the direction of B and the average frequency f_B of B. The amount of this learning is proportional to tf_B. Thus the longer A and B are correlated, and the stronger the activity of B during the correlation, the stronger will be the memory that A retains about B. When A becomes active at some time after it has learned about its correlation with B, the amount of stimulation it will send in the direction of B is proportional to the instantaneous frequency f of A. Thus A will tend to stimulate those columns with which it is correlated by an amount dependent upon its own activity. When A is weakly active, it will only weakly stimulate others; when t is strongly active, the stimulation will be proportionally strong. For any fixed activity f of A, the amount of stimulation A will send in the direction of any B with which it is correlated is proportional to the degree of learning tf_B for that B. Thus A will stimulate each column to which it is correlated by an amount dependent upon its degree of correlation. A column which is strongly correlated with A **will** be strongly stimulated; one which is weakly correlated will be weakly stimulated.

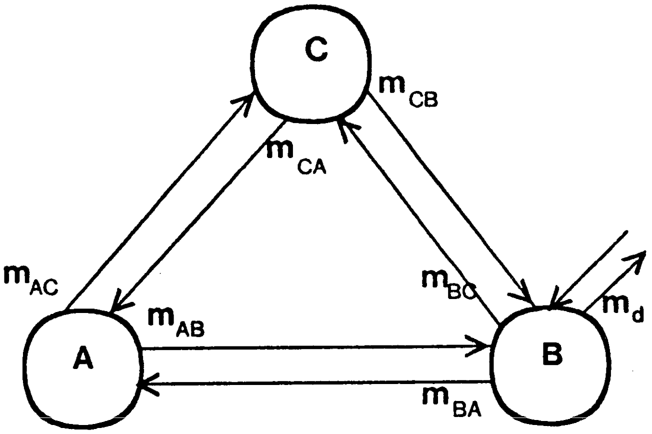

Figure 6 - Three columns whose activity is mutually correlated

It seems that when A is active and it stimulates a correlated column B, it will also stimulate all other columns which happen to be in the same direction as B, even though they are not correlated with A. This is true, but recall that in practice patterns consist of many thousands of correlated columns, rather than just two. The effect can already be seen in the case of a pattern consisting of three mutually correlated columns A, B and C (figure 6). In particular, each pair of columns is correlated and thus they have learned of each other's correlation as described above. If A and C become active simultaneously, each will send a stimulation above the background noise towards B. The stimulation from A will also reach the other columns along the direction AB, and that from C will reach columns along CB. Clearly only B at the intersection of these two lines will receive the combined stimulation from A and C. This effect will be more pronounced for patterns consisting of large numbers of columns.

In general, when a pattern consists of many columns which are active simultaneously, those columns will learn of each other's correlations. At a later time, when only some of those columns become active, they will tend to stimulate the rest of the columns in the pattern., Consider again figure 6, and let f_A, f_B, and f_C be the average frequencies of A, B and C over the times they are active simultaneously. So the components of m_{AB} and m_{CB} above the background noise will each be proportional to f_B. On average, when A, B and C are active simultaneously, B will receive stimulation from A and C above the background noise proportional to $f_A f_B + f_C f_B = f_B(f_A + f_C)$. If it happens that at a later time A and C are simultaneously active with instantaneous frequencies kf_A and kf_C, then B will receive stimulation from them proportional to $kf_A f_B + kf_C f_B = kf_B(f_A + f_C)$. So not only will the activity of a subset of columns in a pattern tend to excite activity in the remainder of the columns in the pattern, but those columns will be excited with strength proportional to the strength of the pattern. In this sense, it is reasonable to say that the pattern has been *stored* on the memory surface. .

This section has described the neuronal level of the model. The term memory surface .has been defined and its relation to the nervous system of an animal has been given. Patterns have been introduced as the unique method in which all concepts are represented. It has been shown that patterns can be active on a memory surface with varying strength, and that patterns can be stored on the memory surface. This neuronal level model reflects the reliability of the brain in a number of ways. The storage of each concept is distributed over thousands of columns, so that destruction of a few columns will lead to a slight degradation of many patterns, rather than to a total loss of one concept. The information contained in communications between columns is encoded as the frequency of a pulse train, so that accidental removal or insertion of a few pulses will not have a catastrophic effect on the overall function [Von Neumann 1958]. Finally, it is the rough correlation between pulse trains which is important, rather than their exact timing relationship, so that small errors of synchronisation between the millions of columns on the memory surface will not greatly impair its operation.

The discussion to this point answers a question of interest concerning what the fundamental unit of memory is, and in what form it is stored. This section has suggested that the memory resides in the degree to which a column is willing to originate the transmission of pulses in a given direction, and (in a shorter time frame) in the habituation of a column. One possible implementation of a column was presented which demonstrates that both these forms of memory can be explained entirely in terms of sensitization and habituation at individual synapses [Kandel 1979].

3. Associations

Concepts are represented in the brain by patterns. Patterns can be active on the memory surface with varying strength, and they can be stored on the memory surface in the sense that if part of a stored pattern becomes active with a certain strength, the remainder of the pattern will tend to become active with the same strength. How do patterns interact with one another, or in other words, how does the brain perform computations with patterns?

Patterns interact only when they are simultaneously active on the memory surface. When any two patterns are simultaneously active on the memory surface, they will become *associated* together by exactly the same mechanism which, in section 2, was responsible for causing two correlated columns to store a memory of the correlation. More specifically, say that the activities of patterns P and Q are correlated, in that P and Q are active simultaneously for a period t with average strengths p and q respectively. Consider any column A with relative activity a in P, and any column B with relative activity b in Q. Then because P and Q are correlated, A and B will be correlated with activities ap and bq, respectively. So the amount of learning in A concerning B will be bqt, and the amount of learning in B about A will be apt. In general, P will learn about its correlation with Q to a degree proportional to qt. Similarly, Q will learn about P to a degree proportional to pt. If P is later activated with strength s, it will tend to excite Q with strength sqt. Thus the strength to which Q is excited by P is proportional to the strength of P and to the degree of association from P to Q which will be represented by $P \rightarrow Q$ and which is equal to qt. Notice that this need not be the same as $Q \rightarrow P$ which is equal to pt.

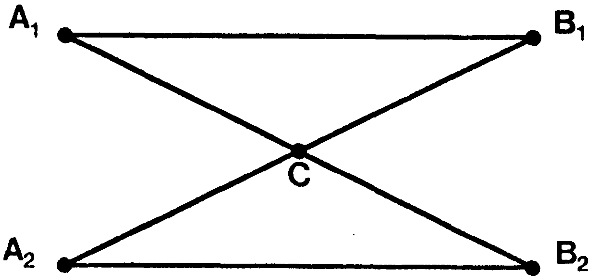

Figure 7 - An extra column becomes activated.

