©1986 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring and Summer 1986, Volume 7, Numbers 2 and 3
Pages 229 [99]-250 [120]
ISSN 0271-0137
ISBN 0-930195-02-7

REM Sleep and Neural Nets

Francis Crick
The Salk Institute

Graeme Mitchison
Kenneth Craik Laboratory
Cambridge, England

The broad features of Rapid Eye Movement (REM) sleep are reviewed. Memory storage in the brain is probably quite unlike that in a digital computer, being distributed, superimposed and robust. Such memory systems are easily overloaded. If the stored memories share common features, random stimulation often produces mixed outputs. Simulations show that such overloading can be reduced by a process we call "reverse learning." We propose that this process is what is happening in REM sleep and that it explains in an unforced manner the condensation commonly found in dreams. Evidence for and against the proposed theory is discussed and several alternative theories are briefly described. The absence of REM sleep in the *Enchidna* and in two species of dolphins (that have relatively large brains) suggests that REM may allow the brain to be smaller than if REM were lacking.

In this short review we aim to explain our ideas about the function of REM sleep (Crick and Mitchison, 1983), to relate them to other contemporary theories and to make a critical assessment of them. As far as possible, we have avoided abstract mathematical arguments, preferring to present our ideas more informally.

The core of our suggestion is that there is a special process related to memory which operates in REM sleep and is in a loose sense the opposite of what happens when we are awake. We have called this process "reverse learning." This idea did not come from an explicit consideration of REM sleep and dreams but from theoretical studies on the way large groups of neurons might interact together. These, as we shall see, make it seem likely that the brain

We thank our colleagues for many helpful comments, in particular Richard Durbin, Robert Haskell, Allan Hobson, John Hopfield and Liam Hudson. This work was supported by the J.W. Kieckhefer Foundation, the System Development Foundation and the Samuel Roberts Noble Foundation. Requests for reprints should be sent to Francis Crick, Ph.D., The Salk Institute, P.O. Box 85800, San Diego, California 92138.

behaves in a rather different way from a modern digital computer. The basic idea can be said to be "obvious," since it occurred to no less than three groups of workers independently, all of whom were studying the behaviour of rather idealized neural nets. Clark, Winston, and Rafelski (1984) termed this conception "brain washing" and later suggested it occurred in non-REM sleep (Clark, Rafelski, and Winston, 1985). Hopfield, Feinstein, and Palmer (1983) called it "unlearning." They made no suggestions as to its possible occurrence in sleep.

We shall first sketch our understanding of REM sleep, non-REM sleep, and dreams, in order to outline what we feel needs to be explained. We next describe what is believed to be the general nature of memory storage in the brain and the very oversimplified models used to study it theoretically. We go on to explain, in a non-mathematical way, the process of reverse learning. Then we point out that the general nature of the bizarre intrusions in REM dreams is just what would be expected from our ideas. Finally, we discuss various criticisms of the theory, touch briefly on the side effects of REM and outline some alternative theories. A recapitulation and summing up concludes the review. We are not concerned in this review with the detailed mechanisms which control the amount of REM and its timing but only with the function of the REM process itself.

General Background

We shall not document here the source of most of our statements since we believe that they are generally accepted by most workers in the field (more extensive references are given in our 1983 paper). Our summary is aimed to underline general points which we feel to be of special importance. We have assumed that the reader is familiar with the methods used to diagnose REM and non-REM sleep.

REM sleep occurs in almost all mammals (the known exceptions are mentioned later) and in most birds, though possibly to a lesser extent. We shall not deal with the difficult topic of sleep in lower vertebrates. The amount of total sleep varies considerably from species to species. An important factor in this determination appears to be the danger to the sleeping animal. Lions sleep many hours a day. Giraffes make do with brief snatches of sleep. Nevertheless, almost all mammals have a reasonable proportion of REM sleep. This is true (as judged by the EEG) even of an animal such as the mole which can hardly move its eyes.

The age distribution for the occurrence of REM is conveniently illustrated by human sleep. Human adults have a total of about one and one half hours of REM sleep per night, whereas newly born babies have approximately eight hours per day. Prematurely born babies have even more, suggesting that a

considerable amount of REM or REM-like sleep occurs in the womb.

Deprivation of REM sleep alone typically produces a large REM rebound. That is, the animal, when allowed to sleep freely after some nights of REM deprivation, has for a time an increased amount of REM.

All these facts make it virtually certain that REM sleep has some important function and that this function is biological in nature and not specifically human. It seems unlikely that the function of REM is solely associated with the psychology of the mature animal. The large amount of REM in the newly born suggests that it also plays some part in the development of the nervous system.

It was originally believed that dreams always occurred during REM sleep and that there were no dreams in non-REM. It now appears that the distinction is not so sharp. Nevertheless, many authorities believe that there is a significant difference between the two states, in that REM dreams are more frequent, more hallucinoid and (if the subject is woken abruptly) more vivid than such dreams as are reported in non-REM.

It seems reasonable to state that most of our dreams are not remembered. The observed facts are that if a person is allowed to sleep freely they report rather few dreams, especially if they wake infrequently in the night. On the other hand, if a person is constantly awoken during REM, they report a relatively enormous amount of dreaming. These facts are usually interpreted along the following lines. It is assumed that, during REM, immediate or very short-term memory storage operates, but that the mechanism for inputting long-term memory is temporarily inactive. On awakening, this latter mechanism slowly recovers. Thus, most dreams are necessarily forgotten, since they are never put into long term memory, unless the dreamer awakes.

This interpretation, if correct, leads to the following important conclusion. It is unlikely that the biological function of REM is related to remembered dreams. That function, whatever it is, is more likely to be related to our host of *unremembered* dreams. There is no evidence to suggest that remembered dreams are anything more than an accidental by-product of this function.

