3

Neural Networks II:

Models of Mind

Conceptual Outline

3.1 The training of a model network that has subdivisions requires a process that can train synapses within subdivisions and between subdivisions without subjecting either to overload. A natural solution to this problem involves taking the network "off-line," so that a filtering of memories can occur when the network is dissociated. This is a possible model for the role of sleep in human information-processing that explains some of the unusual features of sleep and suggests new experiments that can be performed.

3.2 Various features of human information-processing, including the learning of associations, pattern recognition, creativity, individuality and consciousness can be discussed within the context of neural network models.

Efforts to describe and explain the higher information-processing tasks that human beings are capable of performing have always generated tension and concern. There has been a tendency to elevate these processes outside of the domain of the physical world, or to mystify them,through a characterization as infinite and incomprehensible. This tendency may arise from the desire to maintain a uniqueness of and importance to our own capabilities. We will adopt the contrary point of view that our capabilities are fundamentally comprehensible. However, it turns out that this does not diminish a quality of uniqueness and importance. If anything, it shows us how this importance arises.

We begin to tackle the task of explaining aspects of human information-processing in this chapter. However, we will not conclude our discussion until we have described complexity in the context of human civilization in Chapter 9. We start in Section 3.1 by considering the training of subdivided networks and the role of sleep in human information-processing. This section is an essential sequel to the discussion in the last chapter that introduced subdivision. The problem is to develop a systematic approach to the training of subdivided networks.

3.1 Sleep and Subdivision Training

3.1.1 *Training a partially subdivided network*

In Chapter 2 we discussed the role of functional subdivision in the brain. We showed that the storage capacity of a subdivided network was reduced;however, the ability to recall composite states may confer significant advantages on a properly designed network. The subdivided network presents us with a new set of challenges when we consider its training. The approach used in Chapter 2, where we imprinted a set of distinct patterns—each precisely what must be learned—is woefully incomplete. For more realistic modeling of neural networks we should assume that the information presented to the network is not so well-organized. When the information is presented in either a random or, even more realistically, a correlated fashion, there arise problems that relate to the storage capacity of the network and the selection of desired memories.

To illustrate the problem, we can return to the simplest example—the left-right universe (Section 2.4.1). In the first discussion it was assumed that there were no correlations between the left and right halves. The two halves were completely independent and all composite states were equally possible. Now we assume that correlations exist,and that there are some synapses that connect left and right hemispheres to capture these correlations. This means that there are many fewer possible states than $N_L N_R$ but still many more than N_L or N_R, where N_L is the number of possible left halves and N_R is the number of possible right halves. Moreover, we should expect that the number of possible imprints of each subnetwork is greater than its storage capacity ($N_L, N_R >> \alpha N/2$).

Some selection of which states to keep in memory must be made. The point of introducing the subdivided network was to accommodate more of the possible states that can arise. However, if we try to imprint all of them, we will exceed both the capacity of the subnetworks and the capacity of the synapses between the subnetworks. When we were faced with the problem of overload in a uniformly connected network, we used a palimpsest memory to retain only the most recently imprinted memories. However, in a subdivided network this is not the best strategy for keeping memories. If the imprints are correlated,the most recent ones may all happen to have a particular right half, and this will end up being the only right half that will be remembered by the network.

Thus, without control over the number of states that are imprinted and the order in which they are imprinted, we must be concerned about the problem of overload failure. If we must stop the imprinting after only a few imprints sufficient to reach the capacity of the smallest subdivision, we will have limited the training of the network very severely. How can we overcome this problem?

To design a strategy to overcome the overload problem, we must first identify our objective in training the network. The objective should be based on achieving the best utilization of the available capacity: to enable each of the subdivisions of the brain to store a number of patterns that are commensurate with its capacity. These and no others. The stored patterns should be the most important patterns to remember. How do we identify the most important patterns? They should be the patterns that appear most frequently, and the patterns that are most orthogonal—most different. The reason for orthogonality is that it enables more patterns to be stored (Question 3.1.1). Also, if two patterns are similar, we might be able to substitute one for the other without too much loss. Thus if we cannot store two patterns as distinct, the next best thing is to store them as one and classify them together. More generally, when there are highly correlated states, we store one prototype that could be one of the states, or even a spurious state that has maximal overlap with the correlated states. This best utilization strategy enables each subdivision to retain states that are well representative, even though not necessarily exact reproductions, of the possible states.

The next objective is to train synapses that run between subdivisions to achieve correlations between the patterns imprinted in each of the subdivisions. We can think about each subdivision as itself like a neuron. The difference is that the neuron has only two states while the subdivision has approximately $\alpha N'$ possible states, where N' is the number of neurons in a subdivision. Thus we train the synapses between subdivisions to utilize the storage capacity for composite patterns, choosing those that are the most important composite patterns to remember. As before, the most important patterns are those that appear most frequently and those that are most orthogonal.

If we have a subdivision hierarchy, then at the third level of organization, the stable patterns of each of the subnetworks consists of various composite states. The overall objectives that we articulated for the storage of patterns also apply to the storage of states of the third level of the network. These objectives for training the synapses between subdivisions continue to remain the same all the way up to the complete network.

The model of a network of networks suggests a general strategy for its training. Since the training of inter-subnetwork synapses relies upon a well-defined set of subnetwork states, it seems reasonable to train first the subnetworks and then the synapses between them. To achieve this we would separate the subnetworks from each other, train each one, and then attach them and train the synapses between them. In the first stage of training, the subnetworks would be trained to their optimal capacity. Once the subnetworks are trained, the storage of patterns using the inter-subnetwork synapses would have a well-defined significance. Otherwise, if the synapses between subnetworks were trained before the synapses within a subnetwork were set, then modifying synapses within the subnetwork would change the significance of the synapses between subnetworks. Thus it seems that the brain should be trained by first training the smallest subdivisions, then connecting them into larger groupings and training the synapses between them. However, while it might appear to be convenient to train the subdivisions first, this presents us with several practical problems.

We must assume that the training requires many exposures to various environments and circumstances. We cannot wait until the training of subdivisions is complete before the brain is used. Moreover, the sensory information to which the brain is exposed does not reach the brain except through the interaction of action and sensation. In order to act, the brain must be functional, at least to some degree, and therefore we cannot train a brain when disconnected into its small parts.

There is an alternative approach that takes into account the need for separation and training of each subdivision while enabling the functioning of the brain. This approach adds an additional dynamics to the brain. In addition to the neural dynamics and the synaptic dynamics there is a dynamics of subdivision. The dynamics of subdivision is incorporated in a two-step procedure:

1. Imprint the complete network. If the number of synapses between subdivisions is smaller than within subdivisions,this already contains some predisposition to subdivision. Since the network is connected, it can also function.
2. Separate the network into its subdivisions and selectively reinforce (filter) the patterns that satisfy our objectives of optimal utilization of each subnetwork.

The cyclical repetition of steps 1 and 2 should enable the training of the subdivisions to proceed as operation continues. However, it requires the system to go "offline" periodically for the filtering process.

The purpose of the two-stage process is to train the subdivisions separately from the training of the inter-subnetwork synapses.However,there is a need to obtain neural activity patterns for the training. These states must originate from the imprinting that occurs when the system is together. How are the imprinted patterns retrieved for the training of the subdivisions when the system is off-line? By the operation of the network itself. We call the second step reimprinting or relearning. Simulations that build an understanding of its operation are discussed in Section 3.1.2.

Having arrived at this scenario for training the subdivided network, we ask whether there is an analog of this in the actual system. The answer may be that sleep is the time during which the brain performs the subdivision training. This suggests that we consider the phenomenology of sleep and see if it can be reconciled with the possibility that the brain is separated into subdivisions and undergoes a filtering procedure. Because the training of subdivisions is central to their utilization in the brain, as well as in other complex systems, and the purpose of sleep is one of the great unsolved mysteries, we will consider their relationship in some detail in Section 3.1.3.

More generally, it is quite natural to suggest that complex systems that have identifiable function may undergo processes of dissociation and reconnection as part of their developmental process. This enables the subdivisions to develop autonomous capabilities which may then be recombined to achieve new stages of development.

Question 3.1.1 Show that the number of orthogonal states that can be stored in an attractor network is N. This is larger than the number of random states αN derived in Section 2.2.

Solution 3.1.1 The orthogonality of different states may be written as:

$$\sum_{j=1}^{N} \xi_j^{\mu} \xi_j^{\nu} = N\delta_{\mu,\nu} \tag{3.1.1}$$

Using a signal-to-noise analysis as in Section 2.2.5 we can evaluate the stability of a particular neuron $\{s_i | s_i = \xi_i^1\}$ when the states ξ_i^{μ} are orthogonal. We arrive directly at Eq. (2.2.21):

$$\xi_1^1 h_1 = \frac{1}{N}\sum_{j=2}^{N} \xi_1^1 \xi_1^1 \xi_j^1 \xi_j^1 + \frac{1}{N}\sum_{j=2}^{N}\sum_{\mu=2}^{p} \xi_1^1 \xi_1^\mu \xi_j^\mu \xi_j^1 \tag{3.1.2}$$

The first term is the signal term and,as before,is just $(N-1)/N$. The second term, which was the noise term in the previous analysis, is essentially given by the overlap of different states, which is zero in this case by Eq. (3.1.1). However, we must take into consideration that the sum does not include $j = 1$. So we have a correction to the signal

$$\xi_1^1 h_1 = \frac{N-1}{N} - \frac{1}{N}\sum_{\mu=2}^{p} \xi_1^1 \xi_1^\mu \xi_1^\mu \xi_1^1 = \frac{(N-1)}{N} - \frac{(p-1)}{N} = \frac{(N-p)}{N} \tag{3.1.3}$$

We see that the result vanishes when p reaches N. It is impossible to imprint more than N orthogonal states because there are no more than N orthogonal states of N variables. However, this analysis also shows that the basins of attraction vanish in the limit of $p = N$. ▮

Question 3.1.2 Show that after imprinting N orthogonal states, all possible neural states have the same energy in the energy analog.

Solution 3.1.2 N orthogonal states of dimension N are a complete set of states. This can be written as:

$$\frac{1}{N}\sum_{\mu=1}^{N} \xi_i^\mu \xi_j^\mu = \delta_{i,j} \tag{3.1.4}$$

Except for the diagonal terms, this is the same as the synapses of the imprinted neural network. Since all the diagonal terms of the synapses are set to zero, and the off diagonal terms are zero by Eq. (3.1.4), we must have a completely null set of synapses. ▮

3.1.2 *Recovery and reimprinting of memories*

In this section we demonstrate the use of a neural network to recover memories and reimprint them. The reimprinting reinforces some memories at the expense of others.Effectively, the number of memories stored in the network is reduced so that further imprinting does not cause overload. The retrieval process begins from a random initial state which is evolved to a stable state. Using a random initial state means that the retrieval emphasizes the memories with the largest basins of attraction. These memories have been imprinted with the largest weight or are most different (most orthogonal) from other states that have been imprinted. In effect,the process is a filtering of memories that retains the "most important" ones and the ones that provide for best utilization of the storage capacity of the network. There are two advantages of this selection procedure over palimpsest memories discussed in Section 2.2.7. First, selection may be done after the original imprinting of the memories instead of during the imprinting. Second, the selection does not solely rely upon the specific order of imprints, it enables the persistence of particular older memories. However, when imprinting and selection are repeated many times, it is still true that recent imprints are more likely to survive the filtering process.

The reimprinting procedure is a particular way of filtering memories so that overload in the network is prevented and learning can continue. The continued learning is assumed to serve both as continued adaptation to a changing environment and as a refinement of the storage due to selection of more optimal memories—memories that are more likely to be recalled because they appear more frequently in the environment.

Reimprinting is well suited to subdivided neural architectures where each subdivision is expected to have well-defined memories that serve as the building blocks for the complete neural states of the network. In a subdivided network, reimprinting is implemented during a temporary dissociation of the network, which may correspond to sleep (Section 3.1.3). Temporary dissociation is achieved by diminishing the synaptic weights that connect between subdivisions. During the temporary dissociation, reimprinting optimizes the storage of patterns within each subdivision without regard to the associations established by inter-subnetwork synapses. When the inter-subnetwork synapses are reestablished, these associations are also reestablished.

In order to understand how the reimprinting works, we describe simulations that build the process step by step. The reimprinting of memories occurs on top of the imprinting that has already been performed. We must first understand the effect of imprinting states on top of existing memories using a palimpsest approach that weakens the previous memories before imprinting new ones. We can anticipate the results in two limits. The limit of no reduction in synapse strength corresponds to the conventional training. On the other hand, if the synapse strengths are sufficiently reduced in strength, then imprinting new patterns must be equivalent to the case of no prior imprints. Intermediate cases enable us to develop an understanding of the extent to which prior memories affect new imprints, and conversely how new imprints affect prior memories.

Assuming Hebbian imprinting, the combination of reducing the prior synapse values and imprinting a new state is described by modifying the synapses according to:

$$J_{ij}(t) = \chi J_{ij}(t-1) + \frac{1}{N} s_i(t) s_j(t) \qquad\qquad i \neq j \qquad\qquad (3.1.5)$$

Instead of diminishing the synapses with each imprint, we consider an episodic approach. We imprint a prespecified number of states and then reduce the synapses before continuing. For the simulations we use a network of $N = 100$ neurons, and imprint $p_1 = 4$ neural states with equal coefficients. Then the synapse strengths are multiplied by a factor χ. Then p_1 more neural states are imprinted and the synapse values are again multiplied by χ. This is repeated until a total of $p = 16$ neural states are imprinted. The final value of the synapses could be written using the expression:

$$J_{ij} = \chi^{p/p_1 - 1} \sum_{\nu=1}^{p_1} \xi_i^\nu \xi_j^\nu + \chi^{p/p_1 - 2} \sum_{\nu=p_1+1}^{2p_1} \xi_i^\nu \xi_j^\nu + \ldots + \sum_{\nu=p-p_1+1}^{p} \xi_i^\nu \xi_j^\nu \qquad (3.1.6)$$

To analyze the effect of the procedure, we calculate the basin of attraction of the imprinted states. Results averaged over many examples are plotted in Fig. 3.1.1.Each curve shows the distribution of basins of attraction for the imprinted states. The curves are labeled by the factor χ, which scales the synapses between each set of 4 imprints. The results show that for 16 imprints of equal strength (χ = 1) the network is well beyond the optimal number of imprints. Only 10 imprints are stable and the basins of attraction are very small. When we diminish the synapses between successive imprints, χ < 1, then the basin of attraction of the last four imprints are much larger. However, this occurs at the expense of reducing dramatically the basins of the other memories, eventually destabilizing them. Two conclusions from this analysis are: (1) diminishing the synapses is an effective mechanism for ensuring that successive imprints are learned effectively, and (2) the older memories are lost.