It may also occur that when P and Q are active simultaneously, certain columns which are neither part of P nor of Q could become activated. In the example shown in figure 7, A_1 and A_2 are columns of P, B_1 and B_2 are columns of Q, and C is the extra column which becomes activated. As soon as such an extra column C becomes activated, it will begin strengthening its learning of the other extra columns, as well as those in P and Q. The set R of those extra columns can be thought of as representing a new concept, namely the association of the concepts represented by P and Q. For example, if P is the complex concept of a writing implement, and Q is the complex concept of a round roller, then their association R may represent a ball-point pen. In practice, new concepts may be formed by the association of many simpler concepts, each with a particular strength relative to the others. Which simpler concepts are combined in which strengths to make up our more complex concepts is not presently known.

The foregoing discussion indicates that patterns and their associations may conveniently be represented by directed graphs (figure 8). The nodes of such a graph represent patterns, and the numbers labelling the edges represent the strength of association $P \rightarrow Q$ from the pattern P to the pattern Q. Directed graphs must be used (i.e. the edges have arrows) because in general $P \rightarrow$ Q may differ from Q $\rightarrow P$ as shown earlier. In practice, many patterns will have extremely weak associations between them and so the corresponding edges would not need to be included on the diagram.

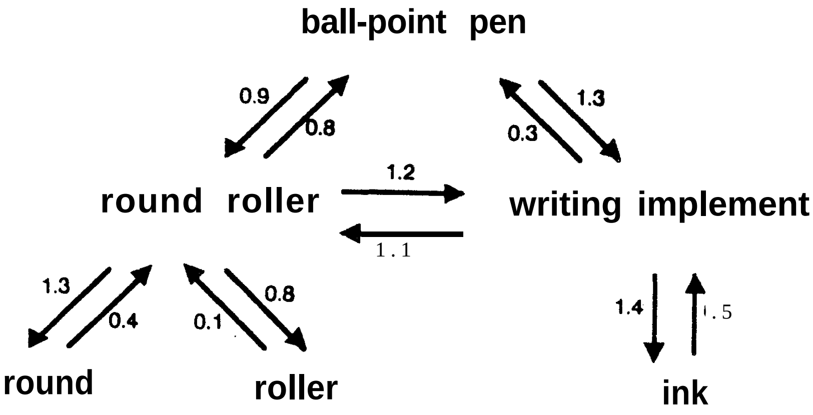

Figure 8 - Directed graph representation

In summary, two patterns P and Q can only interact by being active simultaneously on the memory surface. In that circumstance, P will form an association to Q to a degree depending on the strength of Q and on the duration for which they were simultaneously active. If later P is activated, it will tend to activate Q with a strength depending on the strength of P and on the degree of association of P to Q.

The following paragraphs argue that this mechanism of association which forms more abstract (or higher level) concepts from simpler concepts is sufficient to explain all human concepts. In both his earlier Treatise [Hume 1738] and in his later Enquiry [Hume 1777], David Hume held that there are only three principles whereby concepts are formed, namely by association of simpler concepts which are related by *resemblance*, *contiguity* in time or place, or *cause and effect.* There seems to be no evidence at present which contradicts this view. Rather, these three principles will be examined to see if they can arise from the single mechanism of association described above.

Contiguity in place refers to the fact that when objects are often observed to be physically together they will become associated. The frequent simultaneous observation of the objects will cause the patterns on the memory surface which represent those objects to be simultaneously active for a long total duration, and this explains the forming of an association between the objects.

Contiguity in time refers to the fact that when events often occur in a sequence over time, those events will become associated. Examples are learning the alphabet, and remembering a song. When the alphabet is recited A, B, C,. . ., the patterns representing the concepts A, B, C,. . . will be active in sequence on the memory surface [Kohonen 1984]. As soon as the concept A becomes active, its columns will tend to stimulate each other and thus the pattern will persist on the memory surface for some time. However, the columns comprising the pattern will increasingly habituate, and therefore the strength of the pattern will decrease with time. By the time concept B becomes active, it will be simultaneous with the weaker strength pattern A. Therefore an association $A \rightarrow B$ will be formed which is relatively strong (since B is relatively strong), whereas the association $B \rightarrow A$ will be relatively weak (since A is relatively weak). Similarly, by the time C becomes active on the memory surface, B will have habituated somewhat so $B \rightarrow$ C will be relatively strong while C $\rightarrow B$ is relatively weak. The situation is' illustrated in figure 9. Note that $A \rightarrow$ C is weaker than $B \rightarrow$ C because A will cease to be active before B does and so the duration over which A and C are simultaneously active is less than that over which B and C are simultaneously active. The association C $\rightarrow A$ will be particularly weak because the duration of the simultaneous activation and the strength of A during that period are both small.

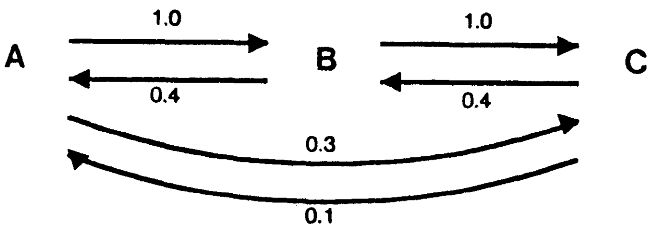

Figure 9 - The sequence A, B, C

After a sequence A, B, C, D,. . . has been learned, as described above, if the pattern B then becomes active, it will tend to stimulate C which will in turn tend to stimulate D and so on. Therefore the remainder of a sequence can be recalled by activating an earlier part of it on the memory surface. Indeed, if instead of just activating B, first A and then B are activated, it will be even easier to recall the sequence because both A and B have associations with C and thus C will receive an even greater stimulation. This effect is particularly noticeable in the case of recalling songs. However, if B is activated first and then A is activated, the stimulation of C will be lower because by that time B which has the stronger associative link with C will have habituated. This model also explains why sequences which are learned in one direction are harder to recall in the reverse direction.

If B becomes active, C will be stimulated more than A because $B \rightarrow C$ is stronger than $B \rightarrow A$.

Cause and effect, another way in which concepts may be associated, is really a special case of sequence. If some observations A, B, C, D of the physical world are made while the world undergoes a continuous change from observation A to observation D, the sequence of observations will be stored on the memory surface as described above. Later activation of the “cause” A will give rise to an activation of the “effect” D and in this manner cause and effect are associated.