It is more difficult to give an adequate description of the content of REM dreams, since that subject is controversial. It is generally agreed that dreams are frequently visual and that smell occurs very rarely. Very surprising (by everyday standards) events take place, and yet the dreamer typically feels very little surprise.

The characteristic features of REM dreams could be called "bizarre intrusions." These are brief incidents that often appear to arise for no apparent reason. It has long been believed that the content of these intrusions is often related to day residue, that is, to incidents occurring in the last day or so, and especially pertaining to those matters which the dreamer has had on his or her mind. However, it is important to notice that one rarely dreams of

a previous event in correct detail. More typically, the intrusion consists of a mixture of features, all or most of which can be related to events which have occurred recently. These bizarre intrusions seem to follow at short intervals, perhaps every second or so.

The other main feature of the dream appears to be the narrative, which possesses greater continuity than the intrusions. It is as if a part of the brain is trying to make some sense of the bizarre intrusions, as indeed one would if such events happened while one was awake. The narrative often has a particular emotional tone (erotic, filled with anxiety, etc.) which one suspects is related to other causes. It is unclear whether the first few intrusions set up the narrative or whether the narrative exists in some latent form before the first intrusion. When one speaks of "a dream" one is usually referring to a single fairly continuous narrative. We tend to regard such mentation as a separate dream if it has a distinct narrative. As will be seen, our theory provides a good explanation of the nature of the bizarre intrusions. It has nothing useful to say about the narrative.

It has been known for many years that during REM sleep a series of impulses, called PGO (ponto-geniculo-occipital) waves, appear in the brain. These impulses originate in the pons and spread to the neocortex via the thalamus. These waves are very frequent, in the cat, for example, as many as 16,000 per day. Moreover, they affect most cortical neurons in the sensory and motor areas. Hobson and McCarley (1977) have suggested that in part of the pons there is a dream state generator which produces these impulses. They postulate that it is these impulses which provide the driving force for REM dreams. Our views are based on their Activation-Synthesis hypothesis.

We can summarize the main points of this section as follows:

(1) REM sleep performs some important biological function for most higher vertebrates.
(2) This function is likely to benefit the developing animal as well as the mature animal.
(3) While REM dreams may give some clues about this function, it is unlikely that remembered dreams have, in themselves, any major biological use since the majority of dreams are unremembered.
(4) REM dreams are probably influenced by the PGO waves from the pons.

Memory Storage and Neural Nets

It seems likely that any memory system can usefully be described under these headings:

(1) Putting the pattern to be remembered into the system.
(2) Storing it over time.
(3) Accessing the system in order to recover it.

While we have no solidly established theory of human memory, the following rather general account is at least plausible. It is assumed that the first and last processes above—that is, inputting and accessing—necessarily involve neurons firing. For long-term memory storage it is assumed that neuronal firing is not required and that the memory is stored in some semi-permanent modification to parts of neurons and especially to synapses. The exact neurological basis of very short-term memory (immediate memory) is obscure.

It is widely believed that the operation of the brain is radically different from the operation of a modern digital computer. The latter uses accurately pulse-coded messages. Information is encoded in the pattern of 0s and 1s sent out at regular time intervals. This enables a particular message to be sent to a particular "address" where information can be put in, stored and accessed. There appears to be no sign of such a system in the brain.

Instead, the information a neuron sends out appears mainly to be contained in a somewhat irregular pattern of spikes in its axon, probably encoded as the average firing rate, although there may be some information in the firing of each spike relative to those from other relevant neurons. This makes it likely that such a system cannot send precise information to a precise address as a computer can. Instead, it is generally thought that memory in the brain is "content addressable" (see below).

Another important difference between most neurons in the brain and the transistors in a computer is that most neurons receive input from very many different sources and each sends its output to many other neurons. For this reason theorists believe that the operation of the brain can only usefully be modeled by systems in which many units (a unit is an idealized neuron) interact with each other.

Such studies have led to the idea of neural nets, in which many units act in parallel and which, in some models, connect back on all the other units in the net. More elaborate models have several tiers of units in series, connected together in various ways. For a general introduction, see chapters one through four by Rumelhart et al. (1986). These studies show that patterns of activity can be stored in the strengths of the connections (the synapses) between all the different units. Thus, a single associative net, suitability adjusted, if given a pattern of activity A (in which some inputs are firing and some are inactive), will produce as its output a different pattern of activity A′. If a sufficiently large part of the pattern A is used as input, it will output the whole of pattern A′. It is the latter behaviour which is described as "content addressible," since fragments of the memory can be used as an address to call up the total memory.

One can ask where the memory is stored in such systems. As we have said, it is stored mainly in the synaptic weights—the strength of all the connections between the neurons. (Notice if there are n neurons in a net, there are

likely to be n^2 synapses between them.) Memory is not stored exclusively in one or even a few connections, but in the net as a whole. That is, the memory is "distributed." Moreover, if we make a few random alterations in the weights, this usually makes little difference to performance, primarily because each unit has a threshold (as does a neuron). The unit is non-linear, in that it will not fire, or only fire at a very low rate, if its effective input is below this threshold. Thus, the system is "robust."

However, most nets have a further remarkable property. Let us assume that the synapses have been adjusted (and there are usually simple rules of how this should be done) so that with an input A it produces an output A′. Now let us make further adjustments to the system so that a different input B will produce some other output B′. Then, if the net is sufficiently large and if A and B (and A′ and B′) are sufficiently distinct, it is found that the system now can do both jobs. An input A produces A′; input B produces B′.

In both cases the information is contained in the strengths of all synapses, but the information for the two associations (A with A′, and B with B′) is superimposed, so that any one synapse is likely to contribute to both associations. This behaviour is not limited to just a pair of associations. Further ones (C with C′, D with D′, etc.) can be added.