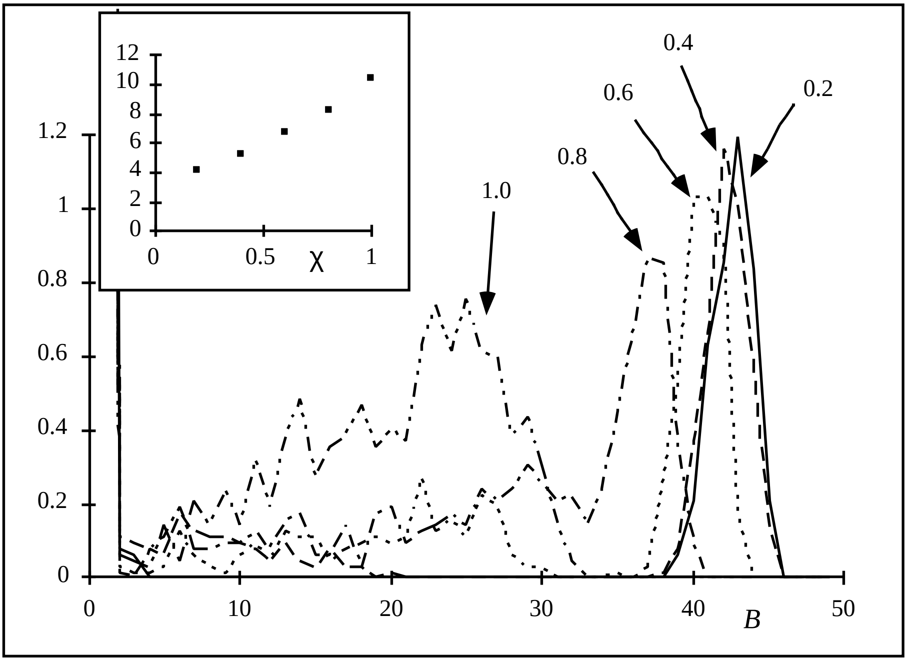


Figure 3.1.1 Simulations of an episodic palimpsest memory using a network of $N = 100$ neurons. Each episode consists of reducing the synaptic strengths by a factor χ then imprinting four new states. Plotted are the resulting histograms of basins of attraction (B). Curves are labeled by the value of χ. The results illustrated have been averaged over many examples. Unstable states are included as having basins of attraction of zero. The curves are normalized to the total number of imprints $p = 16$. By diminishing the strength of synapses (χ < 1) the more recent imprints are better remembered, at the expense of forgetting the earlier sets of memories. The total number of stable states as a function of χ is shown in the inset on the upper left. ▮

The procedure used to generate Fig. 3.1.1 is:

1. Generate p random neural states $\{\xi_i^\mu\}$

$$\xi_i^\mu = \pm 1 \qquad \mu = \{1,\dots,p\},\ i = \{1,\dots,N\} \tag{3.1.7}$$

2. Imprint the first p_1 neural states on the synapses of the neural network (Hebbian rule)

$$J_{ij} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{\mu=1}^{p_1} \xi_i^\mu \xi_j^\mu & i \neq j \\ 0 & i = j \end{cases} \tag{3.1.8}$$

3. Rescale the network synapses by a factor χ.

$$J_{ij} \leftarrow \chi J_{ij} \qquad i,j = \{1,\dots,N\} \tag{3.1.9}$$

4. Repeat steps (2) and (3) for each successive set of p_1 neural states until all p neural states are imprinted.
5. Find the basin of attraction of each of the neural states ξ_i^μ, where an unstable neural state is set to have a basin of attraction of zero (see Section 2.2.6).
6. Make a histogram of the basins of attraction for different ξ_i^μ.

Thus far we have demonstrated the performance of an episodic palimpsest memory. As mentioned above, we prefer to retain selected older memories. They can be retained if we reinforce them before the imprinting continues. We will consider several models of reimprinting.

If we know the states that were imprinted, we can select some of the older memories and reimprint them. This is not a practical approach if the earlier imprints are not known at the time of reimprinting, as when training a subdivided network. However, because it can help us understand the optimal effect of reimprinting, this will be the first case studied below.

If we do not independently know the imprinted states,then we must use the network itself to recover patterns that were imprinted. In this case selective reinforcement of memories may be achieved using the following steps: (1) initialize the network to a random state, (2) update the network a prespecified number of times, and (3) imprint the network. The neural update rule we have discussed in the text is noiseless. Because of the spurious states it is advantageous to add a small amount of noise. The noise helps the network escape from shallow minima associated with the spurious states.Glauber dynamics is just such a noisy neural update rule that was introduced in Section 1.6 and discussed in the context of neural networks in Question 2.2.2. Thus we will compare reimprinting in two cases, using a noiseless update rule and using a noisy update rule.

It is important to perform the reimprinting well before the network reaches overload. As we approach overload, the memory becomes ineffective. This is not only a problem for the network operation, it is also a problem for the filtering process that uses the network to retrieve patterns for reimprinting. If the network is too near overload, the retrieval process will find more spurious states than imprinted states. However, we would like to be able to use a significant fraction of the network capacity. In the simulations below, we balance these considerations by using an initial imprinting with 8 states on a network of 100 neurons. Eight states is below, but a significant fraction of, the maximal capacity of about twelve states.

Figs.3.1.2 and 3.1.3 compare three different reimprinting procedures. Fig. 3.1.2 shows the distribution of basins of attraction.Fig. 3.1.3 shows the integrated number of basins of attraction higher than the value along the abscissa. The starting point, before reimprinting, consists of a network with 8 imprinted neural states.This is shown

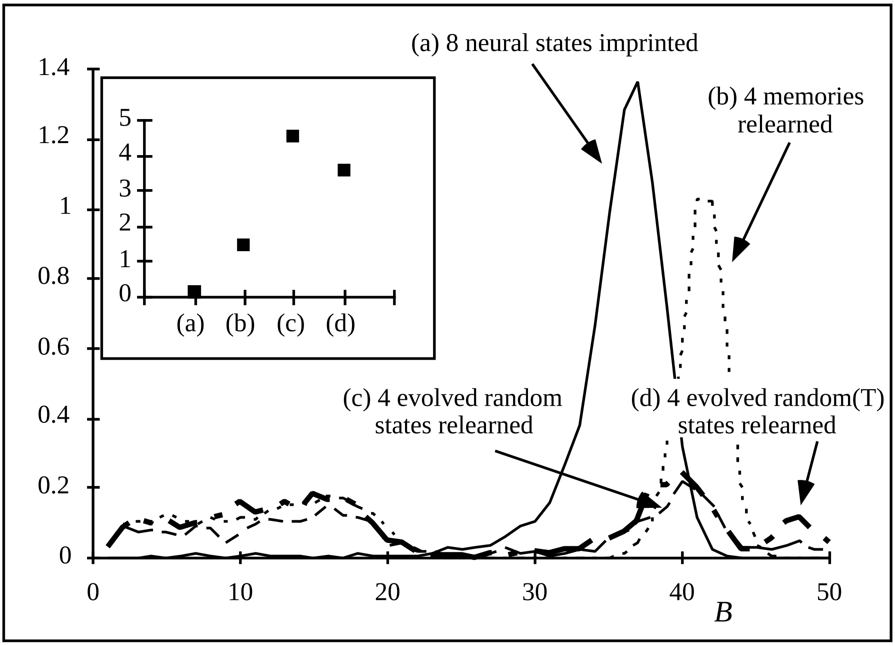


Figure 3.1.2 Plots illustrating tests of reimprinting procedures. Histograms of the basins of attraction are shown. The objective is to strengthen several memories at the expense of the others. The starting point is a network of 100 neurons with 8 imprinted neural states. This is shown as curve (a). Curves (b),(c) and (d) show the results after reimprinting. (b) is the optimal case where four of the originally imprinted states are imprinted again. In curve (c) the 4 imprinted states are evolved random states obtained by applying the deterministic neural update rule to a random initial configuration. Curve (d) is the same as (c) except the neuron evolution includes noise ($\beta = 3.5$) (see Question 2.2.2). In both cases there is an enhancement of some of the basins of attraction. For curve (c) the number of basins enhanced is about 1.5 while for curve (d) it is about 2. Note that the basins of attraction of all of the evolved random states are not necessarily included in the figure since they are not always desired memories. The insert on the upper left shows the number of unstable memories out of the eight originally imprinted states. ∎

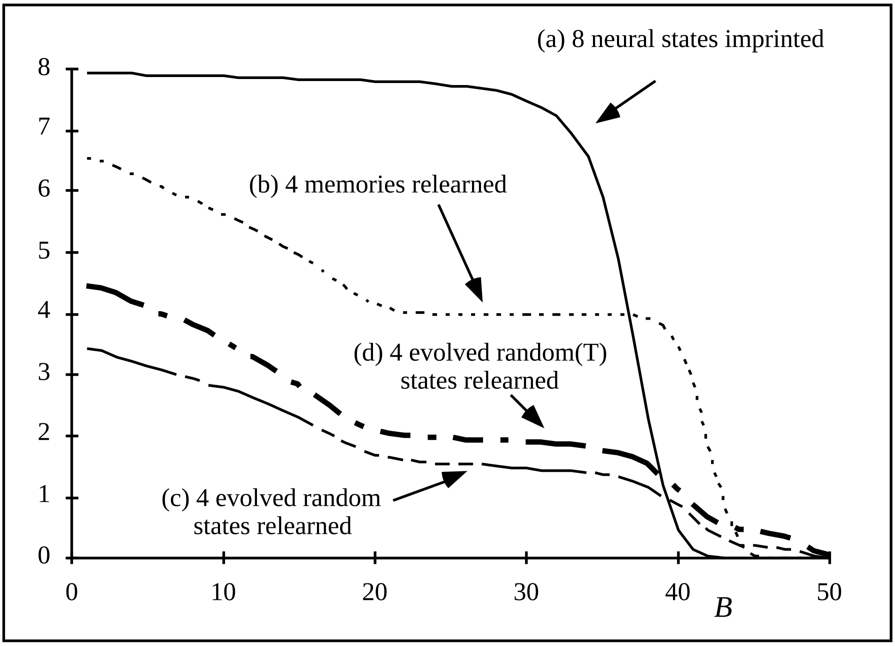


Figure 3.1.3 Plots illustrating tests of reimprinting procedures. This figure is similar to Fig. 3.1.2, except that the plots show the integrated number of imprinted states with a basin of attraction grater than a given value *B*. Flat regions of the curves separate states with large basins of attraction from states with small basins of attraction. This shows the ability of the reimprinting to reinforce some memories at the expense of others. ∎

as curve (a). The objective is to strengthen several memories at the expense of others. Curves (b), (c), and (d) show the results after reimprinting. (b) is the idealized case, where four of the originally imprinted states are imprinted again. This is the same as imprinting four of the states with twice the strength of the other four. As in Fig. 3.1.1, the effect is to reduce the basin of attraction of the neural states that are not reimprinted,and increase the basin of attraction of the ones that are reimprinted. In curve (c) the four states to be imprinted are obtained by applying the deterministic neural update rule to a random initial state. We call these "evolved random" states. In curve (d) the four states to be imprinted are obtained by applying the noisy, or nonzero temperature,neural update rule to a random initial state. The temperature $kT = 0.285$ or $\beta = 3.5$ was chosen based on simulations of the recovery of imprinted states at different temperatures (Question 2.2.2). Whenever the nonzero temperature update rule is used, we also evolve the network by the zero temperature rule to bring the neural state to its local minimum. In both (c) and (d) there is an enhancement of some of the basins of attraction. For curve (c) the number of memories that are enhanced is about 1.5, while for curve (d) it is about 2. The basins of attraction of memories that were reimprinted appear as a peak around the value 40 in Fig. 3.1.2. There is also a small probability that a memory will be reimprinted twice. Such memories appear in a small peak near 50.

The results of the simulations demonstrate that the imprinting of evolved random states enables selective reinforcement of prior imprints. Curve (d) is an improvement over curve (c) because more of the original states are reimprinted. This is explained by the improved retrieval of imprinted states by the noisy evolution.

The procedure for generating Fig. 3.1.2 and Fig. 3.1.3 is:

1. Generate $p = 8$ random neural states $\{\xi_i^\mu\}$

$$\xi_i^\mu = \pm 1 \qquad \mu = \{1,...,p\},\ i = \{1,...,N\} \tag{3.1.10}$$

2. Imprint the states on the synapses of the neural network:

$$J_{ij} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu & i \neq j \\ 0 & i = j \end{cases} \tag{3.1.11}$$

3. Execute the branching instruction:
   For (a): proceed directly to step 7.

   For (b): imprint again the first $p_1 = 4$ neural states on the synapses of the neural network:

$$J_{ij} \leftarrow J_{ij} + \frac{1}{N}\sum_{\mu=1}^{p_1} \xi_i^\mu \xi_j^\mu \qquad i \neq j \tag{3.1.12}$$

   Then proceed to step 7.

   For (c) or (d): proceed:

4. Generate $p_1 = 4$ random neural states $\{w_i^\nu\}$

$$w_i^\nu = \pm 1 \qquad \nu = \{1,...,p_2\},\ i = \{1,...,N\} \tag{3.1.13}$$

5. Execute the branching instruction:

   For (c): update the neural states $\{w_i^\nu\}$ according to the neural update rule 10 times

$$w_i^\nu \leftarrow \mathrm{sign}\left(\sum_j J_{ij} w_j^\nu\right) \tag{3.1.14}$$

   Proceed to step 6.

   For (d): (see Question 2.2.2) update the neural states $\{w_i^\nu\}$ according to the $T \neq 0$ neural update rule 20 times

$$w_i^\nu \leftarrow \mathrm{sign}_T\left(\sum_j J_{ij} w_j^\nu\right) \tag{3.1.15}$$

then update the neural states $\{w_i^\nu\}$ according to the $T = 0$ neural update rule 10 times.

$$w_i^\nu \leftarrow \mathrm{sign}\left(\sum_j J_{ij} w_j^\nu \right) \qquad (3.1.16)$$

6. Imprint $p_1 = 4$ evolved-random neural states $\{w_i^\nu\}$ on the synapses of the neural network:

$$J_{ij} \leftarrow J_{ij} + \frac{1}{N}\sum_{\nu=1}^{p_1} w_i^\nu w_j^\nu \qquad i \neq j \qquad (3.1.17)$$

7. Find the basin of attraction of each of the originally imprinted neural states ξ^μ, where an unstable neural state is set to have a basin of attraction of zero (see Section 2.2.6).
8. Make a histogram of the basins of attraction for different ξ^μ. For Fig. 3.1.3 integrate the histogram from a specified value up to 100.

The previous simulations show the use of imprinting evolved random states to reinforce some of the memories. The next simulation takes the procedure one step further to demonstrate the effect of subsequent imprinting. The simulations consist of four main steps. The first step consists of imprinting eight neural states. The second step consists of selecting four random states, evolving them at a temperature T, then evolving them at $T = 0$ (to bring them to the local minimum) and imprinting the result on the network. The third step consists of diminishing the strength of the synapses by a factor of 2. This ensures that the effective imprinting strength of reimprinted states (which have been imprinted twice) is comparable with that of new states to be imprinted. The fourth step consists of imprinting four new states on the network. The purpose is to demonstrate the ability of the network to continue learning.

The consequences of the full procedure are shown as curve (c) of Fig. 3.1.4 and Fig. 3.1.5. It is the result of imprinting eight memories,then imprinting four evolved random states,then diminishing the strength of the synapses by a factor of 2,and then imprinting four additional memories. In Fig. 3.1.4 the distribution of basins of attraction is shown normalized to 12. Fig. 3.1.5 shows the integrated number of memories with basins of attraction greater than the value along the abscissa.

Two reference curves are included as curves (a) and (b). Curve (a) is the result of imprinting 12 memories on the network. The degree of degradation of the basins of attraction when 12 memories are imprinted is easily seen. This is essentially the maximum capacity of the network, the effectiveness of the memory of these patterns is minimal. Curve (b) is the result of imprinting 8 memories, then diminishing the strength of the synapses by a factor of 2,and then imprinting four additional memories. The total number of imprints is still 12. However, as in the simulations of Fig. 3.1.1, the recent set of 4 imprints have large basins of attraction, while the initial set of 8 imprints are effectively lost, since their basins of attraction have been completely degraded.