It remains to explain how two concepts which resemble one another can come to be associated.’ The clue resides in the meaning of the word “resemble.” Two chairs resemble one another even if one is a stool and the other an armchair. Two numbers resemble one another by virtue of the fact that they are both numbers. Two paintings may resemble one another because they both consist entirely of pencil strokes and the only colors used are red and black. In general, two concepts resemble one another when they share some common set of features. For example, the abstract concept of chair may represent all those objects which can support your weight and which have a horizontal surface (of course, the real concept of chair is far more complex than this), so “support your weight” and “horizontal surface” are common features of all chairs. Because of this common set of features, two different chairs stored on the memory surface will have associative links as shown in figure 10, even if the patterns representing the two chairs were never present simultaneously on the memory surface.

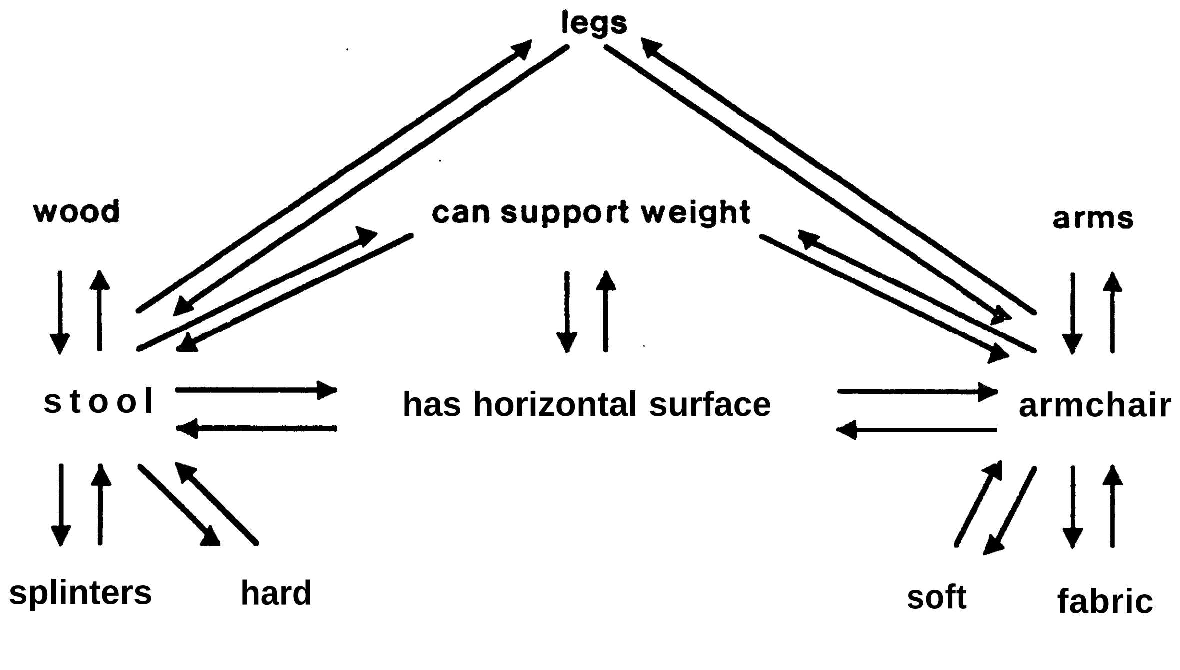

Figure 10 - Two chairs which have associative links

Therefore, there is no difficulty understanding the association of concepts by resemblance, provided that concepts are stored on the memory surface with associative links to their features. But this is exactly the manner in which higher level concepts are formed: by the simultaneous occurrence of simpler concepts (the features) whose very simultaneous occurrence defines the higher level concept. If many red objects are observed, in each case the columns which are connected to the red light sensors will be activated. These columns will form associative links with each other because they are often activated simultaneously. Thus a pattern which consists of these columns will be stored on the memory surface, and this is exactly the pattern which corresponds to, or represents, the concept red. If observations of the color red often coincide with observations of a person's face and hearing heavy breathing, the pattern "heavy-breathing-plus-red-face" will be formed. This pattern represents the concept "physically exhausted". (Of course, the real concept of physically exhausted is far more complex.) Concepts are stored with associative links to the features which define them, and so concepts which resemble one another will be linked associatively via their common features.

The discussion above shows that the ability to generalize is an important computational property of the memory surface. Each time a chair is observed, the set of features which comprise that particular chair will be activated simultaneously. Once the features may be "soft, horizontal-surface, supports-weight", another time "wood, horizontal-surface, supports-weight", a third time "canvas, horizontal-surface, supports-weight". It is clear that the patterns "horizontal-surface" and "supports-weight" will form very strong associative links with each other because they occur simultaneously so often (see figure 10.) The resulting pattern "horizontal-surface, supports-weight" can be regarded as a generalization of the particular instances of chairs which were observed. Thus the process of generalization is the same as that of forming more abstract concepts by the association of simpler concepts. Note that when the generalized concept becomes active at a later time, it does not necessarily follow that the particular instances of the concept will also become activated. There may be so many associated instances that the total activity on the memory surface when spread among them is insufficient to activate any one of them.

The ability of the memory surface to generalize, or abstract, is a reflection of the scientific method, sometimes called induction. One jumps to generalizations from specific instances, whether the generalizations concern causes and effects, or common properties such as mass and momentum. Indeed, the natural tendency of the memory surface to rapidly generalize is observed quite readily, for example when children automatically place an "s" on the end of each noun to form a plural, even though they have always heard the word "mice" and never "mouses."

The tendency of people to become narrow-minded, to form pet topics of conversation and indeed, to fall into many types of habits, can also be seen as a natural consequence of the mechanism of the memory surface. Whenever A and B occur together, $A \rightarrow B$ will be reinforced, but when A and B do not occur together, there will not be much change to $A \rightarrow B$. Similarly, if there is an association $A \rightarrow B$ and A occurs, B will tend to be activated and thus $A \rightarrow B$ will tend to be further strengthened.

For example, due to a coincidence or some other reason, one might begin ever-so-slightly to entertain the notion that when a person spills salt (event A), then something unusual will happen to that person (event B). Now over a number of years, event A will sometimes be followed by event B, and sometimes it won't, purely by chance. However, every time B follows A, it will reinforce $A \rightarrow B$, but when B does not follow A, A $\rightarrow B$ will not be changed. (Of course, a person who is conscious of the effect described in the previous sentence may note when B does not follow A and take conscious steps to reduce the impact of the superstition $A \rightarrow B$. However this is a separate story superimposed on the basic operation of the memory surface as described in this section. Consciousness is discussed in section 5.) In general, it is easy to form habits, but difficult to break them.