The reader will readily surmise that associations cannot be added indefinitely. For any net there is a limit to what it can store before the system begins to misbehave. Broadly speaking, the larger the net, and the more distinct the patterns are from each other, the more associations can be stored.

When a particular net is overloaded it usually misbehaves in a special way. Instead of outputting the required stored pattern x′, it may produce a pattern which can be seen to be a mixture of several of its stored patterns. This is especially likely to happen if the patterns are not totally distinct but have some parts in common.

To summarize, neural nets can store information in a way which is *distributed*, *robust*, and *superimposed*. When overloaded, such nets often output a pattern which is a mixture of some of its stored patterns, especially if the patterns are somewhat related.

It is important to realize that these properties have not been directly imposed on the net by its designer. They are emergent properties of this general type of memory storage system. Whether these models are significantly similar to the arrangements of neurons in the brain remains to be seen, but they appear to be consistent with much that is known, in a broad way, about the neuroanatomy, the neurophysiology, and the overall behaviour of the brain, although in detail the present generation of models is almost certainly highly oversimplified. It should not be thought that we believe the brain is merely a set of simple associative nets.

Reverse Learning

The process of reverse learning is designed to make the storage in an associative net more efficient. The hope is that this process will reduce somewhat the mixed outputs produced by overlapping memories, while leaving intact the unmixed memories which the net was supposed to store. It may also help to remove inappropriate connections made by the somewhat random nature of neural development.

To understand what has been proposed, the reader must grasp the idea behind the normal storage of memories. Let us take a simple example. Consider a pattern of inputs, but for simplicity let us consider only three of the units participating in this input. Let us assume that in this particular case the first two units are active while the third one is inactive. Thus, the activities of the first two units are correlated, while that of the third is anticorrelated to the first two. Again, to simplify matters, let us consider a very simple net which feeds back on itself so that a given input pattern on its incoming axons produces the identical pattern on the outgoing neurons. The rule proposed is that the connections in either direction between the first two units (since this is a net which feeds back on itself) are strongly positive, while those between the third unit and the first two units are absent or negative (inhibitory). This rule, applied to the whole pattern, is used to decide what strengths to give all the synapses in the net. Synapses connecting units whose firing is correlated in the pattern are made strong. Those whose firing is anticorrelated are made zero or negative. It is this property which allows the whole pattern to be generated if only part of the system is input, since the units active in the part then enlist the cooperation of all the rest of the pattern. Notice that the adjustment is a local process. That is, the synapse only has to know what has happened on either side of it. It can ignore more distant events.

For "reverse learning," exactly the opposite rule is used, though probably not in one large step but by only a small increment. This has the effect that any association produced by the net during the reverse learning process is damped down in the storage system of the net. That is, by the approximate adjustment of synaptic weights the correlations appropriate for that particular pattern are weakened somewhat, and any anticorrelations are strengthened somewhat. It is important that the elements of this undesirable association are not completely erased, but merely damped down. Complete erasure would interfere too much with the storage of the desired associations.

For the process to work it must be repeated many times. A typical protocol would then be as follows. The net is first adjusted (either by the experimenter or by some learning process) so that its synapses store a certain number of desired associations. The normal inputs and outputs are then disconnected and the net is given a *random* input. When the net has given a response, the

synapses between inputs and outputs are then adjusted to reduce this association slightly. The net is then set to zero activity and the cycle is repeated with a different random input. This process is repeated many times.

If a net originally had been given a rather random set of synaptic weights it could easily happen that many different inputs tend to produce one particular output. This is especially likely to happen if many of the inputs share a common pattern of activity, since there will then be a strong correlation between these inputs and all the output elements. We may describe this situation by saying that the net develops an obsession. The process of reverse learning will serve to reduce this obsession, since that output will come up many times in the reverse learning process so that its associations will be gradually weakened in the net as reverse learning proceeds, thus removing or at least reducing the obsession.

It is easy to see, therefore, why these ideas, originally devised to improve the performance of these idealized nets, should suggest that the process of reverse learning goes on in REM sleep. The normal inputs and outputs of the brain are indeed largely cut off. The PGO waves from the brain stem could well be a source of rather "random" stimulation. The outputs of the system (which, if we are woken up, show, at least in part, as hallucinoid dreams) do indeed appear to be mixtures of what the system had previously encountered and stored. That idea accounts for the rather nonsensical nature of the bizarre intrusions, since their character is what one would expect from this theory, as we argue later.

We cannot be sure that PGO waves are completely random in their distribution and indeed our theory does not absolutely require this. Their diffuse nature at least makes it plausible that they are semi-random in nature, that they act as broad stimuli to the contents of the higher levels of the brain rather than as precise detailed instructions. There seems to be a deep reluctance to accept the possibility of random elements in the working of the brain, even in sleep. It should be remembered that in waking life it is the superstitious who see deep meaning in every accidental coincidence. However, the way the brain deals with the results of PGO stimulation—especially as far as the narrative may be concerned—need not be at all random.

However it is described, the process is somewhat paradoxical, since it has the general character of forgetting (a trace in the memory is weakened) and yet improving memory. It does this by separating from each other distinct memories which nevertheless have something in common, so that the system is less confused.

In our original paper (Crick and Mitchison, 1983) we suggested that "we dream in order to forget." This of course is a slogan, not a precise definition. Our hope was that the slogan would act as a mnemonic for the reverse learning process. We have since come to regret this slogan since experience has

shown that people assume we meant that the function of REM is to delete all the elements of our unconscious dreams from memory. As we have explained, this is altogether too strong. If we had to produce alternative slogans, we might suggest, "we dream to reduce fantasy," or, "we dream to reduce obsession."

Of course, this confusion of memories is not always undesirable. It is at the root of imagination, fantasy, poetry, etc. However, reverse learning is unlikely to remove completely the interaction of related ideas and concepts but merely to reduce their interaction, so that, as a result, the system becomes less imaginative and more prosaic in its behaviour. It is unlikely that the process could eliminate fantasy completely.