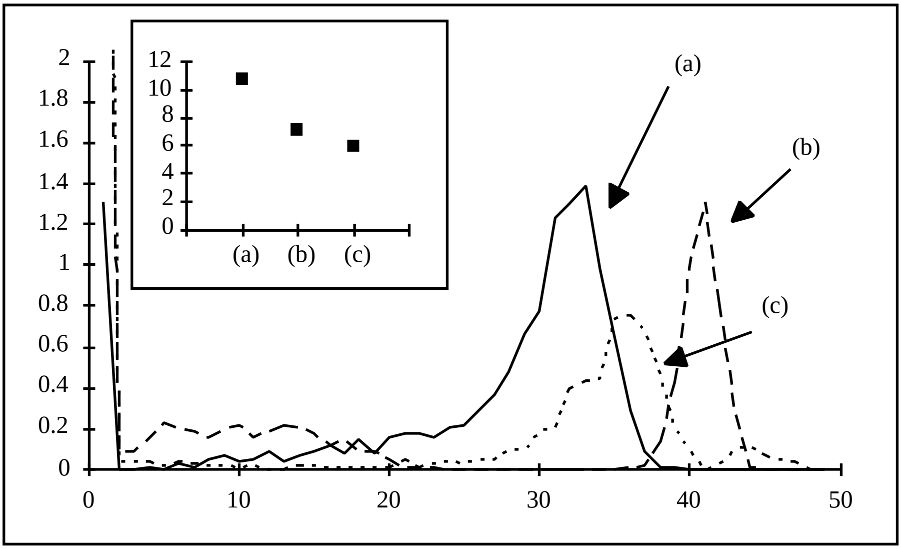


Figure 3.1.4 Continuation of the reimprinting test by imprinting some new states. Curves (a) and (b) are for reference. (a) shows the degree of degradation of the basins of attraction when 12 memories are imprinted all with the same weight. (b) shows the effect of imprinting 8 states, then reducing the strength of the synapses by a factor of 2, then imprinting 4 new states. As explained in Fig. 3.1.1 this results in effective recall of the recently imprinted neural states at the expense of the previously imprinted neural states. The difference between curves (b) and (c) is the insertion of the procedure of evolving four random states with noise, and imprinting the result. This relearning procedure results in the retention of two of the eight original memories for a total of six memories. The others are completely forgotten. ∎

The difference between the simulations leading to curve (b) and curve (c) is only the inclusion of the reimprinting procedure in curve (c). Curve (c) shows that the reimprinting was successful in isolating two of the original memories to retain. These memories survive the imprinting of new states and join them to form a total of approximately six stable memories. This is easiest to see in Fig. 3.1.5, where the relatively flat part of the curve extends down to zero and intercepts the axis at 6 states. Even though curve (b) has more stable states (insert in Fig. 3.1.4), all of these except the newly imprinted ones have very small basins of attraction.

The procedure used to generate Fig. 3.1.4 and Fig. 3.1.5 is:

1. Generate $p_1 = 8$ random neural states $\{\xi_i^\mu\}$

$$\xi_i^\mu = \pm 1 \qquad \mu = \{1,\ldots,p\}, i = \{1,\ldots,N\} \tag{3.1.18}$$

2. Imprint the neural states on the synapses of the neural network

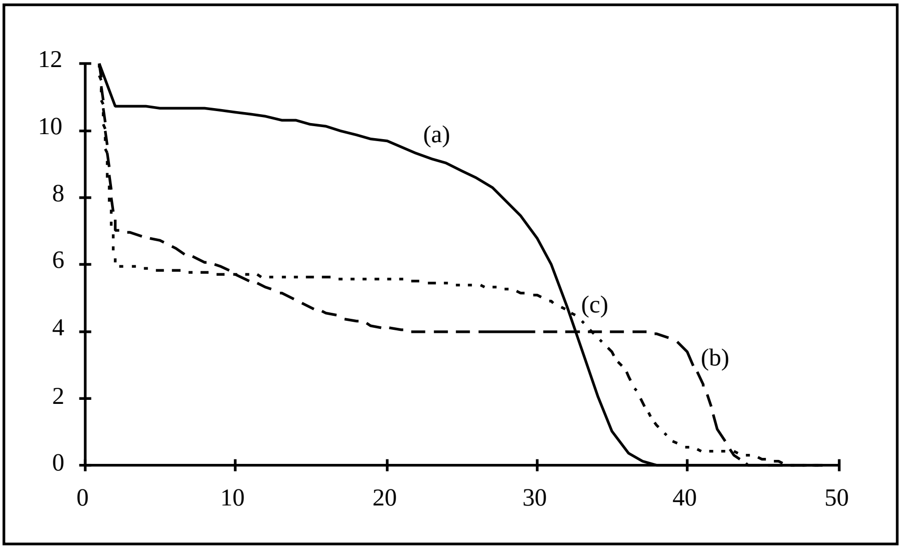


Figure 3.1.5 Similar to Fig. 3.1.4, except that the plots show the integrated number of imprinted states with a basin of attraction greater than a given value. Flat regions separate states with high basins of attraction from states with low basins of attraction. This shows the ability of the reimprinting to reinforce some memories at the expense of others. ∎

$$J_{ij} = \begin{cases} \dfrac{1}{N} \sum\limits_{\mu=1}^{p_1} \xi_i^\mu \xi_j^\mu & i \neq j \\ 0 & i = j \end{cases} \tag{3.1.19}$$

3. Execute the branching instruction:

   For (a): proceed directly to step 5.

   For (b): proceed directly to step 4.

   For (c):

   c1. Generate $p_2 = 4$ random neural states $\{w_i^\nu\}$:

$$w_i^\nu = \pm 1 \qquad \nu = \{1,\ldots,p_2\},\ i = \{1,\ldots,N\} \tag{3.1.20}$$

   c2. Update the neural states $\{w_i^\nu\}$ according to the $T \neq 0$ neural update rule 20 times

$$w_i^\nu \leftarrow \mathrm{sign}_T \left(\sum_j J_{ij} w_j^\nu \right) \tag{3.1.21}$$

   c3. Update the neural states $\{w_i^\nu\}$ according to the $T = 0$ neural update rule 5 times.

$$w_i^{\nu} \leftarrow \mathrm{sign}\left(\sum_j J_{ij} w_j^{\nu}\right) \tag{3.1.22}$$

c4. Imprint $p_1 = 4$ evolved-random neural states $\{w_i^{\nu}\}$ on the synapses of the neural network.

$$J_{ij} \leftarrow J_{ij} + \begin{cases} \frac{1}{N}\sum_{\nu=1}^{p_1} w_i^{\nu} w_j^{\nu} & i \neq j \\ 0 & i = j \end{cases} \tag{3.1.23}$$

4. Rescale the network synapses by a factor $\chi = 1/2$.

$$J_{ij} \leftarrow \chi J_{ij} \qquad i,j = \{1,\ldots,N\} \tag{3.1.24}$$

5. Generate p_2 additional random neural states $\{\xi_i^{\mu}\}$

$$\xi_i^{\mu} = \pm 1 \qquad \mu = \{p_1 + 1,\ldots,p_1 + p_2\},\ i = \{1,\ldots,N\} \tag{3.1.25}$$

6. Imprint the neural states on the synapses of the neural network (Hebbian rule)

$$J_{ij} \leftarrow J_{ij} + \frac{1}{N}\sum_{\mu=p_1+1}^{p_1+p_2} \xi_i^{\mu}\xi_j^{\mu} \qquad i \neq j \tag{3.1.26}$$

7. Find the basin of attraction of each of the $p = p_1 + p_2$ neural states ξ^{μ} (see Section 2.2.6).

8. Make a histogram of the basins of attraction for different ξ^{μ}. For Fig. 3.1.5,integrate the histogram from a specified value up to 100.

3.1.3 *Sleep phenomenology and theory*

Sleep is one of the fundamental phenomena in biological organisms. An excellent review of the phenomenology of sleep and speculations about its nature are given in the book *Why We Sleep* by Horne. The analysis of subdivisions in complex systems in Section 3.1.1 offers an interesting but speculative theory for the role of sleep—that sleep constitutes dissociation with relearning. This theory is consistent with the suggestion advanced in recent years that dreams have a role in "memory consolidation." However, it extends this role to all of sleep. It also provides a constructive framework in which we can discuss the meaning of memory consolidation. In this section we will provide a brief overview of the phenomenology of sleep, challenge two traditional theories for its role and discuss a few modern theories for the role of dreams based upon neural networks. In Section 3.1.4 we compare the theory of sleep as dissociation with the phenomenology of sleep and suggest experiments that can directly test it. While the theory is not directly supported by current experimental evidence,it is consistent with existing results and is an example of a "good theory" because it predicts definite outcomes for novel experiments and would significantly increase our understanding if found to be correct.Finally, in Section 3.1.5, we discuss a new experimental result which provides some support for the predictions.

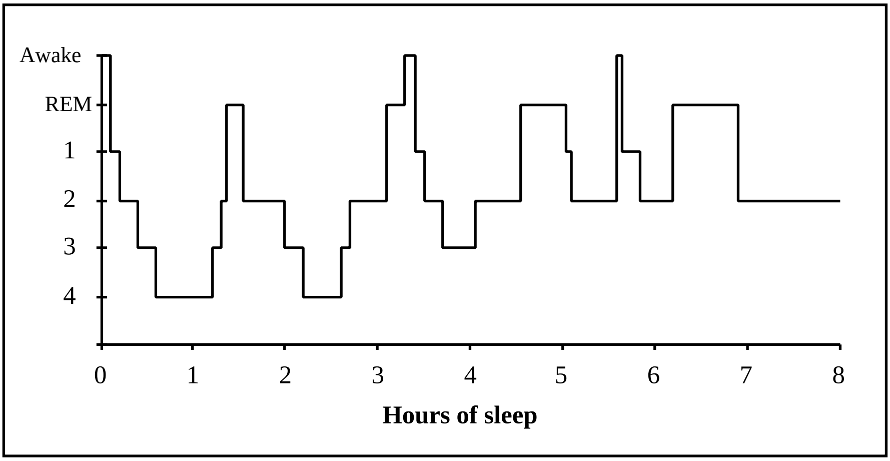


Figure 3.1.6 A sleep "hypnogram" — schematic illustration of the structure of sleep for young human adults showing the different stages as determined by EEG signals. Stages 3 and 4 are called slow wave or SWS sleep. ∎

Human beings spend almost one-third of their lifetimes asleep. Human sleep is known to have several levels identified by quite different brain electrical signals as measured by electroencephalography (EEG). There are at least five recognized levels, the two deepest are together called slow wave sleep (SWS), while the shallowest is rapid eye movement (REM) sleep. Typically, in the first part of sleep, the deepest level is attained rapidly. Then the level of sleep alternates in a pattern of shallower and deeper levels with the average level becoming shallower as sleep progresses (see Fig. 3.1.6). The sleep of animals does not have as many levels. The complexity of sleep increases with the (conventional) evolutionary complexity of the organism.

There are two conflicting popular views of sleep. One is that sleep is necessary for health and well-being. The other is that sleep is a waste of time. Modern society often pays little attention to the significance of sleep. For example, doctors,especially during training, work very extended shifts. Other professions either work long shifts or ignore the natural sleep cycle of day and night. Evidence has accumulated that such practices are counterproductive and cause errors, even fatal errors. There exist efforts to change the training of doctors, and to avoid excessive sleep disruption of airplane pilots. In addition,there are many sleep clinics that are designed to help individuals who suffer from sleep disorders,including various forms of insomnia, an inability to sleep.

Much of our understanding of the role of sleep arises from human sleep-deprivation studies. These studies reveal that sleep loss results in psychofunctional degradation. However, the precise nature of the degradation is not well-understood. Some of the effects of sleep deprivation over several nights include visual illusions or mild hallucinations, and loss of a proper time sense. There are also particular tasks that have been shown to be sensitive to the loss of sleep. However, many others do not appear to be systematically affected. An example of a test that is quite sensitive to the loss of sleep is the vigilance test. In this test a person is asked to monitor a sequence of pictures or sounds for a particular feature that should trigger a response. After sleep deprivation, individuals frequently do not respond to the special feature when it is presented.

Modern theories of sleep have suggested either that it serves a physiological restorative function or that it exists because of genetic adaptation to a survival advantage in removing primitive man from danger. Extensive experimental measurements directed at unveiling the physiological restorative function have not been successful. It appears that after exertion, physical rest rather than sleep is sufficient for reconstruction of tissue damaged during use.

The second suggestion, that sleep serves to remove primitive man from danger, does not coincide with a variety of observations about sleep and its role throughout the animal kingdom. First, sleep consists of a time of reduced awareness of environmental dangers. Even if there were an advantage in inactivity, this lack of activity could be achieved by physical rest rather than the loss of alertness that occurs in sleep. Moreover, even animals that are in constant danger sleep. An extreme example is the case of certain types of dolphins that sleep one half-brain at a time in order to monitor their environment and avoid dangers. Moreover, it is surprising that there are no animals that do not sleep. Nocturnal animals sleep during the day. Predators, whose survival does not depend upon safety from other predators, sleep. Why have no species adapted to the survival advantages of alertness and extra time to find food without sleep? Finally, sleep-deprivation studies on animals show that extended sleep deprivation is fatal. For example, rats die on average after 21 days without sleep. The direct cause of death has not been identified despite substantial efforts.

Neither of the two traditional theories explains the mechanism for psychofunctional degradation after sleep loss. They also do not explain many specific results in sleep-deprivation studies on either humans or animals.

While there has been only limited discussion of the role of sleep in human psychofunction, dreams have evoked more elaborate speculations. Many believe that dreams, or more specifically rapid eye movement (REM) sleep, are the essence of sleep even though they occupy only about one quarter of sleep. Because of their bizarre content, dreams have always invoked mystery. Various theories have suggested that dreams play an important role in human psychology. More recent theories relate dreams to aspects of human information-processing, usually memory. In particular, they suggest that dreams play a role in the conversion of short- to long-term memory—memory consolidation.

There are two specific proposals for the role of dreams that are based upon neural network models. They are precisely opposite. Crick and Mitchison suggested that dreams cause selective forgetting of undesirable or parasitic neural network states.One piece of evidence for this approach is our inability to remember most dreams. More concrete support for this proposal was gained through simulations of attractor networks. Simulations, similar to those in the previous section, were performed by Hopfield, et al. After imprinting a network,the network was initialized to a random configuration and evolved. Instead of imprinting the resulting state, the state was unimprinted, or imprinted with a small negative coefficient of 0.01. This was found to improve the retrieval of imprinted states.The improvement arises because the states with larger basins of attraction are responsible for the instability of the other imprinted states. Reducing the large basins of attraction by a small amount improves the balance between different states. In contrast,Geszti and Pázmándi suggested that dreams are a form of relearning. Their relearning procedure is the one described in the previous section. Its purpose is to filter the memories to enable continued learning.

Both of these models attributed information-processing tasks to rapid eye movement (REM) sleep, or dream sleep. The other parts of sleep, where dreams are infrequent (non-REM sleep), are still generally believed to have a physiological role. However, as described earlier, total sleep deprivation causes psychofunctional, not physiological, deterioration in humans. The primary effects occur with loss of non-REM sleep.

