Taylor & Francis
Taylor & Francis Group

Connection Science

ISSN: 0954-0091 (Print) 1360-0494 (Online) Journal homepage: https://www.tandfonline.com/loi/ccos20

Consolidation in Neural Networks and in the Sleeping Brain

Anthony Robins

To cite this article: Anthony Robins (1996) Consolidation in Neural Networks and in the Sleeping Brain, Connection Science, 8:2, 259-276, DOI: 10.1080/095400996116910

To link to this article: https://doi.org/10.1080/095400996116910

Published online: 01 Jul 2010.

Connection Science, Vol. 8, No. 2, 1996, 259–275

Consolidation in Neural Networks and in the Sleeping Brain

ANTHONY ROBINS

(Received for publication April 1996)

In this paper we explore the topic of the consolidation of information in neural network learning. One problem in particular has limited the ability of a broad range of neural networks to perform ongoing learning and consolidation. This is 'catastrophic forgetting', the tendency for new information, when it is learned, to disrupt old information. We will review and slightly extend the rehearsal and pseudorehearsal solutions to the catastrophic forgetting problem presented in Robins (1995). The main focus of this paper is to then relate these mechanisms to the consolidation processes which have been proposed in the psychological literature regarding sleep. We suggest that the catastrophic forgetting problem in artificial neural networks (ANNs) is a problem that has actually occurred in the evolution of the mammalian brain, and that the pseudorehearsal solution to the problem in ANNs is functionally equivalent to the sleep consolidation solution adopted by the brain. Finally, we review related work by McClelland et al. (1995) and propose a tentative model of learning and sleep that emphasizes consolidation mechanisms and the role of the hippocampus.

KEYWORDS: Consolidation, sleep, learning, hippocampus, catastrophic forgetting, catastrophic interference, rehearsal, pseudorehearsal.

1. Introduction

The topics of transfer and consolidation within learning are closely related. Transfer involves identifying and using information from one task or learning system to facilitate the performance of another learning system. Consolidation focuses on the way in which new information can be successfully integrated into existing information within a given learning system. The consolidation of information is an essential requirement both for many forms of transfer, and for long-term or ongoing 'lifelong' learning. In this paper, we explore the topic of consolidation in artificial neural networks (ANNs).

Existing learning algorithms in many of the most common ANNs do not accommodate consolidation. New information when it is learned significantly disrupts information previously learned by the network. This problem has been

A. Robins, Computer Science Department, University of Otago, PO Box 56, Dunedin, New Zealand. E-mail: coscavr@otago.ac.nz.

0954–0091/96/020259–17 $7.50 © 1996 Journals Oxford Ltd

called 'catastrophic forgetting' (or 'catastrophic interference'). Robins (1995) proposed solutions to the catastrophic forgetting problem based on 'rehearsal' and 'pseudorehearsal' mechanisms that form a possible framework for modelling consolidation. In this paper, we will review and slightly extend the illustration of these mechanisms. No new theoretical material is introduced in this review, but all simulations presented use a structured 'real world' population rather than the randomly constructed populations used in Robins (1995). The main focus of the current paper is to relate these mechanisms to the consolidation processes which have been proposed in the psychological literature regarding sleep.

There are many theories on the function of sleep, ranging from the physiological to the psychoanalytic (see Moffit and Kramer (1993) or Empson (1993) for recent overviews of research on sleep and dreaming). Sleep is a complex and multifaceted phenomena. For our purposes we are interested in exploring one particular aspect of the possible function of sleep, the 'sleep consolidation hypothesis', as proposed for example by Winson (1990). This hypothesis holds that newly acquired information is integrated into existing long term memory during sleep. We explore the relationship between pseudorehearsal and the sleep consolidation hypothesis, identifying many common features. We also summarize related work by McClelland *et al.* (1995) and combine this with our suggestions to propose a (tentative) model of learning and sleep that emphasizes consolidation mechanisms and the role of the hippocampus.

Although these proposals differ from other suggestions linking ANNs and sleep (e.g. Globus, 1993; Hinton & Sejnowski, 1986; Hobson *et al.*, 1992; Hopfield *et al.*, 1993; Sutton *et al.*, 1992), they are not necessarily exclusive of these alternative accounts. A range of interactions is certainly possible between ANN mechanisms and the complex processes of sleep and learning.

In the following two sections, we summarize and present simulations of the catastrophic forgetting effect, and the rehearsal and pseudorehearsal mechanisms respectively. Section 4 outlines the sleep consolidation hypothesis and Section 5 relates pseudorehearsal mechanisms to sleep consolidation. In Section 6, we review related work by McClelland *et al.* (1995) and propose a model of learning and sleep that emphasizes consolidation mechanisms and the role of the hippocampus.