Reverse learning, as we originally presented it, is a purely automatic process. There is no supervisor to determine which fragments of memory should be damped down and which should be left untouched or strengthened. In this simple model, if such fragments have appeared in the (unremembered) dream, they are all damped down equally. Of course, one could easily produce a much more elaborate theory which involves an intelligent supervisor; but at the moment we see no need for this.

Our theory does not automatically account for recurrent dreams. For this we need a subsidiary assumption. It is known that certain inputs, even during sleep, can awaken the sleeper. Examples are one's name being whispered or a sleeping mother responding to the cry of a baby. Thus, there must be some supervisor, in a loose sense, which controls whether the sleeper wakes up under such circumstances. We have therefore assumed that certain dreams, because of their frightening or threatening nature, or whatever, tend to wake the sleeper. Since such dreams are likely to demand the attention of the newly-awakened person, they are more likely to be strengthened in memory, rather than weakened, as the ordinary learning process takes over after the sleeper has awoken. This, we would argue, makes such dreams more likely to recur in the future. We do not regard this suggestion as wholly satisfactory—the real explanation is probably more elaborate—but we suspect it may be along the right lines.

The Content of Dreams

We are reluctant to discuss the interpretation of dreams because, as is well known, it is difficult to deal with any suggestion on this matter in a scientific manner. One may suggest that a particular explanation is not valid—either because the explanation appears forced or because it may be due to coincidence—and be unable to prove this. Critics who doubt this should explain why, after the better part of a century of intense effort, there is still no agreed-

upon method of dream interpretation, even assuming that such an agreed method was correct. For these reasons we said nothing about the interpretation of dreams in our 1983 paper.

However, a significant argument in favour of our theory depends on the *nature* of bizarre intrusions in dreams. As remarked earlier, these tend to be conflations of recent incidents which often have some bearing upon what the dreamer has had "on his or her mind." Let us see how our theory accounts for these features of dreams.

Freud (1900/1976) emphasized the difference between the *manifest* content of a dream and its *latent* content. In oversimplified terms, the manifest content is what the dreamer reports, while the latent content is the reconstructed meaning of the dream using the manifest content and free associations as a basis for interpretation. The distinction is not completely clear-cut, since even some aspect of the manifest content may involve a certain degree of interpretation. We shall only include such an interpretation when we believe there would be general agreement about it.

Let us leave most aspects of the latent content on one side since they are almost always a matter of opinion. There can be fewer arguments about the manifest content. It is of course true that what is reported is not necessarily a true and exact recall of the surface content of the dream and may, for all we know, be systematically distorted in some way; but, it is difficult to argue that all accounts of dreams are simply inventions to amuse and deceive the listener or reader. Moreover, we are not concerned here with the exact details of this or that dream but with the general nature of the manifest content.

Now, there is one aspect of the manifest content about which there is little dispute. It has been known for a hundred years that many of the features which appear in dreams are *mixtures* of objects or events that the dreamer had experienced and stored in memory. Freud (1900/1976) called the process which produced such mixtures "condensation" and gave many explicit examples in *The Interpretation of Dreams*, especially in section VI(A) entitled "The Work of Condensation." For instance, Freud pointed out that in his well-known "Irma's dream" the woman in the dream was a mixture of Irma and several other women. Moreover, the earliest dreams reported by young children (Foulkes, 1985) show condensation even though a narrative is lacking.

An even more striking example is given by Freud in a section of his historical introduction (section I[B]). In one of his own dreams Freud conflated the faces of two men, one a doctor in Freud's native town, the other one of the masters at his secondary school. Freud continues:

> When I woke up I could not discover what connection there was between the two men. I made some enquiries from my mother, however, about this doctor who dated back to the earliest years of my childhood, and learnt that he had only one eye. The school-

> master, whose figure had covered that of the doctor in the dream, was also one-eyed. (1900/1976, p. 76)[1]

Freud's account does not state explicitly whether the face he saw in his dream had only a single eye (though he rather implies that it did); but even if this was not the case we feel that one-eyed men are sufficiently rare that it is unlikely that this putative common feature was merely a coincidence. The important point is that in condensation the objects or events brought together always turn out to have some feature in common. As we have seen, this is exactly what happens when a Hopfield net is overloaded, as indeed Hopfield, Feinstein, and Palmer (1983) pointed out.

There are a host of other possible examples, although in some cases the common feature may be a matter of interpretation (and therefore open to greater argument). Nevertheless, condensation in dreams is such a frequent and universally attested phenomenon that it demands an explanation of some sort or other. It is not always obvious whether the different elements comprising the composite object or event occur strictly simultaneously or in very close succession to each other. The brain is obviously not a static processor, as is a Hopfield net, so we do not feel that at this point this distinction is likely to upset our general argument.

We should state briefly why we are so reluctant to discuss the latent content of dreams. Our reluctance derives from the common human failing of confabulation. This often occurs in post-hypnotic suggestion, as Freud was well aware. A striking modern example of confabulation has been given by Gazzaniga and Le Doux (1978), involving a split-brain patient (a young man whose corpus callosum had been cut for medical reasons). The experimenter was quite certain that the reason given by the left hemisphere for the movement of the hand controlled by the other hemisphere had absolutely no foundation in fact, although the left hemisphere gave these reasons without reservation. A particular example should make this clear. (Recall that the left hand and the left visual field are connected to the right hemisphere.)

> When a snow scene was presented to the right hemisphere and a chicken claw was presented to the left, P.S. quickly and dutifully responded correctly by choosing a picture of a chicken from a series of four cards with his right hand and a picture of a shovel from a series of four cards with his left hand. The subject was then asked, "What did you see?" "I saw a claw and I picked the chicken, and you have to clean out the chicken shed with a shovel."
>
> In trial after trial, we saw this kind of response. The left hemisphere could easily and accurately identify why it had picked the answer, and then subsequently, and without batting an eye, it would incorporate the right hemisphere's response into the framework. While we knew exactly why the right hemisphere had made its choice, the left hemisphere

[1]Published with the permission of Allen & Unwin, London.