Based on the discussion of subdivision in neural networks and training it is reasonable to propose that the stages of sleep correspond to degrees of interconnection between subdivisions of the brain. SWS corresponds to the greatest dissociation, where small neuron groups function independently of each other. At shallower levels of sleep, larger regions of the brain are connected. Ultimately, the waking state is fully connected, including sensory and motor neurons. From EEG measurements it is known that all of the levels of sleep are neurologically active. Consistent with our discussion in the last section, it may be proposed that the activity is a filtering process that reinforces some memories at the expense of others to prevent overload and allow for additional learning. The filtering of memories occurs on all levels of organization. The ultimate purpose of this filtering process is to establish the memories within subdivisions,and the stable composite memories. It also balances the strength of synapses within subdivisions compared to the strength of synapses between them.

There are general consequences of the filtering that we can infer and use to make predictions of its effects on memory. It is to be expected that the strength of associations that are represented by synapses between subdivisions are weakened more rapidly than associations that are represented by synapses within subdivisions. Thus, memories are progressively decomposed into their aspects, stored within subdivisions. Implicit in this architecture is the assumption that the most permanent associations—stable patterns of neural activity—are stored inside the smallest subdivisions. These associations are the elements that are the building blocks for new imprints on the network, and thus are the elements for building new memories.

This theory for the role of sleep is based upon a subdivided attractor network, with no directionality to the synapses, or to the processing as a whole. The presence of directionality in the processing of sensory information should modify this picture, but may not change the essential conclusions.One modification that we can expect is that the triggering of a random neural state will also acquire directionality. The triggering should follow the usual processing path in order to be consistent with the system's natural mechanism for retrieving memories.

3.1.4 *Predictions and experimental tests*

A variety of aspects of the general phenomenology of sleep are consistent with the idea that sleep is a temporary dissociation of the brain into its components. Several of these are described in the following paragraphs.

Sleep itself consists of a dissociation of the cerebral activity from both sensory and motor neurons. This separation is accomplished by sleep substances that control particular interconnecting neurons or synapses. While the dissociation is not complete—we can still respond to sounds and lights during sleep—the degree of correlation between sensory stimuli and the activity of the brain is reduced. Similar controls could be used to further dissociate various subdivisions of the brain.

As mentioned before,sleep is a time during which there is significant neural activity. This is to be contrasted with the lack of explicit memory of this activity. The patterns of neural activity differ qualitatively between different stages of sleep. These changes can be measured using EEG signals, which are used to identify stages of sleep. Systematic differences between patterns of neural activity imply basic changes in either the activity of neurons or their synaptic efficiencies.This requires an explanation. Qualitatively, the greater simplicity of EEG signals in SWS (hence the name slow wave sleep) is consistent with a loss of complexity in the activity patterns due to a lack of correlation between different neuron groups.

The internal triggering of patterns of neural activity occurs in all stages of sleep, but is very apparent during REM sleep, where pulsed neural activity patterns extend through a significant part of the cerebral cortex.

The greater difficulty of waking during SWS is consistent with a greater degree of dissociation in deep sleep than in shallower levels of sleep. It may also be difficult to wake from REM sleep, despite its other characteristics as a shallow stage of sleep. However, in this case the internal triggering of neural activity appears to mask awareness of actual sensory stimuli.

Systematic studies of dream content indicate that specific higher-level critical faculties and a "sense of self" are absent. This includes a lack of surprise at the content of dreams, and an inability to see or perceive one's self. It has been pointed out by Hartmann that this is similar to the waking mental functioning of postlobotomy patients, where connections to the frontal lobes of the brain have been severed. Specific higher-level critical facilities related to self-awareness are often associated with these frontal lobes. This suggests that during REM sleep, specific major sections of the brain are dissociated.

Dissociation during sleep would imply that the neural activity is formed out of composite states that typically would not occur if the brain subdivisions were connected. In REM sleep, when only major sections of the brain are separated from each other, the composite states are formed out of only a few elements. The waking brain with full connections can, at least sometimes, make a kind of sense out of the juxtaposition of elements from the sleep state. This explains the possibility of recalling sleep states from REM sleep in the form of dreams. It also explains their bizarre content. Moreover, it explains why dreams are not always recalled even when experimental subjects are woken during REM sleep. For deeper levels of sleep, with smaller subdivisions, the waking brain can generally make no coherent picture of the sleeping mental state. This explains the absence of recalled dreams from deep sleep despite the ongoing neural activity.

To make further progress in our understanding of sleep and the dissociation model, we will discuss the psychofunctional effects of sleep deprivation. Our discussion will provide some understanding of the deterioration that can result from sleep deprivation. It will also explain why experimental efforts have found it difficult to identify specific psychofunctional tasks that are affected. The central point is recognizing that the deterioration of capabilities is directly linked to activities that are performed during waking. Thus the question, How does sleep deprivation affect the capabilities of an individual? is meaningless without a specification of the activities performed by the individual during the period in which he or she is awake.

The model of sleep as dissociation implies that it is basic to the functioning of the subdivided architecture of the brain. However, the manifestation of sleep deprivation would not be the complete disruption of this architecture. The shorter-term effects of sleep deprivation are related to overload failure. Overload occurs because imprinting is continued during waking hours without a periodic filtering process. Under normal circumstances, there must be a substantial buffer before overload is reached. The buffer exists because of the need to stop imprinting well before the overload threshold. However, if the buffer were very large, then the full capacity of the network would not be utilized. This explains the need for a regular sleep schedule with a consistent structure to the levels of sleep. It also explains why there are dramatic effects of only a few nights of sleep deprivation, which become catastrophic if further extended. We note that no model of the role of sleep based solely on a concept of memory consolidation would account for psychofunctional failure due to sleep deprivation.

The implications of overload failure in a fully connected network were discussed in Section 2.2.7. When overload is reached, various spurious states replace memories as the stable states of the network, and a complete loss of memory results. In order to adapt this picture to describe the effect of sleep deprivation, we must include both the correlated nature of the information that is presented to the brain over any particular period of time, and the subdivided architecture. Their implications may be understood simply. First, the correlated information implies that overload does not affect all imprinted states equally. If newly imprinted states are confined to a particular subspace of all possible neural states, then all states that are not correlated with them will not be affected. The existence of subdivisions leads to similar conclusions by emphasizing that overload should occur in particular subdivisions first, rather than uniformly throughout the network. The conclusion is that the effect of sleep deprivation is primarily confined to activities that are exercised during the waking period. This explains much of the difficulties that have arisen in the efforts to determine specific activities that are strongly affected by sleep deprivation. Many tests evaluate the degradation in an activity, such as intelligence tests, tests of ability to maintain balance, etc. However, even when significant correlations are reported between sleep loss and a particular test in one experiment, these are found not to exist under other experimental conditions. In contrast to the generally ambiguous results on specific abilities, it has been reliably shown that a degradation of ability is found for repeated or similar tests, essentially independent of the nature of the task.

Unlike other activities, the vigilance test is a self-contained test of the ability to persist in a particular activity. The vigilance test requires paying attention to a series of varying sensory images, and responding only to a particular variant. We can understand why a sleep-deprived individual finds this difficult. Imprinting various similar states up to overload would cause, in effect,the basins of attraction of the lack of action to overtake the only slightly different circumstances that require action. The inability to respond differently to a slightly different stimulus may very well be the cause of accidents that occur in early-morning hours. Consider the train conductor who is required to brake the train upon seeing a particular set of lights,after viewing many different panoramas with various sets of lights. Consider the doctor, who after seeing many patients with similar ailments is required to change the treatment based on one of many pieces of information.

The relevance of repeated tasks to the need for sleep does not require sleep-deprivation studies. It is sufficient to note the sleepiness that arises from boredom due to monotonous or repetitive activity. It is also interesting that schools (at all levels) do not schedule classes around learning a particular topic for a whole day. Instead, activities and classes vary through the day. Each subject is taught only for a limited period of time.

Sleep-deprivation studies are generally performed in a monotonous environment without many stimulating or novel activities. Stress, which is likely to increase imprinting (see Section 3.2.4),is also absent. This suggests that aside from the generally necessary activities, clinical sleep-deprivation studies do not capture the psychofunctional degradation from typical daily activities. The most commonly observed difficulties in sleep deprivation arise from visual illusions. This may be understood both from the necessity of vision even under laboratory conditions,the popularity of reading or watching TV during an experiment,as well as the monotonous laboratory environment that implies significant correlations between visual stimuli.

Modeling sleep deprivation by overload failure implies that novel, stimulating, stressful, or boring circumstances lead to an increased effect of sleep deprivation. As will be discussed in Section 3.2.4,all of these, except for boring circumstances,can be related to an increase in imprinting strength and a more rapid approach to overload. Boring circumstances, by virtue of repetition, achieve this result rapidly not because of the strength of imprinting but because of the overlap of different imprints that cause overload in a more limited domain of patterns in the network. Consistent with experience, sleepiness that results from repetitive activity can often be overcome by changing the activity.

If we wish to understand the implications of sustained sleep deprivation, we must look for individuals that inherently possess a particular form of sleep deprivation. The simplest to understand would be a loss of deep sleep, where the basic elements of neural functioning in the smallest neural networks are established. A loss of SWS would be associated with a breakdown in psychofunction well beyond a severe case of sleep deprivation. Experimentally, it has been found that there is a complete lack of SWS in about 50% of individuals diagnosed with schizophrenia. Schizophrenia includes a broad class of severe psychofunctional disorders.

We now turn to discuss new experiments using several distinct methodologies that could directly evaluate the possible role of subdivision during sleep. These tests include clinical studies, imaging and physiological experiments.

The most direct clinical tests would measure the retention of memory of associations at higher levels of the hierarchy. According to the model of temporary dissociation,associations between disparate kinds of information stored in different regions of the brain are preferentially lost during sleep. An experimental test would expose subjects to information that is composed of two or more different aspects. The subjects would be split into two groups, one would sleep and the other would not. The retention of the information would then be tested. Various experiments of this kind have been done but without specific emphasis on correlations of different aspects of information. For example,a visual image and a sound could be presented at the same time. A test would measure the ability to recognize which image-sound pairs were presented. Other combinations of information could also be tested by selecting from known subdivisions in the brain: vision,audition,somatosensory, language,and motor control. Within each category further tests could be performed. For example, tests in vision could measure the ability to retain particular combinations of shape and color. Pictures of people, each with particular color clothes, could be changed by reassigning colors. Tests would determine the ability to recall the association of color with shape.

The development of positron emission tomography (PET) and magnetic resonance imaging (MRI) has enabled more detailed mapping of brain activity in recent years. The ability to map brain activity can also enable mapping of correlations between activity in different parts of the brain. This becomes increasingly feasible as the temporal resolution of imaging is improved. Statistical studies of the correlations in neural firing could directly measure the strength of influence between different parts of the brain while a subject is awake, and during various stages of sleep.

Neurophysiological studies of animals characteristically measure the activity of a neuron under particular stimulus. Using more than one probe at a time,the correlations between neural activities in different parts of the brain could be compared in waking and in various sleep states. Such experiments can also stimulate some neurons,and measure the difference in signal transmission between neurons in different parts of the brain in animals that are awake and asleep.

The dissociation model would require a chemical mechanism for preferential inhibition of the synapses or neurons that interconnect various regions of the brain. The ability to chemically separate different regions of the brain can be directly tested by investigating the impact of sleep substances on neurons and synapses. Synapses or neurons that interconnect different regions of the brain would be expected to have a characteristically distinct sensitivity when compared to neurons and synapses within a particular region of the brain. In order to enable a difference between REM sleep and SWS it would also be necessary for there to be differences in the connections between smaller regions and larger regions. The possibility of sleep substances that preferentially isolate a particular level of the brain structure may become apparent from such tests.

3.1.5 *Recent experimental results*

In one of the first multiprobe experiments, Wilson et al. recently investigated correlations between neural activities in the hippocampus. The hippocampus is an area of the brain that is responsible for representation of information about the organism's spatial position,in particular its location with respect to large objects or boundaries. They found that new correlations in neural activity due to changes in the environment were subsequently repeated during sleep.

This experiment supports a number of aspects of neural network models of brain function. Of principal significance, it supports the attractor network model that memories are stored and can be recovered as a pattern of neural activity. It also supports the discussion in this chapter, that they are recovered during sleep. The idea that waking experiences are reflected in dreams is known. However, this is the first indication of the nature of their representation. Moreover, it is interesting (and consistent with the above discussion) that the recovery of patterns of neural activity was not particularly associated with REM sleep, but rather occurred in SWS.

3.2 Brain Function and Models of Mind

3.2.1 *The fundamental questions*

We use phenomena that are associated with neural networks to understand some of the aspects of brain function by our own recognition of their similarities. In the previous chapter, we briefly mentioned the associative memory function of the attractor network that is reminiscent of human association capabilities. We will expand upon this discussion in the following sections to cover a variety of information-processing tasks. As we do so, we will find that we have to expand our model to include additional features. We start with both the attractor network formed of symmetric synapses,and the feedforward network with unidirectional synapses. We use subdivision to clarify some of the basic issues and expand into the realm of higher information-processing tasks.

As our description of information-processing functions progresses, we must allow ourselves to expand the conventional terminology. We use our model neural network as a model of the brain. The functioning of this network is a model of the mind. We can use terminology such as the subconscious mind to describe the part of the neural network/brain whose function we identify with what is commonly understood to be the subconscious. A sentence that contains such terminology can still possess precise mathematical meaning in the context of the neural network architecture. This is similar to the use of words like "energy"and "work," which have different meanings in scientific and popular contexts.

3.2.2 *Association*

In the attractor neural network model of the human mind,the basic learning process is an imprinting of information. The information may, for example, be a visual image. This information is represented as a state of the neural network (pattern of neural activity),and the synapses between neurons are modified so as to store—remember—this information. The mechanism for retrieval is through imposing only part of the same image. The synapses force the rest of the neurons to recreate the stored pattern of activity, and thus their representation of the stored image.

In order to illustrate how this process manifests itself in behavior, we have to consider the nervous system "output" leading to action as also part of the state of the mind. Part of the pattern of neural activity specifies (controls) the muscles,and therefore behavior. Using the pattern of activity that represents both sensory information and motor control we can,in a simple way, understand how reactions to the environment are learned.

We can consider the example of a child who learns to say the name "Ma" whenever she sees her mother. Let us say that somehow (by smiling, for example) we are able to trigger imprinting. At some time, by pure coincidence, at the sight of her mother the child says something which sounds like Ma (or even quite different at first, subject to later refinement) and we encourage an imprint by smiling. Thereafter the child will say Ma whenever she sees her mother. The pattern of neural activity that arises when the mother is in the visual field has been associated with the pattern of neural activity representing motor control that manifests itself in the word "Ma." Of course this process could be enhanced by all kinds of additions, but this is one essential process for human learning and human functioning that this neural network captures.

We note that the training of a feedforward network discussed in Section 2.3 requires a comparison between the desired output and the output generated by the network. Because both the desired output and the output generated by the network must be represented at the same time,the feedforward network does not by itself provide a model of how responses can be learned.A solution to this problem will appear when we discuss consciousness in Section 3.2.12.

3.2.3 *Objects, pattern recognition and classification*

When we look at a room we do not interpret the image in the form of a mapping of the visual field as a point by point (pixel by pixel) entity. Our interpretation is based on the existence of objects and object relationships that exist in the visual field. The same is true of auditory information, where sounds,notes or auditory representations of words are the entities we differentiate. Similarly, our associations are driven not by direct overlap of sensory information but rather by objects, aspects or relationships. Why is this useful, and how is it possible to identify objects in sensory fields?

The reason objects are used rather than the visual field itself is easy to understand within the neural network model. Consider a particular visual image which is mapped pixel by pixel onto a neural network and imprinted. An attractor network relies upon the Hamming distance of a new image with the imprinted image for recall and therefore for association. Any new image mapped onto the network is characterized by the overlap (similarity measured by direct counting of the number of equivalent pixels) of the image with imprinted images.