2. Consolidation and Catastrophic Forgetting in ANNs

Most neural network learning algorithms involve 'one-shot' learning—a population is presented and trained over a number of iterations. One problem in particular has prevented the use, in some of the most common neural networks, of long-term learning methods capable of the ongoing consolidation of new information (new patterns) into existing information (the population of patterns already learned by the network). This problem is 'catastrophic forgetting', the tendency for new information when it is learned to significantly disrupt old information.

In this section, we briefly describe and illustrate the catastrophic forgetting effect, and in Section 3 we present solutions based on rehearsal and pseudorehearsal. These sections briefly summarize material presented in Robins (1995), which took as a starting point experiments described by Ratcliff (1990). While no new theoretical material is introduced, all simulations presented use a structured 'real world' population rather than the randomly constructed populations used in our earlier paper.

All simulations presented in this paper use the 'Iris' data set, which is well known in the machine learning literature and available from the UCI machine learning database (Murphy & Aha, 1994). This Iris data set consists of 150 items divided into three classes (distinct species of iris) of 50 items each. Each item consists of four real valued measurements of the iris (such as petal length) which in our version of the data set have been scaled to lie in the range 0 to 1. We use a 4:3:4 autoassociative backpropagation network to learn the populations described in each simulation, with a learning constant of 0.05 and a momentum constant of 0.9.

Robins (1995), following Ratcliff (1990), used a 'goodness' measure of the accuracy of learning, with a goodness of 1 equivalent to a perfect match of output and target. Note that in this paper, however, we return to the use of a more standard error function (the average of the sum of the squared errors between output and target vectors) with an error of 0 representing a perfect match. Training is considered complete when the average error falls below a criterion of 0.01. Finally, the results reported for each of the simulations in this paper are averaged over 50 individual replications (using different populations for each replication).

2.1. Catastrophic Forgetting

Ideally the representations developed by neural networks should be 'plastic' enough to change to adapt to changing environments and learn new things, but 'stable' enough so that important information is preserved over time. Unfortunately these requirements are in conflict; stability depends on preserving the structure of representations, plasticity depends on altering it. One consequence of a failure to address this 'stability/plasticity dilemma' (Carpenter & Grossberg, 1988; Grossberg, 1987) in many neural networks is excessive plasticity, usually called 'catastrophic forgetting' (or 'catastrophic interference'). Catastrophic forgetting can be summarized as follows:

> If after its original training is finished a network is exposed to the learning of new information, then the originally learned information will typically be greatly disrupted or lost.

Catastrophic forgetting is an implausible aspect of ANN-based cognitive models and prevents the use of long-term learning to consolidate information. It is also undesirable in practical terms, making it very difficult to modify or extend any given ANN application without completely retraining the network.

While stability/plasticity issues are very general, the term 'catastrophic forgetting' has tended to be associated with a specific class of ANNs, namely static networks employing supervised learning. This broad class includes probably the majority of commonly used and applied networks, such as the very influential backpropagation family and Hopfield nets. A number of recent studies have used these kinds of ANNs to highlight the problem of catastrophic forgetting and explore various issues—these include Nadal *et al.* (1986), McCloskey and Cohen (1989), Hetherington and Seidenberg (1989), Ratcliff (1990), Burgess *et al.* (1991), Lewandowsky (1991), French (1992, 1994), McRae and Hetherington (1993), Lewandowsky and Li (1995), Sharkey and Sharkey (1994a,b) and Robins (1995). For a discussion of those ANNs not affected by catastrophic forgetting, and therefore capable of long-term consolidation, see Robins (1995).