> could merely guess. Yet, the left did not offer its suggestion in a guessing vein but rather as a statement of fact as to why that card had been picked. (Gazzaniga and Le Doux, 1978, p. 148)

Thus, we have no confidence that any interpretation of the latent content of a dream actually corresponds to the real reason the dream occurred.

Further, a dream sequence need not necessarily be explained by some deep insight into the sleeper's mind. The sequence may signify little more than a series of accidental coincidences between items in the previous experience of the subject. This is not to say that one may not learn something about the dreamer's mind from the subject's own interpretation of the dream, or from the subject's reaction to a Rorschach pattern. As this volume testifies, it is more sensible, at this stage, to try to *describe* dreams better, and in particular, in cognitive terms, rather than attempt to *interpret* them.

There appear to be certain cognitive deficiences in dreams as well as sensory deficits, such as the absence of smells. It is well known that a dreamer often lacks certain insights—dreaming about dead relatives, for example, without realizing they are dead. Can a dreamer truly read a written text in a dream? As the nature of cognition in REM sleep becomes clearer it may be possible to provide neurobiological explanations for cognitive peculiarities of dreaming. However, we are concerned in this chapter with the biological reasons for the existence of REM sleep rather than with cognition itself. We feel that the usual anecdotal descriptions of the latent content are of little value to researchers. Whether a physician will eventually be able to use a patient's reported dreams as a reliable source of medical information remains to be seen. At the moment the "silent testimony of the medical profession" does not support this type of diagnostic procedure. A doctor rarely says, "We'd better take a sample of his dreams or he might sue us."

Criticisms of the Theory

Criticisms of the Theoretical Models

It must be conceded, straight away, that the models of neural nets, used to demonstrate that reverse learning can be effective, are grossly oversimplified. For example, the particular neural net originally devised by Hopfield (1982) has the following unrealistic features. The units do not produce spikes in their axons. Rather, each one is in one of two states: positively active or negatively active. The same neuron can, at different times, produce either activation or inhibition at the same synapse. As far as we know this does not occur in the brain. In Hopfield's model any synapse can have either a positive or a negative strength. Moreover, a negative input onto a synapse with a negative

weight produces a positive effect (since $(-)\times(-)=+$) on the postsynaptic unit. All this is highly unrealistic. Moreover, the original net was a forced-choice net. No matter what the input, some output resulted. The net never relaxed into zero activity as more realistic nets can be made to do. In Hopfield's model, the connection from one neuron to another always has exactly the same strength as the reverse connection from the second to the first neuron. This feature is imposed to provide the net with an "energy" function, and is probably not crucial to the model. Another limitation of the model is that a Hopfield net is a static processor. Given an input it produces a single output, not a sequence of distinct outputs as a more elaborate model might do.

The nets used by Clark, Rafelski, and Winston (1985) have units which are somewhat more realistic, although they possess the property that all operations are made simultaneously to all the neurons whereas Hopfield's model makes operations asynchronously, in a random order. Synchronous adjustments are more apt to produce oscillations than are asychronous adjustments, but, as Clark et al. are primarily interested in cyclic activity in the net as a possible basis for short-term memory, they do not regard this as a disadvantage. Moreover, the nets they have explored so far have been set up with mainly random connections. No biologist is likely to regard their nets as truly realistic. However Clark et al. (1985) do show that "brainwashing" (loosely, their term for reverse learning) helps their nets avoid catastrophic behaviour. This result may therefore be relevant to processes needed to reduce the effects of the many semi-random connections made while the brain grows. It should also be noted that Hinton and Sejnowski (1983), in their studies of "annealing" in so-called Boltzmann nets (those with an energy function and a stochastic process equivalent to temperature), have provided theoretical reasons why a free-running reverse learning process forms an important component of their learning algorithm. Thus, while it is probably the opinion of most of the theoretical workers in the field that reverse learning (or brainwashing) is likely to assist many types of theoretical nets to perform better, it may be admitted that a few more examples, and in particular a few more realistic examples, would help to buttress this belief.

There are several technical criticisms arising out of the demonstration of reverse learning by Hopfield et al. (1983). The first is that it is unclear exactly how many distinct random inputs are required to "cover the space," and whether this number is adequate for the function we have postulated in REM sleep. The reason that many inputs are used to test the system is that only in this way can one hope to cover all the various possible states which the net can be in. The net Hopfield et al. used was quite small—only 32 units. It is by no means obvious that the number of distinct PGO waves are sufficient to explore all the previous overlaps which the changing human brain, with its vast number of neurons, is likely to contain. Of course, one should

regard the REM process as a continuing one, not being limited to just one night's sleep, so that a mixed response which was missed by the random inputs on one night might be caught in the next one. In a Hopfield net, many different inputs seem to be needed to test the system. Because of this each quantum of reverse learning was made very small, otherwise the memories originally stored would also be erased.

This brings up another difficulty in the observed behaviour of the Hopfield net, that is, the fact that a fair proportion of the outputs to the random inputs were *not* mixed memories, but stored ones, and this proportion naturally became higher as reverse learning proceeded and the mixed states were damped out. This suggests that in REM sleep the brain is put into a condition such that mixed responses (fantasies) are more common, possibly induced by a reduction of the activity of some or all of the inhibitory neurons. This may serve as an explanation of the remarkable fact that the blood flow in the brain appears to be greater in REM sleep than when the brain is awake. Another alternative is that the shock nature of the PGO waves, perhaps because they are so unlike the normal somewhat structured inputs to the brain, may tend to promote such mixed responses. Whatever the reason, it seems to be a characteristic of bizarre intrusions that they are never, or hardly ever, true recollections, but appear more like mixtures of separate but related items, often of fairly recent occurrence. All this leads us to conclude that our confidence in neural nets, as models for operations in REM sleep, would be increased if nets could be found which more closely mimicked the nature of our dreams.