This means, for example,that if we want the child of the last section to say “Ma” no matter how her mother appears in the visual field, then all possible ways the mother can appear in the visual field must be imprinted independently. “All possible” means essentially independent ways,ways for which the overlap of one with the other is small. This overlap is strictly a Hamming distance overlap—a count of the number of equal pixels. Since there are many ways that the mother can appear in the visual field with only a small overlap between them, this would require a large part of the neural memory. Saying that we identify objects is the same as saying that the child identifies as similar many of the possible realizations of the visual field that contain her mother. We must then ask how this is possible when the visual fields compared pixel by pixel are different.

Underlying the use of objects in describing the visual field is the assumption that objects possess attributes that are unchanged by their different possible presentations in the visual field. The existence of attributes,as discussed in Section 2.4,can be used by a subdivided network to identify the objects. We identify the attributes of a particular object as the states of each of the subnetworks when we are presented with the object. For example,in the separation of visual information into shape, color and motion,the attribute RED would be represented by a particular pattern of neural activity in the subnetwork representing color information. Extracting different aspects of the information and storing them in particular subdivisions of the network enables the object to be identified by a particular set of subnetwork states—by the pattern of common attributes. The suggestion that attributes can provide a mechanism for the identification of objects is not a complete answer to the problem of object identification. It is still necessary to examine how the characteristic attributes of objects are found in the visual field.

In recent years the field of computational vision has been dominated in large part by discussion of computational problems associated, for example, with extracting boundaries of objects. This is important because the extraction of edges provides an important clue as to the existence and nature of objects. This research has been viewed as opposed to the use of attributes for object identification. It may be better understood as providing the computational approach to extracting these attributes. Thus, the extraction of edges provides one (or several) attributes of the visual field that can be used to identify objects; other attributes can be used as well. Rather than relying upon a single algorithm to identify objects,the use of multiple attributes enables several algorithms to act together through associative links, as suggested by Fig. 2.4.4.

Once we have understood the identification of objects through their attributes, we can likewise understand pattern recognition or classification as a process of identifying common attributes. Specifically, elements of a category may be identified by the common state of a particular subnetwork or set of subnetworks. Pattern recognition, viewed in an abstract form, is equivalent to the problem of classification.

The existence of objects is often considered one of the most concrete aspects of reality. We see that the identification of objects is actually an abstraction. It is a basic abstraction central to our ability to interpret sensory information. Moreover, the same methodology of abstraction is also the key to understanding abstract concepts. Abstract concepts,like concrete objects, may be stored in the brain by combinations of aspects or attributes. The attributes are represented by patterns of neural activities of brain subdivisions. This method of representation is also related to other aspects of higher information-processing—generalization,creativity and innovation, to be discussed below.

In summary, because of the many different possible visual fields, it is impossible for the brain to be imprinted with all of the appropriate ones,and associate them with the appropriate response. Instead, the visual fields are reinterpreted as composed of combinations of attributes that reflect the existence of objects and their relationships.

3.2.4 *Emotions and imprinting*

One of the central properties of the neural network that we have not investigated in any detail is the strength of imprinting. Our numerical modeling of the imprinting process generally assumed that each imprint has the same coefficient. However, it is quite reasonable to include the possibility of stronger and weaker imprints, where the strength of imprinting can be controlled in various ways.Stronger imprints result in larger basins of attraction.Larger basins of attraction imply that recall is easier—triggered by a smaller set of common attributes. Even in our discussion of association in Section 3.2.2 it was necessary to invoke a mechanism for triggering stronger imprinting in order to describe the learning of a response.

There are various ways to control the strength of imprinting at a particular synapse.Our concern here is not with an individual synapse,but rather with the overall strength of a particular imprint. In the Hebbian imprinting model,this strength is controlled by the parameter c in Eq.(2.2.6).The control of the strength of imprinting must occur at every synapse in the brain. Chemicals that can be distributed throughout the brain to affect the imprinting would be most easily distributed through the bloodstream. The most natural assumption is that the relevant chemicals are associated with emotions.Emotions affect the general response by the body to external circumstances.At least some of these circumstances imply that imprinting and memory should be enhanced. One indication of this is that new circumstances, or circumstances that are important due to the existence of a threat, or circumstances that are painful, give rise to the release of such chemicals. It would make sense that the emotional reaction governs not only the immediate reaction (the traditional fight or flight response to stress) but also the recollection of such circumstances in the future. Without discussing the process in detail we may conclude that imprinting strength under these circumstances—the coefficient c—is increased by adrenaline (epinephrine/norepinephrine) and affected by other endocrine-system chemicals associated with various emotional states.

A second way to strengthen the imprinting is to repeat the same imprint more than once. In the simplest model of a constant imprinting strength,the total strength of imprinting grows linearly with the number of imprints. We will modify this assumption in the next section, because such continued imprinting is not advisable.

If we assume that selective retention and forgetting of memories is necessary, as discussed in Section 3.1,then the relative strength of the original imprinting will also affect which memories persist. We can expect that memories that persist are those associated with the greatest stress or strongest emotions, or with the largest number of repetitions.A classic example is the persistent memory of traumatic events. In a subdivided network,the persistent memories may be aspects of situations rather than the situations themselves.

We have discussed,thus far, the effect of emotional response on the chemistry of the blood and its consequent effect on imprinting. The source of emotional response must also originate in the nervous system, because the sensory stimuli that describes the environmental circumstances leading to the emotional response are received by the nervous system. We must therefore assume that neural activity affects the adrenal gland and other glands responsible for chemicals that affect the physiological response. The circle of influences between bloodstream chemicals and brain function is an important feedback loop. Part of the brain initiates the emotional state by controlling the bloodstream chemicals, which then affect the functioning and the imprinting of the brain. Physiologically, it is the diencephalon and particularly the hypothalamus, a hybrid nervous-system component and endocrine gland, that bridges between the nervous system and the endocrine system.

3.2.5 *Fixation, compulsion and obsession*

Any model of how the brain works contains within it a model for how the brain may fail. Since there are also many real occurrences of failure, we can compare and try to evaluate whether the model is properly predicting the failure. The storage of various information in the brain with different imprinting strengths enables the possibility that a single imprint will become dominant. The meaning of a single dominant imprint was discussed early in Chapter 2 when the case of a single imprint was described. Under these circumstances any initial state will evolve in time to the attractor that is the dominant imprint. This description is very reminiscent of the behavior of a person who suffers from fixation, compulsion or obsession. Such individuals repeat actions or thoughts regardless of the external circumstances and regardless of the recent history.

Examples of dysfunction include a compulsive repetitive action such as hand washing, fixation on a person or object, and obsession with an idea. In each case the persistent return to the behavior pattern or thought pattern can lead to a severe breakdown in human function.Strong imprinting of a particular thought or behavior itself arises from repetition, so that this is a self-consistent failure of the architecture. Self-consistency arises because repetition strengthens imprinting, and when imprinting is strengthened the tendency to repetition is increased. As discussed in the last section, the existence of strong emotions contributes to imprinting, and indeed strong emotions, passion or anxiety, are often associated with the development of these disorders.

When there is a natural mode of failure, one must expect that the system has built-in safeguards against it. When the failure occurs anyway, it is because of a breakdown in the safeguards. Safeguards against the creation of a dominant imprint may consist of a reduction in strength of repeated imprinting. One approach to reducing the strength of repeated imprints uses an internal feedback mechanism that changes the imprint strength depending on whether the neural activity pattern is already stable. This can be done locally at each neuron—when a neuron activity is consistent with the local field h_i, its own synapses need not be imprinted. Some such modification of the prescription of Hebbian imprinting is likely to exist in the brain to avoid excessive imprinting of existing memories. A second safeguard would require a behavior that avoids exposure to repetitions. Boredom as an emotional state may have a purpose of causing behavior that avoids continued repetition. In order for this safeguard to work,there must be an internal mechanism for recognizing repetition—for recognizing the stability of a state. The problem of recognition is discussed in Section 3.2.6.

While such safeguards may serve to help prevent this mechanism for failure, it should also be understood that the strengthening of imprinting with repetition, and the use of emotions, must not be completely curtailed by the safeguards. Otherwise, the primary function of the brain would be degraded. This limits the implementation of safeguards to an extent that enables function, but also enables failure.

Medical classification of mental disorders distinguishes between neurotic and psychotic conditions. The former are less severe than the latter. Various neurotic compulsions,fixations and obsessions may persist for years without treatment. Psychotic conditions can require severe intervention using drugs or electroshock therapy. The action of these treatments is not well-understood. However, we can speculate about electroshock therapy as a means of shaking, in an uncontrolled way, the energy landscape of the space of states of the brain as we have represented it in the neural network model. This may explain why, despite the grave concerns about its side effects, the treatment continues to be used. Interpreted simply this also suggests that for these disorders the forms of traditional psychotherapy that dwell upon the problem may actually promote it. Therapeutic strategies that emphasize other important (strongly imprinted) areas of a person's life, and change as much as possible of the circumstances of a person's life, would be expected to be more effective.

The distinction between different kinds of repetitive processes—fixation associated with the senses, obsession with abstract thoughts, and compulsion with simple actions—suggests that an overly strong imprint may be localized in different regions of the brain. However, because of the coupling between different parts of the brain these distinctions may not always be maintained.

3.2.6 *Recognition*

One of the standard tests of memory is the recognition test. In this test, for example, a person is shown a number of pictures and asked to recognize them later. The human brain is capable of successfully recognizing at least ten thousand pictures. This would seem to be a natural application of our neural network memory that imprints the images and then recalls them. But is it? The process here is different. The subject is required to identify whether or not he has seen the image before. This is different from reconstructing the image itself. In our model of the associative or content-addressable memory, the task is to provide the missing pieces. In the recognition experiment,the subject says "yes" or if he has, or "no" if he has not, recognized the image. Saying "yes" reflects a particular pattern of neural activity exercising vocal control. The recognition task requires that these neurons have the same firing pattern for all stored images, and a different firing pattern for any image that has not previously been stored.Our model does not contain a single association with all of the images that have been stored. Everything that is recovered in the attractor network is part of the original image.

One way to solve this problem is to suggest that there is a part of the brain that stores the word "yes" along with every image we see. Then we can use this part of the brain to perform the recognition task. We will not adopt this approach here. Instead we will require that the network have some way of identifying whether or not it has imprinted the picture from the behavior of the network itself.

When we impose upon the network a previously imprinted pattern of neural activity, the state of the network is a fixed point of the neural evolution—a stable state. We must find a way for the brain to know—i.e., represent the information—that it is in a fixed point,so that it can say "yes", and then when it is not in a fixed point,it can say "no". In order to act differently if the network is in a fixed point or not we need to have a particular neuron, or set of neurons, that are ON in one case and OFF in the other. Our problem is to construct a set of synapses and neurons that achieves this objective. This problem may appear superficial, but it is not. Let us try to do it in a natural way.

To distinguish between the case of a stable state and an unstable state, it is reasonable to think about comparing the value of a neuron before and after a neural update. We can do this using a synapse that includes a time delay. Biologically, the time delay would be achieved through the axon rather than the synapse, but this is irrelevant to our argument.Using the time delay in the transmission of the signal from one neuron to the next, we can arrange to have the second neuron (the recognition neuron) receive information about the time evolution of the first neuron,in the form of the neuron value at one moment and its value at a time corresponding to one update later. We might think that this is enough to enable the recognition neuron to determine whether the neuron is changing or not. Surprisingly, this is not the case.

The reason that there is still a difficulty can be explained by considering the four pictures at the bottom of Fig. 3.2.1. On the left of each picture are illustrated the two states of the first neuron, before and after the update. The four pictures are the four possibilities.On the right in each picture is shown the value that we want the recognition neuron to take. The essential point is that we would like the recognition neuron to have the same value when the two states of the original neuron are the same, and the opposite value when they are different. This function is equivalent to the logical operation *exclusive or*, XOR, which gives TRUE (ON) if either input but not both is TRUE (ON) and FALSE (OFF) otherwise. What is illustrated in Fig. 3.2.1 is the opposite or negation of XOR if we follow the usual pictorial interpretation of UP as TRUE. We

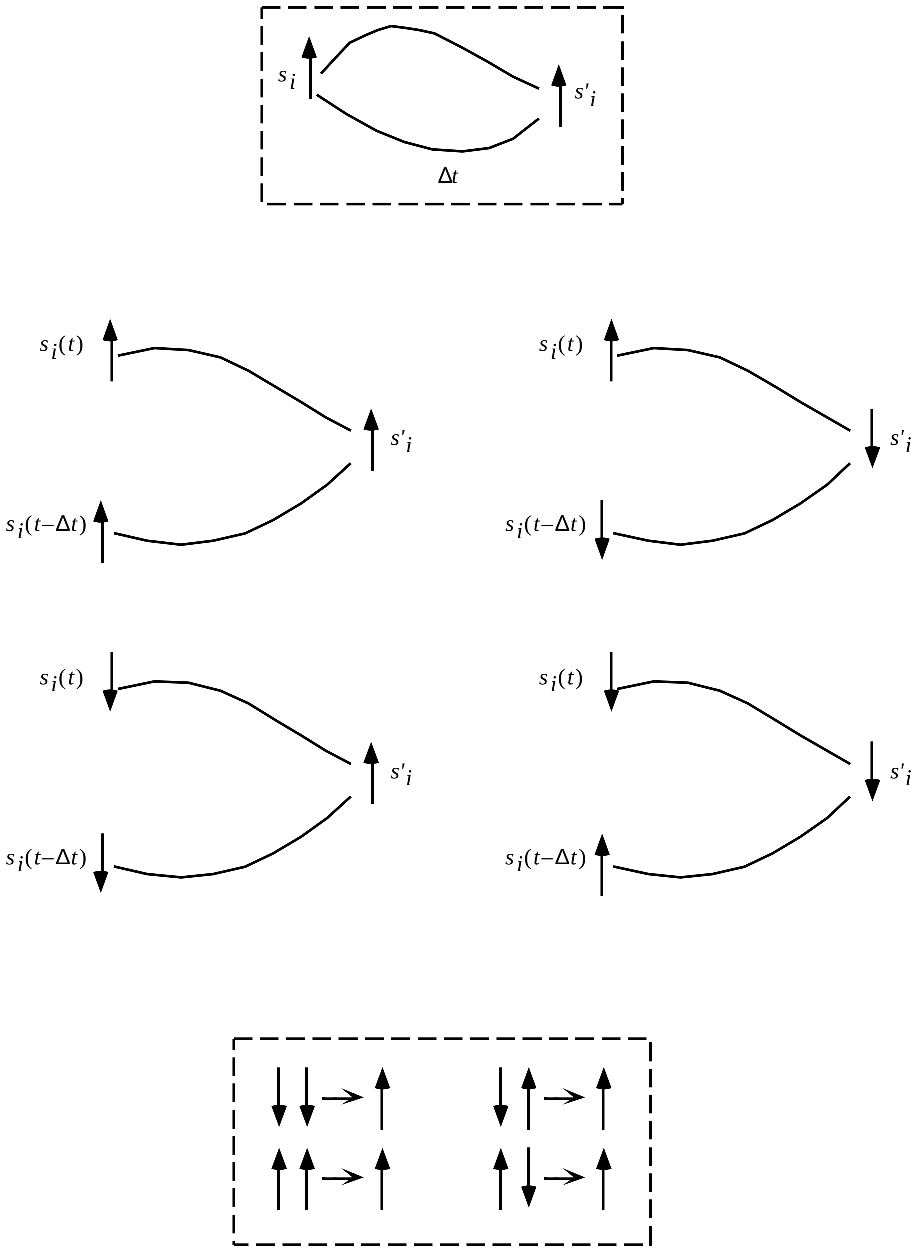


Figure 3.2.1 The problem of recognition requires the network to be able to respond the same to all patterns that have been imprinted, and differently to all patterns that have not. This requires detection of a stable state, which can be found from the time dependence of neural activity. To detect the stability of a particular pattern we use (top) two synapses, one of which is delayed by a time Δt. Both synapses run to a particular other neuron that is supposed to fire only when the two signals it receives are the same. The four possible cases of neuron firings are shown (center) where the left neuron is shown at two different times. Considering the right neuron as a function of two variables, we find (bottom) that it must represent the negation of the logical function exclusive or (XOR). This function cannot be represented by an attractor network. It can be represented by a feedforward network. ∎

can invert the definition or the picture to make the two agree. However, the problem is that the XOR logical operation cannot be built out of symmetric synapses between pairs of neurons. This is apparent when we remember the energy analog of the neural network (Section 2.2.3). Switching the activities of all neurons does not affect the energy, so the inverse of any low-energy conformation is a low-energy conformation. Inverting the XOR operation results in the opposite logical operation, not the same one.Pictorially in Fig. 3.2.1, we see that if the upper left is a minimum energy state, flipping all of the neurons would not lead to the lower left but instead would cause the recognition neuron to be inverted, giving the wrong information.