It has been shown above how the computation of the memory surface serves to form abstract concepts from simpler ones which are associated, whether by resemblance, contiguity in time or space, or cause and effect. Abstraction is one aspect of the operation of the brain. A second aspect is the manner in which the brain progresses from one thought to another. The method has already been foreshadowed in the discussion of sequences above. At any moment of time, some set of patterns will be active with various strengths on the memory surface. Each of these patterns will tend to excite all the patterns to which it is associated with strength proportional to its own strength and to the strength of the associative link. Meanwhile, the currently active patterns will tend to'habituate and their strengths die away. At the same time, the overall level of activation on the memory surface will cause a general background noise which will affect the average level of habituation. Thus the general level of activity will remain within reasonable bounds, as discussed in section 2.

So the memory surface will exhibit a *flow of activity* from concepts to associated concepts. At any time, events in the outside world can activate their corresponding patterns on the memory surface, and these, together with all other currently active patterns, will give rise to all their associated patterns, and so on ad infinitum. Thus, rather than a single train of thought, the memory surface will exhibit simultaneous trains of thought, each one branching out to its many associated concepts. However, the overall level of activity will be kept in check so that weaker associations may not become active at all. Currently active patterns, including those activated by receptors, will habituate and become inactive. If such an inactive pattern is not stimulated for a short period, the habituation will wear off and the pattern may then become active again when sufficiently many of the currently active patterns which are associated to it are present with sufficient strength. For example, in order to temporarily remember a name, one activates the name periodically rather than steadily holding it in mind.

Habituation to input patterns. is easy to observe. One quickly gets used to general background noise and ceases to notice it. One feels one's clothes for only a few moments after dressing. Similarly, it is not easy to concentrate on a particular thought. One's mind tends to drift off along associated tracks. It is also possible to observe that more than a single train of thought can be active on the memory surface at a given time. It is common to be able to drive a car and carry on a conversation at the same time. Of course, if one gets too involved in the conversation the car might crash, and conversely, if the road conditions become very tricky, one may stop conversing for a short time to concentrate more on the road. This illustrates the upper bound on the overall level of activity possible on the memory surface. Another phenomenon which indicates that a number of patterns may be active simultaneously, as well as further illustrating the associative properties of the memory surface, is the well known trick for remembering something. If one sees a face but cannot quite remember the associated name, one can try to assemble other facts which are associated with the face. For example, while looking at the face one might remember that he attended the spring conference last year and he blows his nose loudly. Bingo, the name comes to mind. The face alone did not have a sufficiently strong association to the name to activate the name, but the face plus the nose plus the conference simultaneously active on the memory surface were able to supply a sufficient total stimulation to the name.

The currently active patterns may be thought of as the contents of a short-term *memory*, whereas the associative links between patterns may be considered to be the contents of a *long-term memory*. Forgetting in the short-term memory occurs fairly rapidly through habituation. In the long-term memory, forgetting can take place in two ways. Firstly the degree of sensitization of synapses slowly decreases over time and so patterns and associative links between patterns are slowly forgotten. Of course patterns and associations which occur frequently will be reinforced, and thus tend to be remembered. Secondly, if $A \rightarrow B$ is currently stored in the long-term memory and a new association $A \rightarrow$ C is formed very strongly, then although $A \rightarrow B$ is still present, it may be effectively forgotten (or blocked [De Bono 1969]) because each time A is activated, C will become activated with a great strength and thus prevent B from being activated. In the course of time, $A \rightarrow$ C will be more and more reinforced, while $A \rightarrow B$ will become progressively weaker.

The flow of activity from patterns to associated patterns on the memory surface may also give rise, from time to time, to patterns which cause the effectors to take some action. These may be just single patterns (e.g. representing the simple concept “close the eyes”), or else sequences of patterns in order to carry out a sequence of actions. For example, in learning a motor skill such as knitting or writing the letter “e”, one tries to activate the patterns which give rise to each piece of thc action is their correct sequence. At first the sequence is rehearsed step by step, and the more the sequence is repeated, the stronger becomes the association between the patterns representing adjacent steps. So motor skills are learned in the same way as any sequence. More complex motor skills can be learned in a similar manner by rehearsing a sequence of simpler motor skills. After a motor skill has been practised many times, the associative links between adjacent patterns of the sequence become so strong that only a small amount of activation of the first element of the sequence will give rise to the entire chain of activation along the sequence. Thus very familiar motor skills will not need much activation on the memory surface, and therefore they can proceed simultaneously with other trains of thought on the memory surface. For example, it is easy to hum a familiar tune while doing something else quite unrelated.

‘Not all effectors produce external bodily actions. Some affect internal bodily functions such as release of hormones, and some control portions of the brain other than the cortex such as input preprocessors or brain regions which were earlier on the evolutionary scale (perhaps including regions which are responsible for emotions). Thus the activation of certain patterns may inform the visual preprocessors of what the brain expects to see, so the preprocessors may complete missing lines and compensate for shadows and color of incident light and so on. The activation of other patterns on the memory surface may cause effectors to release chemicals in the memory surface itself and thereby alter parameters like the degree of learning of associations or the rate of habituation of the columns. This could result in ‘altering the speed at which current thoughts give way to their associated thoughts, and could also affect the average level of activation on the memory surface, thereby changing the general level of arousal.

This section has described the basic method of operation of the memory surface. The static aspect (memory) is the association of concepts, and the dynamic aspect (processing) is the flow of activation from concept to associated concept. Both aspects result from the behavior of columns on the memory surface, as discussed in the neuronal level model of section 2. Both the memory and processing are distributed over the memory surface [Lashley 1942, Pribram 1971], in so far as patterns may consist of any set of columns distributed over the memory surface. Thus concepts which originate from quite different sensors and which enter the memory surface at completely different locations can nevertheless become associated, as when thc hearing of a bell becomes associated with the taste of food. Similarly, in the case of damage to the brain, high level skills and concepts are often impaired rather than lost since some portions of the relevant patterns and their associations may remain. On the other hand, low level concepts such as detecting light in a specific region of the visual field or moving the left leg may be totally destroyed by damage to those specific regions of the memory surface which are connected to the appropriate sensors and effectors.

4. Creativity and sleep

The ability to be creative has to a large extent been responsible for the current situation of humans. The realization that sticks and stones could be used as tools, that fire could be created by rubbing sticks together, that crops could be planted by judicious placement of seeds, that crops could be improved by careful selection of seedlings and so on, are all examples of the creative act. Important instances of creativity in both science and art are well known. However, many examples of creativity can be found every day in every person. The child who discovers that the washing basket is a good hiding place has performed a creative act, as has the parent who invents a new entertainment for the childrcn such as rides in the wheclbarrow. Animals also exhibit acts of creativity. Monkeys have been observed discovering completely on their own accord that a branch could be used to retrieve a banana beyond their normal reach. Other monkeys have discovered that food and sand can be separated by throwing handfuls of the mixture into the water. These are all examples of original thinking. Of course, once the original creative act has been performed, the good idea can be passed on to other people or animals. For example, before too much time had elapsed, all the monkeys in the colony (except the very oldest) were using the new method for separating sand and food. Indeed, this method was passed down to successive generations. But the invention of the method is an example of original thinking, or creativity.