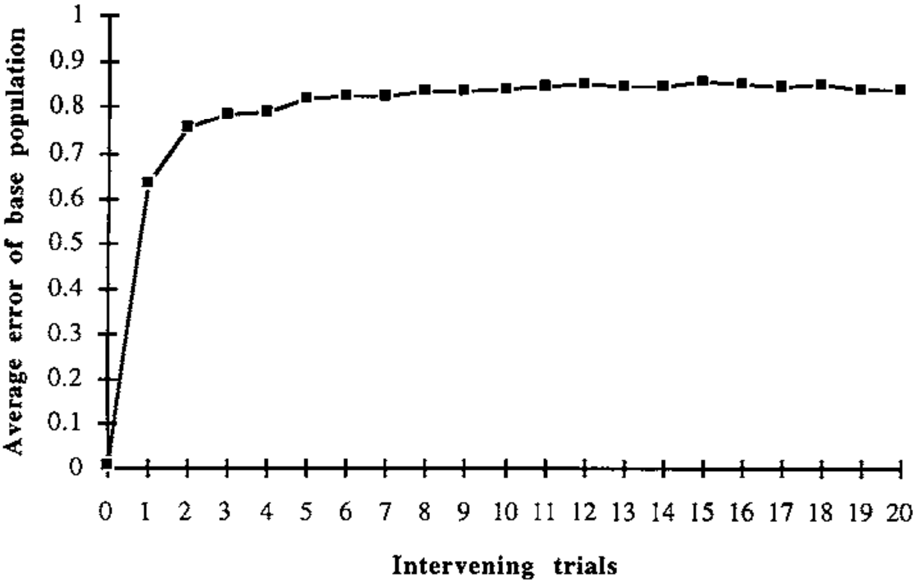

Figure 1. Results from the first simulation: the basic catastrophic forgetting effect.

2.2. Illustrating Catastrophic Forgetting

In the basic illustration of catastrophic forgetting (following Ratcliff (1990)) a backpropagation network is used to learn a 'base population' of items (input/output pairs of vectors) in the usual way (Rumelhart *et al.*, 1986). Subsequently a number of further items are learned one by one; these are the 'intervening trials'. The effect of the intervening trials can be illustrated by plotting a measure (such as goodness or error) of the ability of the network to correctly reproduce the base population after each intervening item.

Our first simulation follows exactly this pattern using the Iris data set and autoassociative network described above. The network was trained to criterion on a base population of 30 examples of one species of iris. At this point (after no intervening trials) the average error of the base population is 0.01. Subsequently, the network was trained one at a time on 20 intervening items drawn from a second species of iris. As shown in Figure 1, after just a single intervening item the error of the base population increases dramatically. The ability to produce the base population outputs accurately continues to decline as more and more intervening items are learned. This illustrates both the form and severity of the classic catastrophic forgetting effect; new information significantly disrupts or even completely wipes out information previously learned by the network.

2.3. Influences on the Extent of Forgetting

While catastrophic forgetting is in general a significant problem, the extent to which the effect occurs for any given population is influenced by a number of factors. Sharkey and Sharkey (1994a,b) provide a useful overview of several practical issues, noting that catastrophic forgetting occurs most significantly in cases where training is sequential and without negative exemplars. French (1992) suggests that the extent to which catastrophic forgetting occurs is largely a

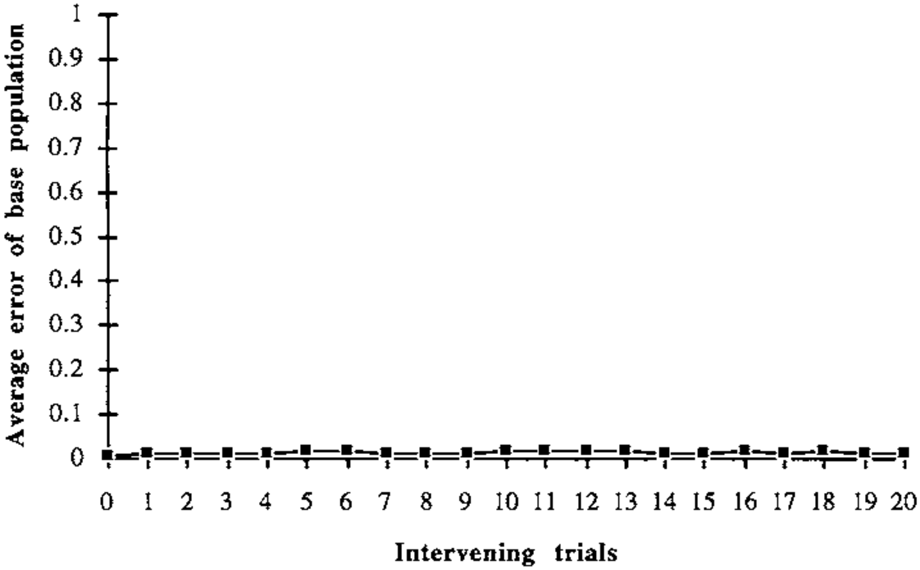

Figure 2. Results from the second simulation; catastrophic forgetting does not occur when intervening items are consistent with the base population.

consequence of the overlap of distributed representations and can be reduced by reducing this overlap.[1] Catastrophic forgetting will be worst when similar inputs, generating similar hidden unit patterns, require very different output patterns to be produced.

Significantly, catastrophic forgetting will not necessarily occur when the intervening items fit consistently into the structure of the base population already learned by the network. In our second simulation we illustrate this point. As before, the network was trained to criterion on a base population of 30 examples of one species of iris. Subsequently, the network was trained on 20 intervening items drawn this time from the same species.[2] As shown in Figure 2, learning these intervening items had almost no impact on the network's ability to produce the base population outputs.