To overcome the problem of representation of the XOR operation and enable recognition requires the introduction of a new kind of synapse. There are many possible ways to do this. One is to use interactions between three neurons $s_is_js_k$. This breaks the inversion symmetry and enables the minimum energy states of the three neurons to correspond to the XOR operation. However, using a symmetric three-way synapse would still lead to some difficulties.A symmetric synapse, where the neurons influence each other reciprocally, does not really make sense when there is a time delay. Moreover, if a symmetric synapse is used,the recognition neuron could affect the other neurons rather than representing their state. This would not be helpful. A directed synapse such as the ones used in a feedforward network would be simpler.

There is a way to introduce an XOR operation using a feedforward network. This is discussed in Question 3.2.1. However, the feedforward network requires two stages for this operation,and it is not particularly convenient. Fortunately, we can probably do just as well with an AND logical operation. The AND operation would detect when a neuron stays ON. Ignoring the neurons that stay OFF, this would be enough to tell us when the state of a neuron is stable. The AND operation can be represented by a feedforward network using one stage (Question 3.2.2). Experimental studies of the biology of neurons also show the existence of individual directional synapses that couple three neurons. In some of these, two neurons must fire in order for the third to fire, thus directly implementing the AND operation. This solves the problem of enabling the recognition task to be performed. The recognition task is fundamental not only for the external recognition test that we have been describing but also for internal processes that lead to other capabilities. For our purposes in continuing to build models of brain function it is sufficient to note the necessity and biological plausibility for such logical operations.

Question 3.2.1 We have shown that an attractor network by itself cannot perform the XOR operation to perform a recognition task. Find a feedforward network with two layers of synapses that can perform the XOR operation. You may supplement the two neurons that you are comparing by a neuron that is always ON. Discuss the biological implementation of this feedforward network.

Solution 3.2.1 The XOR operation requires a comparison of two different binary variables. However, the feedforward network uses neurons represented by real numbers. According to the model we developed in Section 2.3, for two neurons s, s' in the first layer, we can write the value of any second-layer neuron as:

$$s_2 = \tanh(Js + J's' + h) \tag{3.2.1}$$

We can think of the constant term h as arising from a first-layer neuron that is always ON. The two independent linear combinations of the neuron activities that we have are $s + s'$ and $s - s'$. Either can be thought of as a comparator. If we construct a table from different values of s and s' we have:

s	s'	$s + s'$	$s - s'$	XOR(s, s')
1	1	2	0	–1
1	–1	0	2	1
–1	1	0	–2	1
–1	–1	–2	0	–1

(3.2.2)

Comparing the $s + s'$ and $s - s'$ columns with the XOR column we see that the XOR operation requires us to treat a positive sum and a negative sum the same, or a positive difference and a negative difference the same. We must therefore take an absolute value, or square the linear combinations. Two ways to write the XOR operation in terms of floating point operations are:

$$-\mathrm{sign}(|s + s'| - 1) = -\mathrm{sign}((s + s')^2 - 1) \tag{3.2.3}$$

The tanh function can provide us with the square of $s + s'$ by setting up a situation where we make use of its second-order expansion:

$$\begin{aligned} s_2 &= \tanh(h + J(s + s')) \approx \tanh(h) + J\tanh'(h)(s + s') + \frac{1}{2}J^2\tanh''(h)(s + s')^2 + \ldots \\ s'_2 &= \tanh(h - J(s + s')) \approx \tanh(h) - J\tanh'(h)(s + s') + \frac{1}{2}J^2\tanh''(h)(s + s')^2 + \ldots \end{aligned} \tag{3.2.4}$$

The expansion is valid if we use a small enough value of J. Setting up two second-layer neurons with these values, we can take their sum to eliminate the first-order term and keep the second-order term that we need. We use $J = 0.1$, and $h = 0.5$ to obtain the following table:

s	s'	$s_2 = \tanh(h{+}J(s{+}s'))$	$s'_2 = \tanh(h - J(s + s'))$	$s_2 + s'_2$	$\tanh(J'(s_2 + s_2) - 0.9J')$
1	1	0.604	0.291	0.896	−1.000
1	−1	0.462	0.462	0.924	1.000
−1	1	0.462	0.462	0.924	1.000
−1	−1	0.291	0.604	0.896	−1.000

(3.2.5)

The final column is the value of the neuron in the third layer (after two layers of synapses) that gives the XOR operation on the first layer of neurons s, s'. $J' = 1000$ is a large number that makes the tanh function into a sign function as required to obtain ±1. The value 0.9 that appears in the final formula is chosen to lie between the two possible values of $s_2 + s'_2$ shown in the previous column.

There are two difficulties with this representation of the XOR operation that are related to robustness and reliability in a biological context. The first is that it makes use of matched values of h and J on different synapses to ensure that s_2 and s'_2 have consistent values,and a matched value of J'. The sensitivity to different values can be seen, for example, by trying to use $J = 0.11$ only for s'_2 in the above table. The second is that this operation uses a second-order property of the synapse (the second derivative) that may be more variable than first-order properties. ∎

Question 3.2.2 Construct a logical AND using a feedforward network. You may supplement the neurons that you are comparing by a reference neuron that is always ON.

Solution 3.2.2 The second-layer neuron must fire if and only if both neurons of the first layer fire. If we add the activity of the two first-layer neurons, and require the result to be greater than a number greater than zero in order for the second-layer neuron to fire, then we will have the AND operation. This can be achieved in Eq.(3.2.1), for example, by setting J and J' to a large positive number, and h to its negative. Using a large number converts the tanh function into a sign function.

The existence of logical operations such as AND and XOR in the available functions of a neural network is interesting from a computer science point of view. For example,using just AND and negation (NOT), we can construct all possible logical operators (Section 1.8). This might suggest we could construct mathematical or logical operations of the neural network along the same lines as computers.One difficulty with this approach lies in the problem of representation. It is unlikely that the brain represents numbers in a conventional binary fashion. Instead,the word and number "one" are somehow represented as a state of the network involving many neurons. Thus the use of conventional logical operations on individual neurons by synapses is not likely to be the source of the brain's ability to perform addition. At the same time, we should not hesitate to make use of the logical operators at the level of individual neurons to justify the brain's ability to recognize images that have been imprinted. ∎

3.2.7 *Generalization*

One of the important properties of neural networks in pattern recognition or artificial intelligence tasks is their ability to generalize from a few training examples to a large number of cases.Generalization in a fully connected attractor network is simple to understand. The training creates a local minimum in the space of possible network states. The basin of attraction of this state becomes its generalization.

Partially subdivided networks provide an additional layer of generalization (Section 2.4). In addition to trained states, various combinations of substates (composite states) that may appear in the environment are recognized by the network. Since the network has been trained on far fewer states than it recognizes,it may be said to have generalized from the training set to the set of recognized states. This is an advantage ifthe architecture of subdivision of the network is in direct correspondence to the information to be presented. However, it can be a disadvantage if the new combinations are “errors”—states that do not appear in the environment. The advantage of using subdivision for language acquisition through grammatical decomposition of sentences was discussed in Section 2.4.5.

In Section 2.5.3, the simulation of a network with four subdivisions illustrated generalization in a subdivided network. The strength of the inter-subnetwork synapses determines the number and kind of composite states that appear as memories. Moreover, the combinations that are recalled are preferentially those that share some substate combinations with the originally imprinted states. For example, even though the network is divided into four parts,a state that is composed equally of two of the imprinted states is more likely to be stable than other possible combinations. This is not the same, however, as a network with two subdivisions. When there are four subdivisions, the combinations of two imprinted states can occur in three distinct ways. If we consider the substate imprints to correspond to attributes (features) of the information, this implies that novel combinations of features may be recognized if the combinations are not completely different from the imprinted states.

This description of our ability to generalize raises basic questions about the objective of brain function. We should not consider neural network models solely as a model of memory. Traditional evaluations of an individual’s ability, such as in exams, relied upon direct tests of memory. However, the central purpose of the brain is not to remember experiences, but rather to obtain from them knowledge that will serve in future circumstances. The memory of prior experience can serve in future circumstances when there are correlations between them. The purpose of the subdivided network is to abstract the essential aspects of an experience and the relationships between them, enabling this information to be used in future circumstances. In order to do so it is essential both to remember relevant information and to forget information specific to the particular circumstance. As discussed in Section 3.1, a significant role of sleep may be filtering memories, keeping the more persistent associations and forgetting associations that are specific to a particular circumstance. In this model the brain architecture is constructed so that the information to be forgotten largely consists of the associations between information stored in different subnetworks.

3.2.8 *Internal dialogue*

Through most of the twentieth century, behaviorism greatly influenced the field of psychology. Behaviorism attempted to describe all of human behavior in terms of reactions to a set of stimuli. However, it has become more generally accepted in recent years that descriptions of human behavior without invoking complex internal processes (cognition) cannot provide an understanding of more than a limited number of behavioral patterns. We can contrast the behaviorist approach with the concept of an internal dialogue that describes the ongoing language-based processes that occur in the brain without specific sensory stimulation or speech. One reason for modifying the behaviorist approach is the recent ability to measure neurological activity by means other than behavior. Tools for measuring this activity include positron emission tomography (PET) and magnetic resonance imaging (MRI). Even before these techniques, the behaviorist approach was not universally adopted. However, such imaging techniques provide a scientific basis and tools for investigating the internal processes. More fundamentally, adopting a model that includes an internal process, rather than a phenomenological behaviorist approach, is justified when it is easier to describe behavior using the internal process.

In this section we discuss some features of a neural network that are necessary for the existence of processes that are,at least in part,independent of the immediate sensory information. Such independence is not found in a feedforward network, where the input is progressively transmitted through stages to the output. It is also not realized in an attractor network, where the initial state of the whole network is fixed by input and the internal dynamics evolves the state to an attractor. These models are thus incomplete, because thought,and the internal functioning of the brain,is often largely independent of the immediate sensory input. People are able to think about a problem without regard to circumstances unless the circumstances become demanding upon their attention. We rely upon this when exams are given to students, since we do not generally consider most sensory information in the room as relevant to a student's performance,unless there are significant distractions. How is this independence realized in a neural network model?

A natural solution to the problem of introducing an internal dynamics is to assume that there are internal loops whose input is their own output. These loops interact with sensory input,and with motor output, but are not dictated by them. The flow of neural influence is illustrated in Fig. 3.2.2. We can associate the internal processes, for example, with internal dialogue. The existence of such loops leads to several concerns. The first would be that the system becomes stuck in a self-consistent loop. This is the dynamical analog of the excessively strong attractor that was discussed in the context of fixation. To avoid this problem requires some protections against repetition. For example, a neuronal refractory period that is longer than the

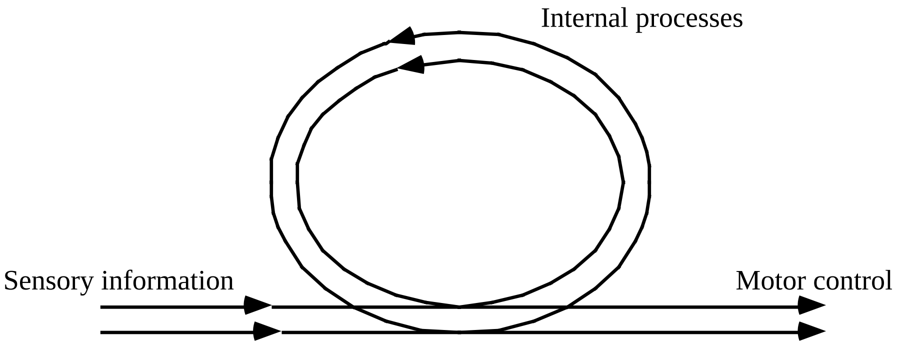


Figure 3.2.2 Internal processes such as internal dialogue can continue largely independent of sensory information. This requires internal dynamical loops that receive input both from the senses as well as from their own recursive structures. The figure illustrates a simple feedforward system with a recursive internal loop. ∎

cycle time could protect against single-cycle repetition. However, it would not protect against double-cycle repetition. While persistent loops are to be avoided,it is also possible for an internal cycle to preserve information over time. This may provide a form of short-term memory. Among other capabilities,a short-term memory enables juxtaposing, at one time in the neural state, events that are separated in time.

A second major area of concern in a discussion of internal loops is the balance that is established between the influence of the senses on the internal processes and their independence. There are dangers associated both with excessive coupling or decoupling of the internal processes from the senses. The stereotype of the absent-minded professor may be a manifestation of a particular balance between connection and independence that might be realized in this model. This balance of connection and independence is realized through the strength of particular sets of synapses. As discussed in Section 3.1, such balances may be maintained through the processes that occur during different stages of sleep.

A related question raised by the model of Fig. 3.2.2 is the relative capacity of the information paths from the senses,as compared to the information paths cycling internally. Specifically, what fraction of the information present in the brain at any one time is a direct consequence of the sensory input? This should play an important role in our understanding of the qualitative behavior of the brain. Is it largely driven by the outside or is it largely internal? When a system is completely determined by the immediate sensory information, we would identify it as a reactive system. When the sensory information determines only a part of the internal state, we can talk about the external and internal worlds as they are manifest in the state of the network. The problem then becomes to identify the relative complexity and interactions between the external and internal worlds. We will discuss the quantitative characterization of complexity in Chapter 8.

3.2.9 *Imagination, creativity and error*

There are two forms of creativity that are often discussed separately. The first is the general human ability to create new sentences, or to respond to circumstances that have not been experienced before. The second form of creativity is considered to be rare and is associated with particularly “creative” individuals. In this section we discuss the first more general creativity. The second is related to the first, but also requires an understanding of individual differences, and therefore it will be discussed in the next section.

The term “create” implies an act after which something exists that was not present before. However, acts of creation involve bringing together elements that were previously in existence, but juxtaposed in new ways. The elements may be objects,attributes or relationships. In order to create there must be an external manifestation—the act of creation. However, a precursor to the act of creation is imagination. The ability to imagine is the ability to represent internally a combination of elements that was not previously experienced. Creativity thus involves both imagination and implementation. Imagination requires a partial independence of internal representations from the external world. Otherwise, the internal state of the brain would only reflect the external reality and there would be no imagining or creativity. The previous section on internal dialogue described how such independence can be implemented. Here we focus on the problem of generating internal representations that differ from imprinted information.