Experimental Criticisms

The most powerful criticism of the idea of reverse learning is that deprivation of REM sleep does not produce the effects the theory would predict. Because our brain is so very complicated it is not clear exactly what these predictions should be—although one might reasonably expect an increase in imagination and fantasy if REM sleep is prevented. In a child whose brain is still developing one might predict an increase in obsessions. In experimental animals one would not be surprised if REM deprivation appeared to make the cortex more excitable. REM deprivation might also be expected to make subjects' performance on memory tests somewhat confused.

There are hints that some of these effects do occur. It is claimed (Cartwright and Ratzel, 1972) that for certain individuals REM deprivation does indeed produce more fantasy. There certainly are experiments (Cohen, Thomas, and Dement, 1970) which demonstrate that the electro-convulsive threshold of the cortex of the cat was lowered after REM deprivation. Unfortunately, the literature on the effects of REM deprivation upon memory,

tested in one way or another, is confused, and we shall not attempt to come to grips with this topic here. By and large, however, the general opinion is that when people or animals are deprived of REM their behaviour is not affected in any obvious or reproducible way.

If this were true, then *any* theory about the function of REM would have to face this criticism. Moreover, taken by itself, this finding suggests that REM sleep may serve no important purpose. This is in clear contradiction to the biological evidence, outlined earlier, which suggests exactly the opposite. For this reason one must look rather critically at the experimental evidence.

There are two main ways to deprive a subject of REM. One is to wake the subject up every time they enter REM sleep. The other is by the use of drugs. There are severe limitations to the first method. It is not easy to wake someone as soon as they enter REM. After several nights of REM deprivation the effects of REM rebound are so great that one may have to awaken the subjects fifty times a night. Some subjects may refuse to continue with the experiment. For these and other reasons, very few subjects have been deprived of REM for more than a week. Moreover, even in these cases it is suspected that subjects may be "snatching" fragments of REM during the night or day. If REM deprivation takes some time to demonstrate its effects (and the theoretical models at least hint that this might be the case, since very many PGO waves may be needed to "cover the space"), those effects may have been missed.

On the other hand, many drugs seem to reduce REM, and certain drugs used for treating depression can abolish REM entirely for many weeks (Wyatt, Fram, Kupfer, and Snyder, 1971). (Indeed it has been claimed that if a depressed person is deprived of REM by waking them up, their depression is reduced.) This evidence again poses difficulties for any theory of REM. However, it should be noticed that (mainly for ethical reasons) there have been few, if any, cases in which normal subjects have been totally deprived of REM by means of drugs. Any drug is likely to have side effects which are apt to confuse the interpretation of its actions. For all those reasons, we do not feel that experiments on REM deprivation constitute decisive evidence against our theory. Nevertheless, we would much prefer it if better experiments on REM deprivation, conducted either upon people or upon animals, did show systematic psychological or physiological defect, since we find it difficult to believe, in the face of all the biological evidence, that some such defect does not exist.

Nature's Experiments

Another approach to the study of REM is to turn to nature's experiments. In our original paper we mentioned only one mammal which had been shown

not to have REM sleep. This is the *Echidna* (an Australian monotreme), the Spiny Anteater. The *Echidna*, for a monotreme, has an abnormally large neocortex for its body size. However, considering that one way to avoid overlaps in a net is to make it larger, with more units, we thus wondered whether this might explain the relatively large size of the *Echidna's* brain. It was therefore natural for us to ask whether there were any other mammals with unexpectedly large brains. The most obvious example is *homo sapiens*, but it seemed more reasonable to attribute the size of our brain to our exceptional cognitive ability. The other example was the *Cetacea*, the dolphins and whales. Since whales are not the most convenient experimental animals to study, we wondered whether dolphins had been shown to have REM sleep. Inquiries failed to trace any experiments in the western world, but thanks to Dr. Theodore Bullock, we learned that experiments on sleep in dolphins had been in progress for some time in the U.S.S.R. A recent review (Mukhametov, 1984) presents the results of this research in English. In brief, two species of dolphins have been tested, the bottlenosed dolphin and the porpoise. Both showed non-REM sleep (though their sleep possessed certain unusual features), but in no case (about thirty animals in all) were there any signs of REM. Nor could any PGO waves be detected. It might be thought that this was solely a phenomenon characteristic of sea mammals, but parallel studies on two species of seals demonstrated normal REM.

These results are of considerable interest biologically. An important difference between dolphins and seals is that the former never have to support their heads out of water. Thus, there is less evolutionary pressure for them to have small heads. The results hint rather clearly that the function of REM is to make advanced brains more efficient and, in particular, to allow these brains to have a smaller size than they would otherwise have. This conclusion, if true, would be broadly compatible with our theory, but it cannot be said to provide very strong support. We feel our theory deserves one mark (but not more) for "predicting" this effect.

It would clearly be useful to look for other species in which REM is absent. An obvious candidate is the duck-billed platypus. This is not an easy experimental animal to study, although modern telemetering methods might overcome these limitations. More types of small *Cetacea* (such as the killer whale) could usefully be studied. Perhaps some heroic experimenter might apply telemetering to one of the larger whales.

Birds, because of flight, have strong evolutionary pressures to keep their weight down. One might ask whether any of the large flightless birds have lost REM, especially if they evolved on isolated islands with an absense of the usual predators. Did the Dodo dream? Alas, we shall never know, but it would be useful to have a sound estimate of the size of its brain in relation to its body weight.