Arthur Koestler [Koestler 1965] argued that all creativity, including artistic creativity, scientific discovery and even humor, is the combination of concepts which are not usually combined. Thus original thinking does not involve creating something out of nothing; rather it is a novel combination of old ideas. There seems to be no evidence currently known to contradict this view. In terms of the model presented here, creativity is the simultaneous activation of two patterns which do not have strong associative links with each other, but which do have strong associative links with a common set of patterns. When two such patterns A and B are simultaneously active, their common set of patterns will receive an unusually strong stimulation since they are being stimulated from both A and B. Thus an unusually large amount of activation will be concentrated in the columns belonging to the common set of patterns. The larger the common set of patterns, the larger will be the sudden jump in activation. Such a sudden jump in the level of activation can be detected by sensors and is probably what is responsible for the "AHA reaction" [Koestler 1965].

An example of a creative act is illustrated in figure 11. When Kepler was studying the orbits of planets, he knew that the orbits were egg-shaped and he discovered certain other properties of the orbits, such as the fact that in constant time, the planets swept out constant area. He also knew that the mathematical curve called the ellipse is egg-shaped and he may have known properties of ellipses that matched those of orbits, such as the constant area property. However it took him a very long time to connect the two concepts of ellipse and orbit. When he thought about orbits, the associated concepts of egg-shape, constant-area, and many others would have been stimulated, and those in turn would have provided some stimulation to the concept of ellipse, among very many others. But for a long time, ellipse did not become sufficiently active for the connection to become apparent. If one day orbit and ellipse became simultaneously active, as apparently happened, their common patterns, egg-shape and constant-area, would have received double their usual stimulation and this increase in activity may have caused Kepler to feel the AHA reaction.

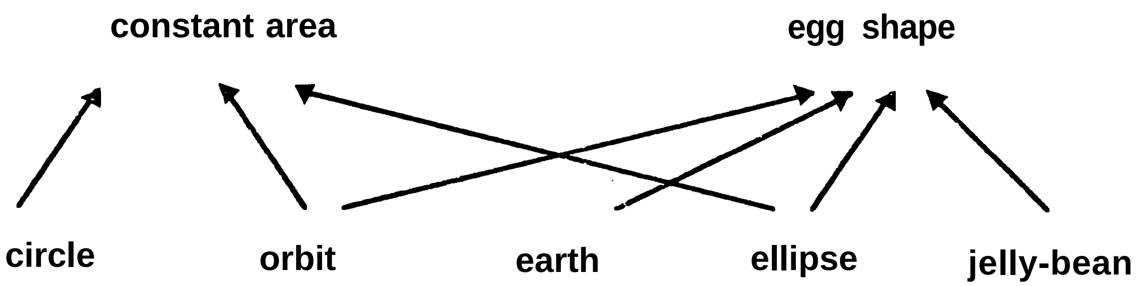

Figure 1 1 - A set-up ripe for a creative leap

A classical case of the AHA reaction is Archimedes yelling eureka as he jumped naked out of the bath. The reason for his pleasure was his sudden association of the problem of measuring the volume of complicated objects with the rise in the level of water as objects are inserted. Again the creative step (or "intuitive leap") can be explained by the simultaneous activation of two concepts with a number of common features, the sudden increase in activation of the common patterns being felt as the AHA reaction. But why do the two concepts become simultaneously activated in the first place? This is the result of pure chance. In Archimedes' case, the chance consisted of him thinking about his problem at the same time that he was observing the water level as he sat in the bath. In the case of the child hiding in the washing basket, the chance event may have been tripping over the washing basket while thinking about where to hide. The banana-reaching monkey may have luckily happened to look at a branch while thinking about having too short an arm to reach the banana. ' .

In every case, the creative act involves some luck which is respdnsible for getting two previously unassociated (or at most weakly associated) concepts simultaneously active on the memory surface, in conjunction with a certain degree of preparation on the part of the brain, consisting of having appropriate features associated with the concepts and being able to recognize the AHA reaction when it occurs. All the examples of luck which have been mentioned so far have been supplied by the environment outside the body. Chance arises essentially from the fact that the environment is sufficiently varied that patterns stimulated on the memory surface by the sensors are not always correlated with the patterns currently active on the memory surface. Of course, there is no need for animals to just wait around until they are impressed by lucky inputs. Animals actively seek out varied inputs in order to increase the likelihood of forming useful associations. Thus animals are naturally inquisitive, children engage in play, newborn babies scan the environment with their eyes [Haith 1980] and techniques for encouraging lateral thinking [de Bono 1969] are indeed effective in promoting creative ideas.

Creativity is such a useful activity for day-to-day problem solving (and thus survival) that it would seem to be a good idea for a brain to be equipped with a mechanism which would regularly stimulate the memory surface in a random manner, in order to encourage creative thinking and break down the natural tendency to form fixed habits of thought and action. Unfortunately, the requirements of survival in a competitive environment (e.g. a pedestrian crossing a road) are such that a high level of attention must be continuously directed at the inputs arriving from the environment. Thus the sensors must strongly stimulate the columns to which they are connected, and it will be important to think clearly and quickly. It would be inappropriate to randomly stimulate the memory surface under these conditions. However, there may arise periods during which it is less dangerous than usual to reduce this level of attention to the external environment, such as when it is too dark or too cold to hunt or to be hunted. During such periods, the possible advantages to the memory surface of reducing the level of awareness of the environment may outweigh the obvious disadvantages. Indeed, almost all animals do reduce their level of awareness while sleeping.

Sleep consists of two quite distinct types, active (or "rapid eye movement") sleep and quiet sleep, which are interleaved throughout the night. People woken from active sleep often report dreams, whereas those awoken during quiet sleep report their thinking drifting "from idea to idea, merely by a process of association, an experience very similar to that of the daydream" [Meddis 1977]. Babies sleep for a large proportion of the time, and this proportion progressively reduces with age. Babies also appear to have a large fraction of their sleep devoted to active sleep, with this fraction again reducing with age. How can these findings be explained?

From earlier discussion, it is apparent that when a chain of patterns is activated on the memory surface, the corresponding associations are strengthened or reinforced, whereas if an association is not used for a lengthy period it will tend to diminish in strength, or be forgotten. Thus during normal daily operation, chains of patterns closely connected to the input patterns which are continuously arriving with great strength on the memory surface will tend to be reinforced the most, whereas memories or skills which are not so immediately important, but which nevertheless may be of some benefit at a later time, will tend to be forgotten. The role of quiet sleep may be to give some reinforcement to those less immediately important patterns. The strength of the incoming inputs is reduced in order that it will dominate the total activity of the memory surface to a lesser extent. Thus longer chains of patterns can be activated without being suppressed by the stronger activity of the latest sensations, as would happen while awake. Associations between patterns which are more distant from input patterns will be reinforced during quiet sleep.