The ability to imagine novel combinations of elements is implicit in the subdivided network we have been investigating. The network can generate a set of stable composite states that are,in effect,untrue memories. Assuming that the internal neural dynamics described in Section 3.2.8 explores various possible states of the network, these composite states appear from time to time as "imagined" possible combinations of partial states that were not imprinted. For example,having seen a bird in flight and a walking human, one might imagine a composite consisting of a flying human.

The extent to which composite patterns appear is controlled by the relative strength of inter-subnetwork synapses and intra-subnetwork synapses (the parameter *g*) discussed in Chapter 2. The progressive decomposition of memories during sleep, discussed in Section 3.1, suggests that sleep is also intimately related to the emergence of composite patterns. Composite patterns would appear first during sleep in the form of dreams, most of which would not be remembered. The partially subdivided network reflects both the concepts of divergent and convergent thinking. Divergent thinking is the ability to imagine new combinations. Convergent thinking is reflected in the inter-subnetwork synapses that limit them.

One question that might be asked is: How does the network distinguish between imagined states and real memories? A possible answer may be found in the relative strength of their basins of attraction. It can be shown that the basin of attraction of composite states is smaller than that of imprinted states. This may enable the network to distinguish them using a strategy similar to that described in the section on recognition. However, it is also apparent that some degree of confusion may arise. Isolated occurrences would result in false memories. In extreme cases, this confusion may give rise to functional disorders. This is consistent with the existence of a variety of psychological disorders involving hallucination. Thus the possibility of hallucination is rooted in the basic nature of the network architecture that enables imagination to occur.

Another consequence of the model of imagination is a trade-off between memory and creativity. In order for new composite states to be formed,the strength of associations between subdivisions must be reduced. The relationship between the elements that were originally imprinted tends to be lost. The trade-off between storage of more composite states and more imprinted states discussed in Section 2.4 appears here as a trade-off between memory and imagination, or even memory and creativity. Thus, for example, the ability to combine words into new sentences also requires a forgetting of sentences that were heard or spoken before. Memory requires maintaining the associations, while creativity requires loss of associations so that novel combinations can be imagined.

We should also make a connection between creativity and error. Even the most basic form of creativity—the application of prior experience to new circumstances—requires the possibility of error. More generally, in any act of creation there must be a possibility for error. An error can be defined as a creation that is not consistent with the external world.

The possibility of error implies the importance of limiting creativity. To manifest all possible combinations of elements, while in some sense creative, would not be effective. Creativity is only effective when the many possible combinations are limited to those that are more likely to be correct.A partially subdivided network appears to be an effective approach. It limits the number and type of composite states.Limiting creativity in this way reduces the probability of errors; however, it does not eliminate them.

The interdependence of creativity and error, two characteristics of human activity, should not be considered a limitation of our neural network model;it appears instead as a fundamental relationship. This relationship should persist despite improvements in the modeling and understanding of brain function.

Using the picture we have developed for creativity and error, we are able to begin to describe individual differences. The degree to which subdivisions of the network are isolated—the parameter *g*—can describe a one-parameter variability between individuals. Individuals who have a smaller value of *g* will be more forgetful,more creative and more prone to error. Individuals with a larger value of *g* will retain more memories, be less creative and less prone to error. This prediction could be tested by psychofunctional tests of a group of individuals. Here we can consider allegorical evidence from conventional stereotypes of various professions. The conventional stereotype of the most creative profession—artists—as also the least practical,can be contrasted with professions requiring few errors, such as accounting. Since the consequences of error are diminished,the arts would be expected to attract more creative individuals (lower *g*), with weaker memories and higher susceptibility to error. On the other hand,individuals with lower levels of creativity (higher *g*) and greater memory retention would be expected to be more successful in professions where consequences for error are higher.

The preceding paragraph begins to identify distinctions between individuals;however, this is only a small step toward understanding individuality or, more specifically, the second form of creativity that is attributed to specific individuals. The artistic creativity of Picasso is typically considered to be a completely different phenomenon from the commonplace ability to form new sentences. Nevertheless,it is possible to suggest they are quite similar. To do so, however, requires us to go further into an understanding of the subdivided network architecture and the source of individuality.

Question 3.2.3 Why didn't we consider spurious states introduced in Section 2.2.7 as a source of imagination/creativity?

Solution 3.2.3 Spurious states, like composite states, are formed from combinations of imprinted states.Spurious states,however, do not generally retain identifiable aspects of the original states. This is because they are formed by combining individual neuron activities from each of the imprinted patterns, rather than neurons associated with a particular attribute. The only structure imposed upon spurious patterns is by virtue of their overlap with imprinted states.Spurious states may be stable states of the network, and therefore may be imagined. However, unlike composite states, in general they will not have a coherent interpretation. ∎

3.2.10 *Individuality*

The design of modern computers relies upon a set of models that perform all computational tasks (Section 1.9). Despite various architectural differences,there is a uniformity of function.Often the objective of new models of computation,including neural network models, is to demonstrate that they have sufficient capability to be classified with computers—they are capable of universal computation. We have argued already in Section 1.3 that one of the essential characteristics of complex systems is the distinction between different realizations of the same architecture. Consistent with this, the subdivided neural network suggests an entirely different approach to computation based on a nonuniversal computation strategy. This nonuniversal strategy is the subject of this section and forms a basis for understanding human individuality.

Before proceeding, we mention that different computers, or a computer at different times operating on different information, behaves in different ways. We might suppose that this would allow us to use the universal computation approach to account for individual differences. However, one aspect of the concept of universal computation is that the basic capabilities are universal even though the particular data and the particular hardware are not. For example,certain problems that are inherently difficult for one computer running one computer program will also be inherently difficult for any other computer running any other program. There are various assumptions inherent in this statement,and it would be more correct to formulate it in terms of computational complexity classes. However, in the case of the human architecture, it appears that the capabilities are fundamentally different between different realizations of the architecture.

The reason for nonuniversality is rooted in the original motivation for subdivision—correlations and independence in information. Our environment manifests correlations that are nonuniversal. Tree leaves could be any color, or could be colored at random. Objects need not retain their shape over time. Subdivision exists because of the correlations in the information that is presented to the individual by the external world. By structuring the information internally in a way that is compatible with the structure of the external information, the subdivided architecture is designed to accommodate to it, or take advantage of it. However, from a computation theory point of view, there is no reason for the information to be structured in a particular way.

Once again our simplest example,the left-right universe (Section 2.4),is helpful. We contrasted the capabilities of a network divided right from left and the network divided top from bottom. These networks had radically different capabilities in the left-right universe. This demonstrates in a simple way how the capabilities of distinct individual realizations of the same architecture may vary drastically.

The inherent nonuniversality of the architecture of subdivision is modified by the effect of selection due to fitness,which can lead to commonality between individuals. Thus, for organisms in the left-right universe we would expect to find only left-right subdivided networks and no top-bottom subdivided networks. Similar commonalities should also be expected among people. Thus the variability in brain architecture and the resulting variation in capabilities is limited to the degree that selection imposes the architecture as a result of evolutionary processes (Chapter 6). Thus we have argued that there can be an environmental/evolutionary pressure toward commonality in brain architecture because of a commonality in the environment of different individuals. However, this commonality is limited to the actual impact of selective forces.

The nonuniversality of the subdivided network becomes clearer when we think about the hierarchical structure motivated by the 7±2 rule and the large variety of possible mappings of sensory and motor information onto this structure. Consider the many different filters of information that might be useful under different circumstances. It is possible for a single individual to have many of them and to selectively use them. In the extreme case we can ask: If there are many possible filters of information that might be useful, why doesn't each individual make use of all of them? The first answer to this is that the number of such mappings grows exponentially with the amount of information,so it would be impossible to contain them in a single realization of the architecture.

We also recall that much of the usefulness of subdivision is lost when the number of subdivisions becomes greater than seven. In a hierarchy, we can use more than seven distinct filters; however, choosing how to arrange them matters. The strongest associations are maintained between information that is connected at the lowest level of the hierarchy. Progressively weaker associations exist between subdivisions that are connected at higher levels of the hierarchy. Depending on how the filtered information is mapped onto the subdivisions, an individual will retain distinct associations leading to a wide variety of possible individual differences.

Using the individualized hierarchically subdivided architecture as a model of the brain we can return to a consideration of imagination, creativity and memory. The functional hierarchy corresponds to a nonunique selection of attributes distributed in a tree of stronger and weaker interconnection. This nonuniqueness suggests that different individuals will remember different associations and also be creative in different ways. For example,some individuals will find it easy to remember the association of names and faces while others will not. Those who remember these associations have these attributes strongly connected to each other. In this model, generic capabilities of an individual are directly related to the organization of information within the architecture of the brain.

Our conclusion is that unlike modern computation theory, the subdivided architecture of the human brain is a nonuniversal architecture whose individual realizations have widely different task-dependent capabilities. We also surmise that a universal strategy may not be effective at many human information-processing tasks. The nonuniversal architecture is consistent with the uniqueness of individuals.

3.2.11 *Nature versus nurture*

A central controversy in modern science revolves around the relative importance of the genetic code as compared to environmental influence in determining human behavior and various aspects of brain function and human information-processing. This is often called the nature versus nurture controversy. The model of the brain formed from a hierarchy of functional subdivision also provides us with a model for the relative influence of nature and nurture.

To extend the model for individuality to a first model of the influence of nature and nurture we need only suggest that the subdivided architecture itself is genetically controlled. On the other hand, the information that is imprinted upon it is a direct result of the environment. Thus,the aspects of information that are retained by a particular individual are guided by genetics. Genetics controls the type of associations that are strongly retained,and those that are readily forgotten. The environmental influence is contained in the actual associations and information that is present. This model exhibits a complementary influence of genetics and environment on an individual. It shows explicitly how genetics influences potential qualities of an individual, and how the environment influences the actual qualities.

It is important to emphasize that while this picture is appealing in its simplicity, it must be considered only as a first approximation. Aspects of the subdivided architecture are susceptible to environmental influence. For example, even the development of basic interconnections of the visual system are influenced by exposure to light. A limited amount of specific information may also be built in due to genetic programming. These include instinctive behaviors that are more prevalent in animals.

3.2.12 *Consciousness and self-awareness*

Of all the traits associated with human beings,the ability to be self-aware,and the related concept of consciousness present some of the most difficult philosophical dilemmas. The practical implications of these dilemmas are related to the concepts of free will and determinism and the related responsibility of an individual for action. We separate consciousness from the problem of selective awareness (“I am conscious of …”) which is described in the following section. In this section, after discussing some conceptual obstacles, we will construct a neural network model of consciousness. We then discuss practical reasons for its existence,a test of the model’s ability to recognize self, and modes of failure of the model which can be compared with psychofunctional failure.

Conceptually, a paradox in considering the problem of consciousness arises from the problem of recursive signal processing. Consider, for a moment, the problem of consciousness as that of being aware of sensory input, and consider the neural network as a form of information processor. An example is the sensory processing that is performed by the neurons that receive and process visual information. There is no indication that this provides any awareness. It is simply a mapping of sensory information into another form. This information is transferred to the input of another neural system. The second system is now handed the job of providing the awareness. However, no process that transforms the sensory information further would do anything qualitatively different. Thus we are perpetually left to the problem of deferring the consciousness to later, more internal stages, without resolution. Some have argued that the ultimate recursion must lead to something that is unphysical and therefore outside the domain of scientific inquiry.

The problem with this model of the process of consciousness is that is places the consciousness in the wrong place—as the primary recipient and interpreter of sensory information. It also uses only a limited view of the function of a neural network. We will address both from a different perspective.

We begin by considering a model that differentiates in an essential way between the conscious and subconscious mind. There is generally no disagreement that such a differentiation should be made,since casual introspection shows that the conscious mind is not aware of—does not contain—all of the internal processes taking place in the brain. Indeed,it is aware of very limited aspects of these processes. Thus we begin by constructing a subconscious part of the brain. The responsibility of the subconscious region of the brain is to receive sensory input and to act upon it. This may seem strange at first sight; isn't the conscious mind necessary? The answer is, largely, no. Habitual acts and most of the details of daily activity are performed directly without apparent input from the conscious part of the brain. The easiest way for us to represent the subconscious brain is as a conventional feedforward network that takes the sensory input and determines motor control based on this sensory input. Thus far we have done nothing at all unusual except to claim that this model does not possess consciousness, which we knew from the outset.

Now we would like to construct consciousness. We do this from a pragmatic point of view by asking, What is the information that the conscious mind possesses? Introspection suggests that the conscious mind possesses sensory information. It also possesses knowledge of motor activity. However, it does not possess information about the internal processes that lead from sensory to motor activity. This suggests that we construct a new part of the network model that represents both the sensory information and motor activity information, but not the intermediate stages. The next question to ask is, What does the conscious mind control? The first answer is that it has no primary control function. By this we mean that control is not continuous in the same way that it is for the feedforward network. We can see this from the terminology—the awareness or consciousness do not convey the meaning of action. They are rather passive terms describing the possession of information. No action is required on the basis of this information.

There is, however, a secondary control function. The awareness is capable of exercising control over the motor activity. However, this control is circumscribed. It acts as a corrective process rather than a control over moment-by-moment action. Thus the direct control over action is performed by the subconscious network, while the conscious network acts by redirecting the subconscious feedforward network.

How does the conscious network decide to exercise control over actions of the subconscious network? We answer by considering the conscious mind as an attractor network (Fig. 3.2.3). The pattern of neural activity in the attractor network represents both sensory and motor activity. It's task is to recognize their "compatibility." As discussed in Section 3.2.6, recognition can be performed by measuring the dynamics of the attractor network. If the juxtaposition of the sensory and motor activity is recog-

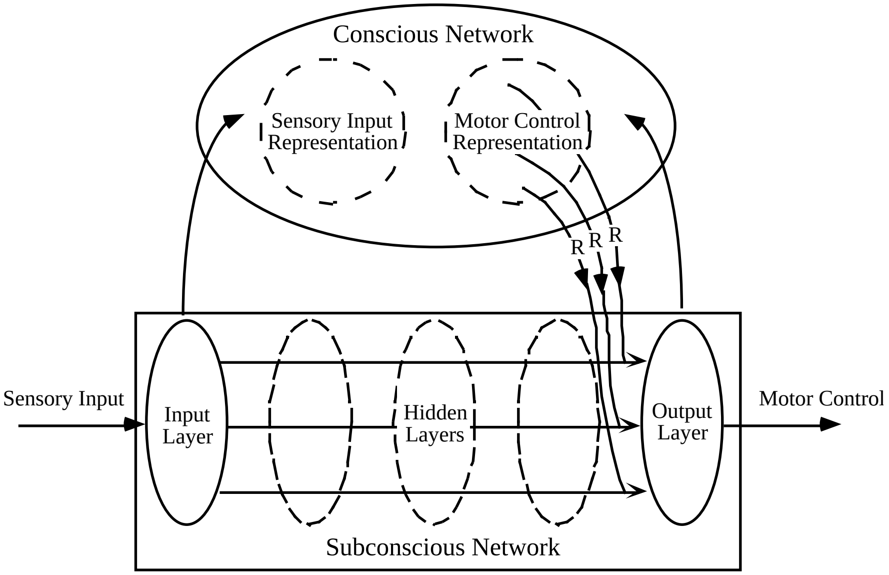


Figure 3.2.3 A model that captures some of the essential features of self-awareness and consciousness can be constructed out of two parts. The first part, representing the subconscious mind, is a feedforward network that is a sensory motor control system. The input is sensory information and the output is motor control. The second part, representing the conscious mind, is an attractor network whose state is composed of input both from the sensory information and from the motor control. It has no direct control function. However, when the sensory and motor information is not recognized as an imprinted state the network exercises control over the actions through recognition synapses similar to those discussed in Section 3.2.6 and Fig. 3.2.1. In effect, the imprinted states in the conscious network represent a model of the self. When the actions are not consistent with the model it intervenes to change the behavior. The network function is illustrated schematically in Fig. 3.2.4. ∎

nized,then the conscious mind does not interfere. However, when the state of the sensory and motor activities are not recognized,then the conscious mind acts by causing the feedforward network to modify its actions. This occurs over a longer time scale than direct action by the subconscious network.