While we note this effect for future discussion, in most cases of interest the extra items to be added to a network will not be further examples of the regularities captured by the base population. In general catastrophic forgetting will be, as the name implies, a significant practical problem.

3. The Rehearsal and Pseudorehearsal Solutions

In Robins (1995), we showed that catastrophic forgetting can be alleviated by the use of rehearsal, the relearning of a subset of previously learned items at the same time that each new item is introduced. This forces the new intervening items to be incorporated into the context of the old base population items, preserving the structure of the base population instead of just 'overwriting' it.

The rehearsal mechanisms that we explored were set in the context of learning a base population followed by intervening items just as described above. Rehearsal was implemented by introducing each new intervening item not alone, but in a 'rehearsal buffer' along with a fixed number of old items (base population or previously learned intervening items). The items in the rehearsal buffer (including the new intervening item) were then trained in the usual way until they were learned to criterion. The various possible ways of selecting and managing the old items in a rehearsal buffer define a large family of possible 'rehearsal regimes'.

3.1. Random Rehearsal and Sweep Rehearsal

In this section, we will illustrate the use of two rehearsal regimes, random rehearsal and sweep rehearsal, to alleviate catastrophic forgetting. Our baseline (no rehearsal) is the experiment described in the first simulation as shown in Figure 1. For the current simulations we again use a base population consisting of 30 items of one species of iris, and 20 intervening items drawn from a second species. Each intervening item, however, is trained in a rehearsal buffer which also contains five old items.[3]

Random rehearsal involves choosing the five old items for each rehearsal buffer at random from all previously learned items. As each new intervening item is introduced a rehearsal buffer containing that item and the five randomly chosen old items is created, and all items in the buffer are trained to criterion. Our third simulation implements this regime, the results of which are shown in Figure 3. Performance on the base population is maintained at a high level of accuracy. In other words, rehearsing random old items as each new item is learned is an effective way of reducing catastrophic forgetting.

Random rehearsal, like most other regimes explored in Robins (1995), uses a 'fixed buffer'. For a given new item, the old items (however they are selected) remain constant in the buffer and are trained to criterion with the new item. Sweep rehearsal is based on the use of a 'dynamic buffer'. Intervening items are trained in a buffer with previously learned items as usual, however, the previously learned items are chosen at random for each epoch. Training progresses over a number of epochs until only the single new item, which is always in the buffer, is trained to criterion.

Our fourth simulation repeats the conditions of the third simulation and implements the sweep rehearsal method. As each new intervening item is introduced a rehearsal buffer containing that item and the five randomly chosen old items is created every epoch, and training continues (repeating this process) until the new intervening item is trained to criterion. The results are shown in Figure 3. Performance on the base population is maintained even more effectively than in the case of random rehearsal.

Sweep rehearsal is the most effective rehearsal regime that we have explored, despite the fact that it does not explicitly retrain old items to criterion. Instead, for every new item introduced, sweep rehearsal's dynamic buffer exposes many old items to just one or more training presentations (epochs).

3.2. Random Pseudorehearsal and Sweep Pseudorehearsal

It is possible to achieve the benefits of rehearsal described above even when there is no access to the base population on which the network was trained. In other words, we can do rehearsal even when we do not have the old items to rehearse! This mechanism, which we called pseudorehearsal, is based on the use of artificially constructed populations of 'pseudoitems' instead of the 'actual' previously learned items in the rehearsal process.