Unfortunately, the reinforcement which takes place during quiet sleep suffers from the same problem as the reinforcement which takes place while awake. Namely, it is the strongest associations which are activated the most and which therefore are reinforced the most. The memory surface has a tendency towards forming fixed habits of thought, thereby blocking alternative paths of associations which may be useful at odd times. The solution to this problem is to incorporate a mechanism into the brain which will provide random stimulations of patterns so that weaker chains of associations which would normally be blocked can be activated from time to time and therefore reinforced and not completely forgotten. For reasons discussed above, it would be inappropriate for such random stimulations to occur while the animal is struggling to survive. However, they can occur while the animal is asleep, and this then is the role of active sleep.

During active sleep, a component of the brain outside the memory surface periodically provides random stimulations to the memory surface. The stimulation is random in the sense that there is no correlation between the columns which are stimulated and the patterns and associations which were being formed and reinforced while the animal was awake. Each burst of random stimulation will cause a random set of patterns to become active, and these active patterns will then lead to a flow of activation from pattern to *associated pattern in the normal manner, until the next burst of random stimulation arrives. Hence possibly unusual chains of associations and unusual combinations of concepts will be activated. Indeed, dreams seem to be the playing out of unusual combinations of concepts, with periodic jumps (or addition of new dream material) as the random bursts of activation occur [Crick and Mitchison 1983]. Recurring dreams may be caused by strong chains of associations whose starting point is normally blocked [Freud 1900]. As soon as the starting point is activated by a random stimulation, the dream will recur i.e. the flow of activation will proceed from the starting point along the strong chain of associations. Thus recurring dreams will tend to be reinforced, just as all activated associations will be reinforced. In general, active sleep helps the memory surface to retain events, concepts and skills which are not necessarily used regularly or frequently by the animal. Nevertheless, it may be difficult to remember the thought content of dreams after awakening, simply because there may not be a strong path of associations leading to the start of each segment of the dream.

The memory surfaces of young children differ from those of older people mainly in the number of patterns and associations which have been stored. The patterns which become active on a young memory surface will not have as many existing patterns to moderate them. The young child will be more impressionable, will tend to form generalizations faster and will be more vulnerable to forming strong associations too early, before all the evidence is in. At the same time, children have the protection of their parents, and so are less likely than adults to be in danger while sleeping. Therefore younger children will sleep more, and in particular will have more active sleep, in order to increase the random stimulation to their memory surfaces and thereby reduce the natural tendency to form narrow habits of thought (such as overgeneralizations) too quickly.

As a corollary of its memory function, active sleep brings together concepts in unusual combinations. That is, a random burst can activate simultaneously two quite separate concepts which are not usually associated. If these two concepts happen to have many features in common, the AHA reaction will be felt. It is therefore possible for people to wake from sleep with a good idea in mind [Hadamard 1954, Fishbein 1981]. The commonplace advice "sleep on it" for people with a problem to solve or a decision to make is thus not without some basis. If a person has a problem "foremost in their mind", i.e. the patterns and associations relating to that problem have been reinforced recently by much concentration on the problem, then those patterns and associations increase their likelihood of being activated as part of a random stimulation. Such patterns may thus be stimulated more often in conjunction with other random patterns and therefore have an increased chance of a lucky strike upon a solution.

In summary, creativity consists of a lucky simultaneous activation of two patterns with many features in common. Such an activation is felt as the AHA reaction. The luck may be provided by the environment external to the animal, or by a random stimulation during active sleep. Sleep has two survival values, namely to help memory of infrequently used concepts and associations and to encourage creativity. These functions of sleep make necessary a reduction in the impact of the senses on the memory surface. The unusual nature of the patterns which may become active during sleep (especially during random bursts) also necessitates the inhibition of the effectors controlling gross motor actions, to prevent random movements which may inflict self-injury. This inhibition can be observed as the relaxed posture of sleeping humans. There is no need to inhibit minor motor actions, and in some cases they are not inhibited, as in the case of the rapid eye movements during active sleep.

5. Self and Consciousness

In the model of brain function presented in previous sections, all concepts are represented as patterns on the memory surface. The simplest concepts are those most closely linked to sensors or effectors. All more complex concepts are built up from the association of simpler concepts. However the exact details of which simple concepts are combined to give a particular complex concept are poorly understood at present. Examples include such high level concepts as number, well-dressed, volume and playful. Another example is the concept of self.

Although the concept of self is not fully understood, neither is it particularly mysterious compared to other high level concepts. As discussed earlier, a useful analogy may be drawn between scientists modelling aspects of the universe and the memory surface forming a hierarchy of progressively more abstract concepts by abstracting common fea-. tures from patterns which arrive from the environment. Such a hierarchy is just a model of those aspects of thc universe which may be relevant to the survival of the animal. The more accurate the model, the more likely is the animal to interact with the environment in such a manner as to promote its own goals. (For this reason, although it is true that the models formed by our brains and indeed the patterns arriving from our senses are only approximations and distortions of the environment, nevertheless the distortions are probably not too bad, otherwise we would have difficulty surviving. For example, it would be most unpleasant if every time we walked towards what we perceived to be a doorway, we would bang our noses on a solid wall.) The concept of self is such a model of part of the universe, namely that part of the universe which the brain can control. At a young age children form an abstraction concerning all those things they can directly control (namely parts of their bodies), as opposed to things which they discover that they cannot directly control (namely the environment outside their bodies). As people learn more about their bodies and their environment, they may refine their concept of self to some extent. When they are presented with an unusual circumstance (e.g. their brain is physically separated from their body [Hofstader and Dennett 1981]), they will be somewhat confused since their model of reality is no longer consistent with their observations of reality. But this is no more mystical than the confusion which might be experienced by a Newtonian physicist on observing that particles moving at high relative speeds do not obey the simple law of addition of velocities. The concept of self is a more or less accurate model which the memory surface forms to distinguish the objects it can directly control from those it can't.

Another high level concept of interest is that of consciousness. The word consciousness actually has a number of different meanings, but all relate to the idea of awareness. The different meanings of consciousness are awareness of one's environment, self-awareness, and awareness of one's train of thought. Awareness of environment (e.g. the sun is currently shining through the window) is simply the activation on the memory surface of the appropriate patterns which model that aspect of the environment. Similarly, self-awareness refers to the fact that the concept of self is currently active on the memory surface. If a person is writing a sentence and does not have the concept of self active at that time, then the person will be conscious (i.e. aware) only of the act of writing taking place. On the other hand, if the concept of self is simultaneously active, then the person will be conscious of their own self performing the act of writing.