The Side Effects of REM

In spite of the great activity in the brain in REM sleep there is very little output to the musculature. It is generally agreed that this is due to inhibition from the brainstem acting in the spinal cord. It is perhaps not surprising that this inhibition is not absolutely complete. The extremeties of our limbs may twitch a little from time to time. It is rather more unexpected that eye movement is not inhibited at all—although this may be because eye movement causes no problems to the sleeper. The nervous impulses which produce rapid eye movements probably come directly from the brainstem, rather than by the more circuitous path from brainstem to cortex and then back again to brainstem.

An effect which appears to be correlated with the REM state is penile erection in males, though this may persist for a short time after the sleeper awakes. It is not solely correlated with erotic dreams. There seems no obvious reason why erection should not also be inhibited in sleep. One may wonder whether this phenomenon might have selective advantage in evolution. It would not, after all, be too surprising if this particular state of male readiness for sexual intercourse contributed to the number of offspring fathered. However, this assumes that the female of the species will be in easy proximity to the male during or shortly after sleep. One wonders, therefore, if REM penile erection has been observed in animals such as tigers, which are solitary except for brief periods of sexual intercourse, as opposed to lions, which constantly sleep in close proximity to each other.

Some Alternative Theories

We shall not attempt to review here all the many theories which have been proposed to explain REM and the nature of dreams. There seems little point in dwelling on Freud's ideas since they hardly fit the biological data. Freudians have no unforced explanation for the large amount of REM in the unborn and newly born and the occurrence of REM in so many different species of mammals and birds. To a modern neuroscientist Freud's theories, in spite of their appeal to the contemporary imagination, seem little better than the common belief in earlier times that dreams foretold the future, a belief which also held strong intuitive appeal.

Other general theories of REM function seem hardly more plausible. For example, Jouvet (1980) has suggested that REM dreams help to tune up instinctive behavior laid down crudely in epigenesis, an idea originally proposed by Maeder in 1912 (cited in Freud, 1900/1976). The actions taking place in dreams are supposed to help this tuning up process by acting out such behavior. This seems to us very unlikely, since to do this effectively it would

seem essential that there be considerable muscular output in order to receive feedback from the environment. Moreover, there is already a mechanism, highly developed in mammals, which seems to perform just this function in the awake animal, namely, play.

We shall refer briefly to three more recent suggestions concerning REM since they are somewhat related to our own approach. Evans and Newman (1964) proposed a function for REM based on the analogy of the brain as a modern digital computer. These authors suggested that during sleep the brain was "off line" and could use that time to reprogram the events of the day, to remove redundant material, etc. Although we believe that there are important differences between the brain and a digital computer, one might accept this suggestion as an extremely loose analogy. However, our own suggestions are based on a rather different model, the associative net, which stores information, as we have explained, in a radically different way from digital computers. Our explanation of the nature of dreams (or at least the bizarre intrusions in them) depends crucially on the distributed and superimposed nature of that storage. Thus, we feel that the proposals of Evans and Newman, though superficially somewhat similar to ours, are somewhat different.

Another recent suggestion is made by Davis (1985). He poses the problem of how our memories are maintained over long periods of time in the face of relentless molecular turnover. Davis suggests that in sleep our memories are relearned, so that decaying synaptic strength can be renewed.

We are reluctant to accept this idea for several reasons. In the first place, we feel that the problem of molecular turnover is more likely to be avoided because the synapse, or elements in it, will turn out to have a *cooperative* structure. (See the recent discussion by Crick [1984] on this topic.) Second, if Davis is correct, then the *Echidna* and dolphins, which lack REM, should have poor long-term memories. We doubt that this is so (there is no folk-saying "a dolphin always forgets"), but the hypothesis could be tested. Third, the present authors find it hard to reconcile the character of REM dreams with the suggestion that they reinforce memory. Such memories seem to us a very odd choice to reinforce. We feel our suggestion that these bizarre intrusions are somewhat weakened in the memory store makes more sense.

However, it is possible that the other type of dream, more typically found in non-REM sleep, may be part of a memory consolidation process, especially for more recent events. It will be recalled that in normal circumstances a period of REM is preceded by a period of non-REM, and it would not be unreasonable if recent memories were "consolidated" before they were "edited" by the postulated reverse learning process. However, consolidation seems less likely to be able to deal effectively with the longer-term rehearsal of memory needed to counteract molecular turnover. We mention finally the suggestion of Clark et al. (1985) that their "brainwashing" process occurs in non-REM sleep fol-

lowed by a period of positive learning in REM. Again, we feel that the nature of REM dreams makes this proposal less likely than ours.

Recapitulation and Summing Up

There is one point which is so familiar to sleep workers that we have not mentioned it thus far: the very odd and unexpected nature of REM sleep, as shown by the term "paradoxical sleep." It is remarkable that for part of the night the human brain should appear, from viewing the EEG, as if it were wide awake, and yet, the sleeper, as judged by their unresponsiveness as well as by the relaxation of many of their muscles, seems to be in an especially deep sleep. This demands some special explanation. As we have said, the wide distribution of REM sleep among mammals and birds suggests that any explanation must be a biological one, and not something peculiar to human beings.

We believe our theory is attractive because it starts with very general assumptions about the manner in which the brain is wired and the way it operates. The fact that most neurons receive many distinct inputs, that an axon typically branches extensively to form many distinct synapses—some of which may feed back onto similar neurons, together with the absence in neuronal firing of any obvious form of pulse code—certainly suggests that the brain is rather unlike a digital computer. It appears more likely that memory is stored in the brain in a radically different way. Simple models of associative nets show that such a memory store can be distributed, superimposed, robust, and content-addressible. Overloading such a net often produces outputs which are a combination of stored associations. These combinations are most likely to occur if some of the stored associations have something in common. Thus, these rather naive theoretical ideas show that mixed outputs are likely to occur as an emergent property, and not merely something specifically imposed on the initial design of the net. This idea, together with the semi-random nature of PGO waves, is the core of our interpretation of REM dreams.