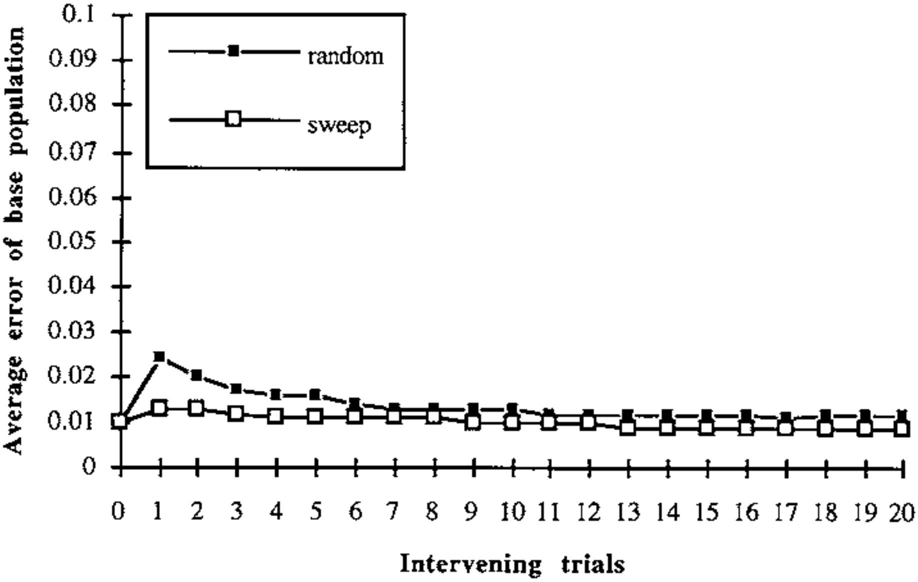

Figure 3. Results from the third and fourth simulations comparing random and sweep rehearsal (cf. the no-rehearsal condition shown in Figure 1). Note the different scale on the *y* axis compared with Figures 1 and 2.

A pseudoitem is constructed by generating a new input vector at random and passing it forward through the network in the standard way. Whatever output vector this input generates becomes the associated target output. A population of pseudoitems constructed in this way can be used instead of the actual items in any rehearsal regime. The pseudopopulations used in our simulations contained 128 real-valued random pseudoitem pairs.

Pseudorehearsal proceeds in the same way as the ordinary rehearsal regimes described earlier—a network is trained on a base population then new intervening items are introduced and trained one at a time using a rehearsal buffer. The given regime proceeds as usual, except that whenever it is necessary to choose a previously learned item or items for rehearsal a pseudopopulation is constructed (as described above) and a pseudoitem or items are chosen instead. Pseudoitems can be thought of as sampling or mapping the output behaviour or function learned by the network in the process of learning the old items, and as taking the place of the old learned items during further learning.

We now consider the random pseudorehearsal and sweep pseudorehearsal variants of the random rehearsal and sweep rehearsal regimes presented earlier. Once again, we use the first simulation (no rehearsal) as our baseline condition. In the fifth simulation we repeat the procedure of the third simulation (random rehearsal), except that pseudoitems are used instead of 'real' old items in the rehearsal buffer. Before each new intervening item is learned the network is used to construct a new pseudopopulation. The new item is then learned in a rehearsal buffer containing five randomly chosen pseudoitems, and all items in the buffer are trained to criterion. The results are shown in Figure 4. While not as effective as random rehearsal of real old items, random pseudorehearsal is still fairly successful at maintaining performance on the base population.

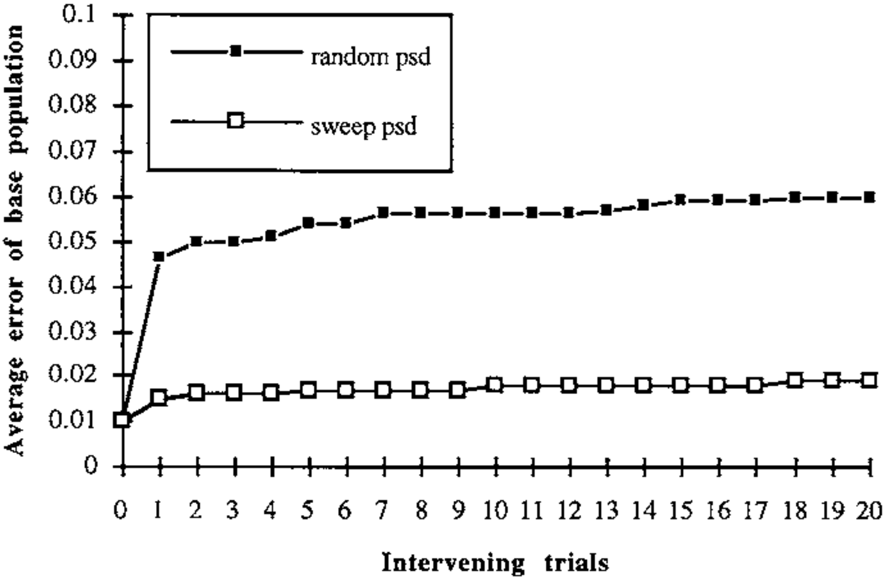

Figure 4. Results from the fifth and sixth simulations comparing random and sweep pseudorehearsal (cf. random and sweep rehearsal in Figure 3 and the no-rehearsal condition shown in Figure 1).

In the sixth simulation, we repeat the procedure of the fourth (using sweep rehearsal), except that pseudoitems are used. Before each new intervening item is learned a new pseudopopulation is constructed. As each new intervening item is introduced, a rehearsal buffer containing that item and the five randomly chosen pseudoitems is created every epoch, and training continues (repeating this process) until the new intervening item is trained to criterion. The results are shown in Figure 4. Sweep pseudorehearsal remains highly effective at preserving performance on the base population. After the 20th intervening item the error is roughly 2% of the error of the no-rehearsal condition and increasing only very gradually.