An interesting way to summarize the recognition process that the attractor network performs is as a question. The question, in this case, is: Is this me? Or specifically: Is the current situation and my actions within it consistent with my self-image? The self-image is the set of stable states of the attractor network. Summarized in this way, we see how the notion of self-awareness and consciousness are related and are represented by this model.

An additional concept that can be described is the concept of a will. It is easiest to identify the will by noting the use of the modifiers that describe an individual as having a strong or weak will. The will represents the ability of the conscious part of the mind to control the subconscious. In this model this is represented by the strength of the synapses that originate in the conscious attractor network and act to modify the state of activity of the subconscious feedforward network.

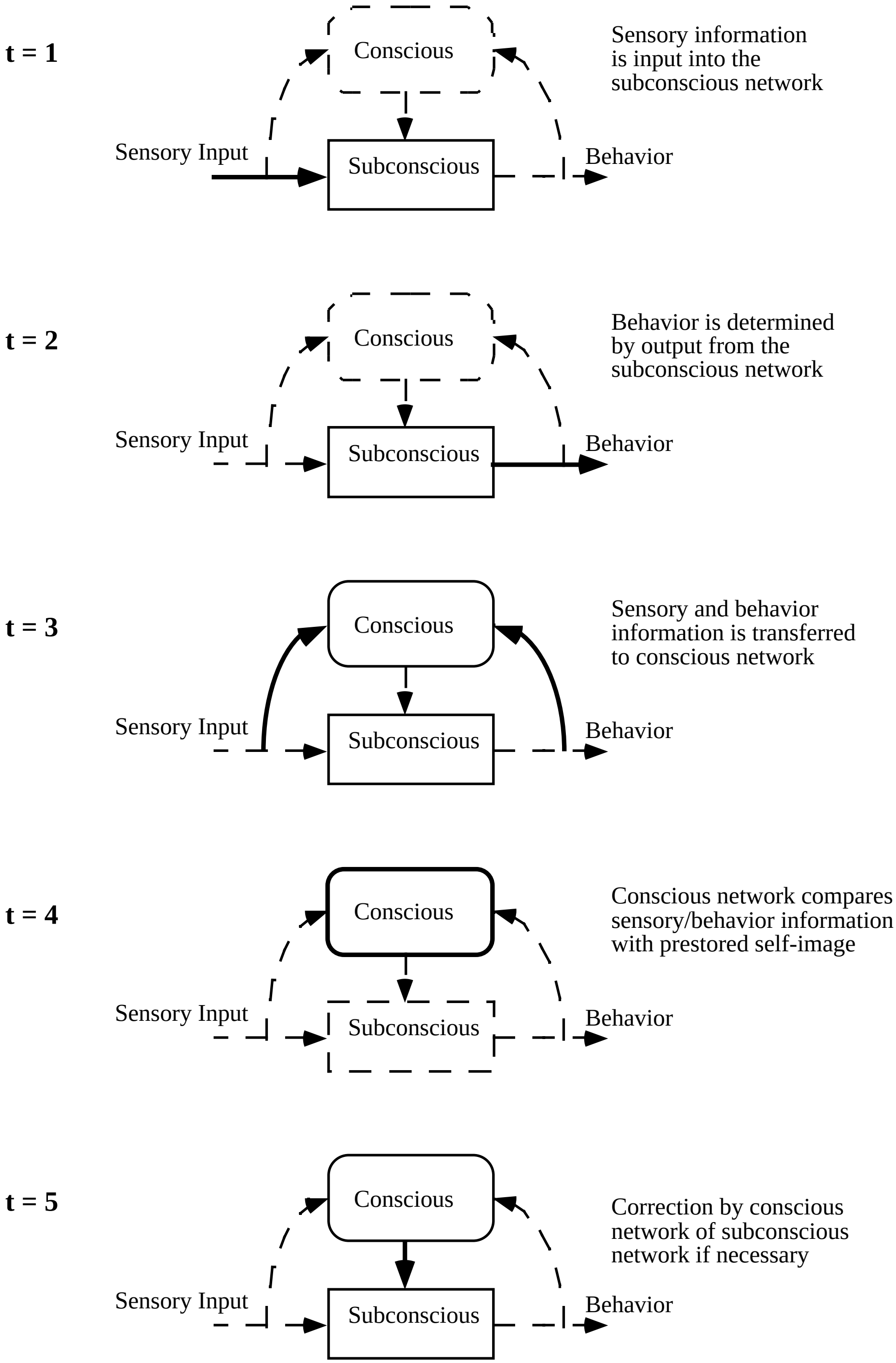
The architecture we have constructed,shown in Fig. 3.2.3,is a comparatively simple model that captures some of the features that we attribute to the function of the conscious and subconscious parts of the mind as well as the interactions between them. The interactions are described pictorially in Fig. 3.2.4. Given the conflicts that have arisen out of the concept of consciousness, the possibility of discussing a concrete model provides some new opportunities for progress in our understanding.

The practicality of consciousness can be considered by realizing that the combination of feedforward and attractor networks provides a solution to the limitations of each of these network architectures. As discussed in Sections 2.3 and 3.2.2,the training of a feedforward network requires storage of the desired input-output pairs. In our model of consciousness, this storage is performed using the attractor network. Moreover, the training of the feedforward network must be done incrementally, while that of the attractor network may be achieved by a single imprint.On the other hand, the attractor network is not capable of complex processing and will suffer overload if too many patterns are stored in it. In this model, complex processing can be left to the feedforward network, and once it is trained,the size of the basin of attraction of the attractor network pattern can decrease. Significantly, it is not necessary for the attractor network by itself to be able to generate the pattern representing a response,it must only verify and provide corrections to this response.

Consciousness is characterized by the recognition of oneself. Experimentally it is manifest in the ability to recognize oneself in a mirror. This ability is not present for animals other than apes and man. Even monkeys appear unable to recognize themselves. The neural network model indicates how self-recognition can occur. By virtue of juxtaposing sensory and motor information in an associative memory (attractor network), it enables correlations between them to be imprinted. When moving and seeing this motion in a mirror, the imprinted information is recognized by the conscious network, and thus the answer to the question, Is this me? is yes.

The physiological location of the conscious and subconscious networks can be tentatively identified. The frontal lobes that are much more developed in apes and man than in other animals have generally been identified with consciousness and planning. They also have a topographic map of the body that serves as an area of motor control. However, the motor control due to the frontal lobes has been associated with voluntary rather than habitual motion. Involuntary movement and the coordination of voluntary and involuntary movement are both centered in the cerebellum,

Figure 3.2.4 Schematic illustration of the model of consciousness described in the text. This model assumes two components of the mind—the conscious mind and the subconscious mind. The activity and interactions of the two components are illustrated in the figure as a time sequence. The subconscious mind is directly responsible for receiving sensory input and acting upon it. The conscious network corrects the subconscious based on an internal representation of the self, and triggers retraining of the subconscious network. ∎



Sensory information is input into the subconscious network

Behavior is determined by output from the subconscious network

Sensory and behavior information is transferred to conscious network

Conscious network compares sensory/behavior information with prestored self-image

Correction by conscious network of subconscious network if necessary to which there are many projections from the motor areas of the frontal lobes. These observations are consistent with the neural network model.

While most animals do not recognize themselves in a mirror, they do have a sense of self associated with location/territory or smell. The part of the brain associated with representing information about spatial location is the hippocampus. This part of the brain may serve as the associative network that enables an identification of self related to location, as in:"I am here" or "This is my place." Recent experiments discussed in Section 3.1.5 identify the hippocampus as a network that stores information in correlated patterns of neural activity—an associative network. This also suggests that consciousness is not a monolithic entity; it may have various aspects related to different parts of the brain.

As with other aspects of the models we have discussed, the model of consciousness provides an understanding of some of the failure mechanisms of the network. One failure mechanism is found by considering the strength of the control by the conscious over the subconscious—the will. We see that the subconscious mind is essentially reactive. When the will is weak, the behavior would be characterized as impulsive. On the other hand, if the will is too strong, then the attractor network, which does not have the ability to process information through several layers of synapses, takes over the reactive function. This implies that actions are based upon relatively simple conscious processing. When all actions are based upon simple conscious processing, behavior is characterized as fanatic.

We can take the discussion one step further by discussing changes in the will. Similar to other synapses, the will is likely to be changed by imprinting. Thus it is strengthened by action and weakened by inaction. Since the exercise of the will is under conscious control,it may be strengthened by consciously exercising control even when the control is unnecessary, or it may be weakened by passivity when the control would otherwise be exercised.

In this section we have emphasized the limited control that the conscious mind exercises over action. However, the conscious mind appears to exercise direct control over what we are paying attention to, as discussed in the following section.

3.2.13 *Attention*

One of the phenomena associated with both internal dialogue and response to sensory stimuli is that of attention. We are able to be aware of various aspects of sensory information, or focus on a particular thought. How would we design a system that can achieve this? One approach is used by computers, where a central processing unit receives information from different parts of the memory according to its instructions. The central processing unit must label the information according to where it is taken from in the memory. This requires an addressing/labeling that distinguishes one part of the brain from another. We will discuss an alternate strategy that leaves the information in place but acts as a kind of spotlight. This approach is better suited to the neural network models we have been describing, because the nature of information is established by its location in the brain rather than by retrieval and labeling. We continue to avoid an explicit labeling scheme in this manner.

Until now we have been comfortable with the model that neurons are firing or not firing with roughly equal probability, and the pattern of activity or inactivity represents the information that is present in the mind at a particular time. We now need an additional intrinsic label that will enable us to identify which region of the brain has our attention.One way to achieve this is to give up the symmetry between activity and inactivity and assume that significantly more of the neurons are inactive; it is then the neurons with significant activity that are representing the information. This helps because we can then control the overall activity level in a particular region of the brain. If we raise the overall activity we are drawing attention to it, and if we reduce the level of activity we are reducing our attention to it. It is indeed well established that the neurons in the brain are active less than half of the time. Moreover, imaging experiments that are assumed to measure which parts of the brain are utilized at a particular time measure their average neural activity, which is higher than in other regions of the brain.

How does this change affect all of our previous analyses of the storage of patterns in attractor networks? The answer is that qualitatively very little changes. A pattern that is to be imprinted consists of a pattern of neural activity where the fraction of active ($s_i(t)$ = 1) neurons is less than one-half. The imprinting rule may be modified slightly to prevent the bias itself from being imprinted. If the average activity of the neurons is consistently *m*, then the Hebbian imprinting (Eq. (2.2.6)) may be modified to read:

$$J_{ij}(t) = J_{ij}(t-1) + c(s_i(t-1) - m)(s_j(t-1) - m) \qquad i \neq j \tag{3.2.6}$$

This means that imprinting results from deviations from the average activity. The network capacity as measured by the number of patterns that can be imprinted and retrieved actually goes up slightly, because, in effect, the patterns do not interfere with each other as much since they involve different sets of firing neurons. However, the overall amount of information that can be stored is diminished because each pattern does not contain as much information. For our purposes, these are minor adjustments to the results that we have already found in Chapter 2.

In order to make use of the bias in neural activity for the purpose of attention, there must be a mechanism by which the overall activity within a particular region of the brain is controlled. We will describe a mechanism for such control. The mechanism must be independent of the neural activity and synaptic transmission that we have been describing. Throughout the brain there are found cells whose function is not understood. These cells,called glial cells,may have some function in maintaining the structural integrity of the neural system. We will invest these cells with a model for how the attention system might work, recognizing that there are other possible embodiments. The essential property that we are using is that these cells are not part of the neural representation themselves. This role could also be taken by a separate set of neurons. The reason it appears natural that the cells involved would not actually be neurons is that they do not require specificity of interaction with a particular neuron. Instead they should interact more generally with a whole region of neurons. There are, however, some classes of neurons that do this as well.

In order to fulfill their function,the glial cells would need two capabilities: to control the overall activity of the cells and to measure their overall activity. Control over the activity might be achieved by control over the local blood flow supplying nutrients to a region of cells, or by control over the chemical contents of the blood. Alternatively, chemicals in the intercellular fluid might be involved. The purpose of these chemicals would be similar to that of neurotransmitters in that they inhibit or excite neuron activity; however, they are likely to be quite different in detail,since they have an effect on the overall level of activity rather than serving as one of many signals to a cell. Measuring the overall activity of a region of neurons may be achieved by sensing various by-products of their electrochemical activity. This measurement would take longer than the transmission of an individual pulse along an axon; however, the activity itself is an average over many transmission pulses.

The behavior of the glial cell then becomes similar to the behavior of a neuron, in the sense that it has either an ON or an OFF state. In the ON state it promotes the activity of the neuronal assembly it is in contact with;in the OFF state it suppresses it. It acts as a metaneuron that is related to the average neuronal activity. Control over the glial cell may then be exercised in several ways. For example, a self-consistent attention mechanism could be formed by glial cells attempting to activate the assembly of neurons they are in contact with whenever the cells are significantly active. The glial cell measures the activity present with respect to the expected activity. If the glial cell is OFF, the neurons would be generally inactive. If the cells become significantly more active than expected,the glial cell turns ON and promotes the activity of the region of cells. If the glial cell is ON and the activity falls below that expected, then the glial cell turns OFF. Interactions between glial cells that suppress one glial cell when another is active would lead to an exclusive attention mechanism.

An important part of the phenomenon of attention is that it is coupled to consciousness. We can implement this coupling by assuming that the conscious part of the brain, discussed in the previous section, controls the glial cells and thus the regional neural activity. We emphasize again that the term “glial cell”as used here might be substituted by another biological analog without changing the essence of this discussion. Moreover, it should be clear that the mechanism we have described for attention is not the only approach. It is one of the ways that are consistent with the spirit of the neural network models we have been developing.

One of the interesting outcomes of this model of attention is that it provides a mechanism for a new level of dynamics that would be related to a sequential activation of glial cells causing a sequential activation of particular regions of neurons. This can provide a missing piece in our discussion of language in the subdivided architecture. In Chapter 2 we suggested that different subdivisions of the brain are responsible for storage of distinct parts of speech. We can now suggest that a sequential firing of glial cells results in sequential activation of different parts of speech that trigger verbalized speech,are triggered by hearing speech, or represent internal dialogue in the form of organized strings of words—sentences.

3.2.14 *Summary of brain function*

We have devoted this chapter to building a relationship between our neural network models and several of the basic phenomena of brain and mind. The relationships have not only described some of the interesting phenomena, but also created a framework in which poorly understood concepts can at least be discussed. These include such diverse concepts as sleep, creativity and consciousness. To incorporate these into our model, we expanded the basic neural network to include various additional features. The comparison of many of these with the actual brain has yet to be performed. However, it is helpful to have theories that can be tested both experimentally and through simulations.