The third meaning of the word consciousness, awareness of one's own train of thought, is often called introspection. Introspection is the forming of an association between the concept of self and the other concepts currently active on the memory surface. When the concept of self is simultaneously active with another concept, the pattern resulting from their'association represents the concept of the self having that particular thought. So the activation of that resulting pattern is a particular introspection. Of course one can also be aware that one is introspecting, and aware that one is aware that one is introspecting and so on ad infinitum. But this is not an infinite regress, or a viscious circle. It simply represents the ability of the memory surface to form abstract concepts from simpler concepts, and then more abstract concepts out of those concepts and so on to arbitrary (but at any moment of time, finite) depth. If awareness that the self is writing is one pattern, then the association of that pattern with the concept of self is awareness that one is introspecting about the self writing, and the connection of that association with the concept of self is awareness that one is aware that one is introspecting and so on to an arbitrary, but at any moment finite, depth.

In summary, the concepts of self ‘and consciousness are highly abstract and complicated concepts, but they can be studied within the framework of the model presented here. In a sense conscious thoughts can be regarded as those patterns which are active simultaneously with the concept of self and whose activity is sufficiently strong to form an association with the concept of self. Unconscious thoughts are those which are active when the concept of self does not happen to be active (such as driving a car home without being aware of doing it), and those patterns whose activity is too weak to form an association with the concept of self (such as one step in a learned and heavily reinforced sequence). Subconscious mental activities are those which can never be associated with the concept of self, and therefore introspection will not shed any direct light on them. These are the activities which occur in regions of the brain other than the memory surface, such as the pre- and postprocessors and earlier evolutionary structures.

6. Free will

The brain model presented here is computational and can therefore be considered as mechanistic and deterministic. Randomness was discussed in relation to creativity, but this arises simply because activities outside the memory surface (including other brain structures and the external environment) are not necessarily correlated with events on the memory surface. However, in principle, if all inputs to the memory surface were known, its behavior would be entirely deterministic.

The fact that the brain is deterministic seems to answer the question whether humans have free will in the negative. But this is a somewhat unsatisfactory answer because humans certainly do feel that they have free will. Why then do humans feel that they have free will? To tackle this question it is first necessary to clarify the concept of free will. It is argued that humans ascribe free will to an object in proportion to the lack of predictability of the behavior of that object, provided it is believed that the object exercises control over its behavior. Stated more simply, free will is related to unpredictability. The more predictable the behavior of an object is, the less free will will be attributed to it. Conversely, an organism whose behavior is highly unpredictable will be felt to be exercising a good deal of free will over its activities.

A clue that free will is related to unpredictability comes from observing people who design and use computer programs. Simple computer programs whose behavior is easy to understand usually evoke no feelings in their users. But complicated programs which produce all sorts of results which their users find unexpected begin to be thought of by those users in personal terms. A user might think “I wonder why he moved his rook just now. Perhaps he thinks I won’t see how to counter the subtle pressure he’s exerting on my queen side.” Computer programs are in general very complex objects and it is easy to design a short program which even the designer cannot understand. It is not unusual for computer programs to surprise their own designers.

Inanimate objects can, if they are sufficiently complex, exhibit unexpected behaviors and so they can begin to have free will attributed to them. What happens in the converse case when people have very predictable behaviors ? Imagine that a very large, very fast computer is constructed which has every single detail of your brain wired into it (a task which is impossible in practice because of the vast number of details involved). Now the neurons of your brain operate on a time scale of thousandths of a second but computers can operate in billionths of a second. So if the computer were given complete control over every single input entering your brain, it could make a very rapid calculation of exactly what your behavior will be for the next minute, print a description of the predicted behavior on paper, seal it inside an envelope and hand it to you (say after twenty seconds) with strict instructions not to open the envelope until one minute has elapsed. Of course, handing you the envelope after twenty seconds and giving the instructions are part of the input to your brain which the computer took into account in its calculations. After one minute you open the envelope and discover that the computer's prediction was exactly correct. The experiment is repeated and this time you decide to trick the computer by doing something quite unexpected. Your knee is not at all itchy, so you decide to scratch it. After one minute, you open the envelope and it includes everything, even the scratching of your knee. T,he experiment is repeated and this time you decide to really trick the computer and so you open the envelope prematurely. It reads "you will open this envelope after 47 seconds." Defeated, you realize that your behavior is entirely predictable. Your model of yourself no longer seems to accord with reality. You begin to give up the feeling that you can exercise free will.

Given that free will is indeed directly related to unpredictability, then why do people feel that they have free will? It is because they cannot accurately predict their own behavior. The memory surface retains memories of past events and holds a view of the present in the form of currently active patterns. So it is possible to review your activities in the past right up to the present instant: but there is no view of what is ahead in the future. Of course, by forming a sufficiently accurate model of yourself, it is possible to predict to some extent your future behavior You can predict, more or less accurately, that if you poke yourself with a ball point pen you will feel pain, probably say "ouch" and later feel annoyed at your stupidity. To the extent to which you know that your behavior is predictable, you feel a reduction in your free will. Everyone feels that although they have some free choice, their range of choices is constrained by reality and therefore their degree of free will is limited.

However, it is impossible even in principle to form such an accurate model of yourself that you can predict your behavior exactly and infallibly, like the hypothetical computer. Informally, the reason is that you have some fixed speed of computation, and you cannot compute faster than that. So if you had a complete blueprint of all the details concerning your brain and a complete list of all the inputs which will be applied to your brain over the next minute, it would be impossible to simulate the blueprint on those inputs faster than your own brain was operating. Therefore, although you could eventually compute all your behavior up to some particular time, the computation would not be completed until well after that time. In fact, you do have a blueprint of your brain which is the brain itself and you are effectively simulating it continuously. Therefore you are computing every detail of your behavior. But the result of the computation is always available just in time to execute that behavior. So you can look **backward** in time and you can introspect concerning the current instant, but you can never get ahead of yourself to predict your own exact behavior before it occurs.

More formally, say that you could accurately predict your behavior in the future. Then you could see what that behavior was going to be and you could decide to differ from it. At the appropriate time, you could then perform the different ‘behavior. So your prediction must have been incorrect. Therefore there is no way in which you can accurately predict your own behavior.