3.3. Discussion

The results of these rehearsal and pseudorehearsal regimes can be interpreted as follows. A trained network can be thought of as embodying a function that encodes the base population. Learning new items changes the shape of this function, disrupting performance on the base population. In rehearsal, training old items (including base population items) along with new items forces the network to preserve the shape of the base population function (and therefore maintain performance on the base population items). We can still achieve this effect, however, even without access to the old items. Pseudoitems randomly sample the structure of the base population function originally learned by the network. Training pseudoitems along with new items still forces the network to preserve the shape of the base population function.

The results presented here using the Iris data set replicate, in every important respect, the results described by Robins (1995) using a randomly constructed data set. Performance of the random regimes, however, is somewhat improved in the current simulations, as the relearning of old items exploits the structure of the population. Performance of the sweep regimes remains nearly optimal. In general terms the 'broad and shallow' approach of the sweep regimes is a much more successful rehearsal strategy than the 'narrow and deep' approach of the random regimes (or other fixed buffer regimes). This suggests that, in general, rehearsal should be broad (include lots of items), but it does not need to be deep (rehearse items to criterion).

As well as preserving the base population, the rehearsal regimes presented here also maintain the good generalization performance characteristic of both ANN learning methods and human cognition. For each of the four rehearsal regimes, during the training and testing described above the average error of a generalization population was tested every time the average error of the base population was tested (i.e. for each test from 0 to 20 intervening items). The generalization population consisted of a further 20 examples of iris drawn from the same species as the 30 base population examples. For the random and sweep rehearsal regimes the average error of the generalization population typically exceeded the average error of the base population by no more than 0.003 to 0.005. For the random pseudorehearsal the average error of the test population typically exceeded the average error of the base population by no more than 0.002, and for sweep pseudorehearsal the errors were indistinguishable.

To summarize, catastrophic forgetting is a very general consequence of an ANN style of learning affecting a wide range of networks. A rehearsal or pseudorehearsal based solution to the catastrophic forgetting problem offers a mechanism for adding information to ANNs without disrupting existing information. In short, it provides a possible basis for modelling the consolidation of information over long-term learning. In the following sections we explore the relationship between pseudorehearsal and 'sleep consolidation', a mechanism which it has been suggested is the basis of the long-term consolidation of information in human learning.

4. The Sleep Consolidation Hypothesis

One theory regarding the function of sleep, particularly rapid eye movement (REM) sleep, is that it is a period when newly acquired information is integrated into existing long-term memory (see for example Greenberg and Pearlman (1974), Winson (1990), Wilson and McNaughton (1994)). We refer to this as the 'sleep consolidation hypothesis' (or just 'sleep consolidation').

Winson (1972) suggested that essential information gathered during the day was 'reprocessed into memory' during REM sleep. In a more recent overview, Winson (1990) reviews evidence from a number of studies and observations which support this proposal. The review focuses initially on the theta rhythm—a regular cycle observed in the hippocampus of mammals (including primates and humans, e.g., Sano *et al.* (1970), Stewart and Fox (1991)). Winson reviews studies suggesting that disruptions to the theta rhythm destroy spatial memory, and that the theta rhythm facilitates the long-term potentiation (LTP) mechanisms assumed to underlie learning. Theta rhythms occur during the apprehension of significant or changing environmental information, and also during REM sleep, suggesting a link between waking and sleeping learning mechanisms. Winson also considers evidence at the level of individual neurons, reviewing a study of place field cells—hippocampal cells that fire maximally when the animal is at specific locations in a given environment. Specific place field cells which happen to become active as an animal explores a new environment become active again during sleep at significantly greater than background levels, suggesting that the animal is reprocessing or strengthening the information encoded while it was awake. Winson then briefly presents evolutionary arguments linking the evolution of the theta rhythm to neuroanatomical changes in the mammalian brain underlying "advanced perceptual and cognitive features", and developmental arguments linking the patterns of REM sleep in infants with their information processing requirements. In summarizing the various animal studies, Winson states "In REM sleep this [important] information may be accessed again and integrated with past experience to provide an ongoing strategy for behaviour" (Winson, 1990, p. 47). Winson speculates that in humans "Dreams may reflect a memory processing mechanism inherited from lower species, in which information important for survival is reprocessed during REM sleep" (Winson, 1990, p. 47).