The fact that some nets can be made to behave in a more orderly way by the semi-automatic process we have called "reverse learning," is the basis for our suggestion concerning what happens to the brain during REM sleep. However, the undoubted value of some degree of fantasy and imagination suggests that the process should not remove all cross-associations in the memory store. Because of the way the brain develops—the connections being dictated in a rough and ready manner by the animal's genes but otherwise made in a semi-random way and then tuned up by experience—the "reverse learning" mechanism might also be used to remove "obsessional" behaviour in the developing brain.

Thus, our theory, starting from simple assumptions about the general way the brain works, explains many of the major phenomena related to REM sleep. What it does *not* explain is the narrative aspect of REM dreams. From our waking behaviour we know that there are processes in our brain which try to make sense of a succession of bizarre happenings—and it is not unreasonable that these also operate in REM sleep—but to explain them in neural terms requires a more elaborate theoretical model than our simple associative net. Our theory would, of course, predict that REM deprivation should produce psychological deficits of one sort or another, and the absence of any clear demonstration of such deficits is unexplained by our proposal, although we do not feel the evidence is solid enough at this stage to discount our model. In passing, we should note that all competing models account for the facts somewhat less well than ours does. However, the absence of REM in certain mammals, together with the unexpectedly large size of their neocortices, is quite compatible with reverse learning and suggests an important idea independent of our theory, namely, that the function of REM, whatever it may be, is to allow an animal to have a more efficient brain and, in particular, a rather smaller one than it would if REM were lacking. This suggestion appears to have been widely overlooked.

Nevertheless, we view our theory as a tentative first step towards developing a scientific model of the function of REM sleep and of the interpretation of dreams. Its basic defect is easy to see: we attempt to explain the complex behaviour of a complicated structure by an extremely simple model. The strength of such an approach, however, is also its weakness. It is practically impossible to test such a theory decisively at the macroscopic level. The brain is so complicated that conflicting evidence is all too easy to dismiss, as psychoanalysts have unwittingly demonstrated for the better part of a century.

However, our theory is not intrinsically untestable. Being essentially a micro theory, any decisive test must be conducted partly at the micro level. In our case this involves knowing how synapses and neural thresholds are modified in REM sleep. At the moment, however, we do not know just how synapses are modified while we are awake, though much experimental effort is now being devoted to that problem.

In time, it should be possible to refute or confirm our ideas in an unambiguous way. Meanwhile, we think it might be useful if brain researchers bear the idea of "reverse learning" in mind. We believe this construct to be sufficiently attractive to make worthwhile the small effort needed to understand it more properly. Researchers at many levels, from those studying the nature and the interpretation of dreams, to molecular biologists working on synapses, as well as theorists devising models of the brain, could usefully remember that "reverse learning," or some process similar to it, might be occurring in REM sleep.

References

Cartwright, R.D., and Ratzel, R.W. (1972). Effects of dream loss on waking behaviors. *Archives of General Psychiatry, 27,* 277–280.

Clark, J.W., Winston, J.V., and Rafelski, J. (1984). Self-organization of neural networks. *Physics Letters, 102A,* 207–211.

Clark, J.W., Rafelski, J., and Winston, J.V. (1985). Brain without mind: Computer simulation of neural networks with modifiable neuronal interactions. *Physics Reports, 123,* 216–273.

Cohen, H., Thomas, J., and Dement, W.C. (1970). Sleep stages, REM deprivation and electroconvulsive threshold in the cat. *Brain Research, 19,* 317–321.

Crick, F. (1984). Memory and molecular turnover. *Nature, 312,* 101.

Crick, F., and Mitchison, G. (1983). The function of dream sleep. *Nature, 304,* 111–114.

Davis, B.D. (1985). Sleep and the maintenance of memory. *Perspectives in Biology and Medicine, 28,* 457–464.

Evans, C.R., and Newman, E.A. (1964). Dreaming: An analogy from computers. *New Scientist, 419,* 577–580.

Freud, S. (1976). *The interpretation of dreams.* New York: Penguin Books. (Originally published 1900)

Foulkes, D. (1985). *Dreaming: A cognitive-psychological analysis.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Gazzaniga, M.S., and Le Doux, J.E. (1978). *The integrated mind.* New York and London: Plenum Press.

Hinton, G.E., and Sejnowski, T.J. (1983). Optimal perceptual inference (pp. 448–453). *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Washington, D.C.

Hobson, J.A., and McCarley, R.W. (1977). The brain as a dream state generator: An activation-synthesis hypothesis of the dream process. *American Journal of Psychiatry, 134* (12), 1335–1348.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79,* 2554–2558.

Hopfield, J.J., Feinstein, D.I., and Palmer, R.G. (1983). "Unlearning" has a stabilizing effect in collective memories. *Nature, 304,* 158–159.

Jouvet, M. (1980). Paradoxical sleep and the nature-nurture controversy. In P.S. McConnell, G.J. Boer, H.J. Romijn, N.E. van dePoll, and M.A. Corner (Eds.), *Adaptive capabilities of the nervous system* (pp. 331–346). Amsterdam: Elsevier/North-Holland Biomedical Press.

Mukhametov, L.M. (1984). Sleep in marine mammals. *Experimental Brain Research, Supplement, 8,* 227–238.

Rumelhart, D.E., McClelland, J.L., et al. (Eds.) (1986). *Parallel distributed processing, volume 1: Foundations.* Cambridge: MIT Press/Bradford Books.

Wyatt, R.J., Fram, D.H., Kupfer, D.J., and Snyder, F. (1971). Total prolonged drug-induced REM sleep suppression in anxious-depressed patients. *Archives of General Psychiatry, 24,* 145–155.