Say in order to help you predict your own behavior you employ the hypothetical computer mentioned earlier. Now this computer can accurately predict your behavior, because it can simulate your blueprint so quickly. As soon as it hands you the envelope describing your behavior for the next minute, you tear it open, read the description and purposely perform a different behavior. You seem to have fooled the computer. What has gone wrong? The fact is that the computer can only predict your behavior up to the instant you open the envelope. Part of the input to your brain which the computer was using for its simulation was the fact that you would receive a sealed envelope after exactly 20 seconds. If you leave the envelope sealed, there is no problem and the computer can proceed to simulate your behavior for the remaining time and then produce the contents for the envelope. But if you open the envelope, then the computer cannot proceed with its simulation beyond that point because part of the input to your brain will then consist of the contents of the envelope, and the computer does not know the contents of the envelope at this stage of the simulation. In fact, in order to proceed, the computer would have to predict its own behavior to work out the contents which it will produce for the envelope,- but of course it cannot predict its own behavior for the same reason that you cannot predict your own behavior.

In summary, people never predict completely accurately their own behavior. They can look to the past and the present, but not to the future with complete accuracy. To the extent that they cannot predict their own behavior, they feel that they have free will. Indeed, not only do people never predict completely accurately their own behavior, but it is logically impossible for them to do so, even with the aid of the fastest and largest computing machinery that could ever be built. So in this sense, it can be said that people really do have free will. On the other hand, a faster computer can exactly predict the behavior of a slower computer, such as a human, provided that the slower computer is not told about the prediction before the end of the period to which the prediction applies. In this sense, people do not have free will.

7. *Summary*

Higher brain functions seem to be the responsibility of the cortex. The cortex can be modelled as a memory surface consisting of a large number of columns. Columns can be active to various extents at various times. The strengths of the activities of a set of columns is called a pattern and represents a concept or model of some aspect of the real world. The properties of each individual column and their method of communication causes patterns which are active simultaneously to become associated. All human concepts can be represented by the association of simpler concepts. The fundamental mode of thought of the human brain is the flow of activation from the currently active concepts to the concepts which are associated with them.

Creativity is a chance association of concepts which have many things in common. Creativity can be fostered by lucky occurrences in the environment, but can also be stimulated during sleep, Sleep also serves to reinforce memories which would not otherwise be frequently activated.

The exact combination of concepts which comprise each of our high level concepts, including the concept of self and the many concepts needed to explain human language such as having au approximate model of the listener's concepts [Cohen and Levesque 1980], are not presently well understood although they can be studied within the framework of this model. Acquired modes of thinking, such as deductive reasoning, and acquired algorithms, such as for the addition of two long numbers, can also be studied within this framework. The ability to introspect and the feeling of free will are also consistent with this model.

Brain functions which are not performed by the memory surface require a different framework for their explication. These include the complicated operations of feature detection performed by the preprocessors, and the highly complex feedback control mechanisms of the postprocessors. Also not explained are emotions, drives, pain and so on which may be the domain of brain structures which appeared earlier in the evolutionary chain than the cortex. Finally, real animals probably do not begin life with a completely blank memory surface, but may rather have certain patterns "prestrengthened" so that they can begin a useful interaction with their environment sooner after birth. What these patterns are for each animal is also an interesting area of study. However, the more preconceived ideas the animal begins with, the less flexible will it be in its adaptation to a different environment. For this reason, the model presented here has taken the extreme case of beginning with a blank memory surface.

It may be somewhat disturbing to regard the brain in the mechanistic terms discussed here. However, this feeling of disturbance probably arises from having a model of mechanism as stupid, clumsy and fairly predictable in accordance with most of the actual machines currently in existence. The brain is not a stupid, clumsy or fairly predictable mechanism. But it is convenient to think in terms of a continuum of machinery, with simple and predictable machines at one end and brains at the other. The continuum of machinery is a very rich class, and includes members which are highly complex, self-referential, elegant and extremely difficult to predict. An analogy with animals illustrates the point. Humans are animals, but they are not at the stupid, simple and primitive end of the continuum of animals, such as snails. Humans are not just animals; they are very particular animals with very intricate and interesting properties. Similarly, humans are machines, but they are not just machines [Ryle 1949].

Acknowledgements

I would like to thank Esther Goldschlager and Dimitrios Gounelas for the enormous amount of both help and encouragement which they have given me over the course of this research. I also thank Henry Bankier, Richard Green, Michael Halmagyi, Andrew Lenart, Adolfo Lopez-Otero and Janet Wiles for stimulating discussions. The University of Sydney and Stanford University provided facilities, financial support and, above all, an appropriate atmosphere for the pursuit of this study.

References

Albus, James S., Brains, Behavior and Robotics, BYTE Books, 1981,

Arbib, Michael, The Metaphysical Brain, Wiley-Interscience, NY 1972.

Cohen, Philip R., and Levesque, Hector J., Speech Acts and the Recognition of Shared Plans, Proc. 3^{rd} Canadian Conf. for the Computational Study of Intelligence, Victoria, B.C., 1980, (263-271).

Condillac, Etienne Bonnot, Treatise on the Sensations, 1754.

Crick, Francis and Mitchison, Graeme, The Function of Dream Sleep, Nature, Vol. 304, 1983.

De Bono, Edward, The Mechanism of Mind, Penguin, 1969.

Fishbein, William (ed.), Sleep, Dreams and Memory, Advances in Sleep Research, Vol. 6, 1981.

Freud, Sigmund, Interpretation of Dreams, 1900.

Hadamard, Jacques, An Essay on the Psychology of Invention in the Mathematical Field, Dover, 1954.

Haith, Marshall M., Rules that Babies Look By, Hillsdale, N.J., 1980.

Hofstader, Douglas R., and Dennett, Daniel C., The Mind's I, Basic Books, NY, 1981.

Hume, David, Treatise of Human Nature, 1738.

Hume, David, An Enquiry Concerning Human Understanding, 1777.

Kandel, Eric R., Small Systems of Neurons, Scientific American, Vol. 241, No. 3, 1979, (60-70).

Koestler, Arthur, The Act of Creation, Macmillan, 1965.

Kohonen, Teuvo, Self-Organization and Associative Memory, Springer-Verlag; 1984.

Lashley, Karl S., The Problem of Cerebral Organization in Vision, in Biological Symposia (Cattell, Jaques ed.), Vol. VII, 1942.

Meddis, Ray, The Sleep Instinct, Routledge & Paul, London, 1977.

Pribram, Karl H., Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology, Prentice-Hall, N. J., 1971.

Ryle, Gilbert, The Concept of Mind, Hutchinson, 1949.

Stevens, Charles F., The Neuron, Scientific American, Vol. 241, No. 3, 1979, (48-59).

Ullman, Jeffrey D., Some Thoughts About Supercomputer Organization, Stanford University, 1984.

Von Neumann, John, The Computer and the Brain, Yale University Press, 1958.