Smith (1993) summarizes results from a number of studies relating to REM sleep and learning. These studies include both animal and human studies of the relationship between training/learning experiences and subsequent REM sleep, and the effect of sleep deprivation on the efficacy of learning. Smith concludes with the following summary points.

> (1) REM sleep increases or increases in number of REMs appear after substantial learning has taken place.
> (2) These REM sleep/REMs increases appear to persist for many hours or days after the end of acquisition.
> (3) At certain times, which seem to coincide with the beginning of expected REM sleep or REMs increase, application of REMD [REM sleep deprivation] results in relatively permament learning–memory deficits. These vulnerable time periods have been named *REM windows*.
> (4) Not all learning is sensitive to REMD. Material that is relatively simple and with which the subject is generally familiar does not appear to be affected by REMD. More likely to be involved with REM sleep is material that requires the learning and understanding of new concepts that were previously unfamiliar (Smith, 1993, p. 356).

These findings are consistent with the claims of the sleep consolidation hypothesis. The first three points suggest a gradual process with a fairly direct correspondence between the amount of material to be learned and the amount of REM sleep necessary for effective learning. The fourth point introduces a more subtle observation which we shall return to later.

Wilson and McNaughton (1994) further develop evidence linking the hippocampus to sleep consolidation, although the mechanisms that they describe do not occur during the REM phase, and suggest that consolidation may involve the interactions of processes in different phases of sleep. Recording from multiple hippocampal place cells (in the rat) during spatial behavioural tasks, they found that cells which fired together when the rat occupied a particular location in the environment were more likely to fire together during subsequent sleep. This effect declined gradually during each subsequent sleep session. From this and other observations Wilson and McNaughton conclude:

> . . . initial storage of event memory occurs through rapid synaptic modification, primarily within the hippocampus. During subsequent slow-wave sleep synaptic modification within the hippocampus itself is suppressed and the neuronal states

> encoded within the hippocampus are 'played back' as part of a consolidation process by which hippocampal information is gradually transferred to the neocortex (Wilson & McNaughton, 1994, p. 678).

To summarize, sleep is a complex and multi-faceted phenomena, and there are a wide range of theories on the nature and function of sleep. For the purposes of this paper we are interested in a particular hypothesis, that new information is integrated with existing memories during sleep, which we refer to as the 'sleep consolidation hypothesis'.

5. Pseudorehearsal and the Sleep Consolidation Hypothesis

We are now in a position to state one of the main claims of this paper. We suggest that the results summarized in the foregoing, particularly those relating to pseudorehearsal, provide converging evidence for the veracity of the sleep consolidation hypothesis. We suggest that the catastrophic forgetting problem in ANNs is a problem that has actually occurred in the evolution of the mammalian brain, and that the pseudorehearsal solution to the problem in ANNs is functionally equivalent to the sleep consolidation solution adopted by the brain.

5.1. Catastrophic Forgetting and the Need for Sleep Consolidation

The stability/plasticity dilemma is a fundamental one for any learning system, artificial or natural, faced with the task of integrating new information with old. In ANNs the catastrophic forgetting problem is a basic consequence of the plasticity of the representing medium, connection weights. As neural representations are also thought to be mediated by plastic connection weights (synapses) it seems probable that catastrophic forgetting has also been encountered during the evolution of the brain (see, for example, Winson's (1990) discussion of the evolution of the mammalian cortex).

Catastrophic forgetting implies that a special learning mechanism is required to explicitly manage the integration of new information.[4] We suggest that in mammals sleep consolidation is exactly this mechanism. Currently, the sleep consolidation hypothesis describes a process, but it does not explain the function of the process. (Why not just integrate new information during waking cognition as it is learned?) Our proposal extends the sleep consolidation hypothesis with the suggestion that a major function of/motivation for sleep consolidation is the management of catastrophic forgetting. We further suggest that sleep consolidation is functionally equivalent to the pseudorehearsal solution in ANNs, and that a consideration of pseudorehearsal mechanisms may therefore supply at least initial hypotheses about the details of sleep consolidation. In the remainder of this section, we draw out the parallels between these two mechanisms.

5.2. Parallels Between Sleep Consolidation and Pseudorehearsal

5.2.1. Preserving the structure of old information. While operating in different domains, both pseudorehearsal and sleep consolidation have the same underlying function—the integration of new information into old information while preserving the structure of the old information.

5.2.2. Rehearsing approximations of old information. While it is an effective method in ANNs, rehearsal (cf. pseudorehearsal) is unlikely to be a realistic model of biological learning mechanisms, as in this context the actual old information (accurate and complete representations of all items ever learned by the organism) is not available. Pseudorehearsal is significantly more likely to be a mechanism which could actually be employed by organisms, as it does not require access to this old information, it just requires a way of approximating it.

Pseudorehearsal approximates old information by using randomly constructed inputs to sample the behaviour of (the function encoded by) the weights that embody it. While it is not currently possible to confirm in detail that sleep consolidation is using this mechanism, there is evidence which is at least suggestive of such a process. Hobson and McCarley (cited in Winson (1990)) proposed that dreaming consists of associations and memories elicited from the forebrain as a result of random or 'chaotic' inputs from the brain stem such as PGO (pontine-geniculate-occipital cortex) spikes. The theme of random neocortical activation was central to the Crick and Mitchison (1983) account of the function of dream sleep, and Moffit and Cramer (1993, pp. 3–4) also cite the apparent randomness of dreaming at the physiological level.

Pseudorehearsal uses the approximations of old information as a context within which new information must be integrated during learning. If the neocortex is indeed being randomly stimulated as a method of approximating old information as proposed above, then data supporting the sleep consolidation hypothesis and the role of the hippocampus in learning (see Section 6) both support the proposal that new information is being integrated into this context.

Apart from the broad similarities described above, we can extrapolate from the superiority of sweep rehearsal regimes over fixed buffer rehearsal regimes (see Section 3) to two predictions about sleep consolidation. First, that sleep consolidation will be most effective if it utilizes a broad sampling of old information (many old items), but second that this sampling may also be shallow (not rehearse any specific item for an extended period). This implies that the process which samples old information during sleep consolidation can create quite transitory approximations.

5.2.3. Not necessary for 'compatible' information. Recall from the second simulation that catastrophic forgetting does not occur in ANNs when new information is added which is consistent with the information already learned by the network. In this case the weight changes created in learning the new information do not disrupt the old information. Catastrophic forgetting occurs when new input/output mappings are significantly different from those already learned by the network.

Similarly, sleep consolidation does not appear to be necessary for learning new information which is already consistent with an organism's knowledge. Recall (see Section 4) that in summarizing a range of studies of REM sleep and learning one of Smith's conclusions is that:

> Not all learning is sensitive to REMD [REM sleep deprivation]. Material that is relatively simple and with which the subject is generally familiar does not appear to be affected by REMD. More likely to be involved with REM sleep is material that requires the learning and understanding of new concepts that were previously unfamiliar (Smith, 1993, p. 356).

It is interesting to note that both pseudorehearsal and sleep consolidation share this specific property.

5.2.4. An 'off-line' process. Pseudorehearsal represents a specific phase of ANN learning unlike standard learning or processing states. Sleep is also a distinct and characteristic state. It could be described as an 'off-line' state which incurs several costs for an organism in terms of reduced perception, mobility and time to forage. Such a state would be necessary, however, if the sleep consolidation process were disruptive to and incompatible with ongoing cognition. Winson (1990, p. 47) notes, for example, that the stimulation arising during sleep would be hard to 'dissociate' from waking activity, and that locomotion is suppressed during REM sleep presumably to prevent the stimulation of motor neurons from actually causing movement.

If sleep consolidation is indeed similar to pseudorehearsal as we have proposed, then it involves significant, random and unfocused 'activation' of long-term memory. It could certainly be argued that this would cause severe disruption to ongoing cognition and necessitate an off-line sleep consolidation process.

6. Towards a Model of Consolidation

McClelland *et al.* (1995) also consider the implications of ANN style learning and the catastrophic forgetting problem,[5] and also suggest that special learning mechanisms are required in principle. Considering a broad range of empirical and theoretical results, they draw detailed conclusions about the properties of such learning mechanisms in organisms, emphasizing the following points in particular:

- In order to extract underlying structure successfully, information must be learned slowly.
- In order to preserve the structure of old information, new information must be learned 'interleaved' (mixed together) with old information.
- 'Complementary learning systems' are required, a fast hippocampal system for learning new information, and a slow neocortical system for storing old information and its structure long term.

McClelland *et al.* suggest that the neocortex learns slowly in order to discover the underlying structure of incoming information. The hippocampal system allows for the rapid learning of new information without disrupting this structure. This new information is then 'reinstated' by the hippocampus in the neocortex and integrated with the old information/long-term memory via interleaved learning.

The reinstatement and learning process is assumed to occur "in off-line situations including active rehearsal, reminiscence, and other inactive states including sleep" (McClelland *et al.*, 1995, p. 424). Although McClelland *et al.* do not further pursue this suggestion, their account is obviously consistent with the sleep consolidation hypothesis and its supporting empirical evidence linking the hippocampus with the integration of new and old information during sleep, particularly the proposals of Wilson and McNaughton (1994) noted earlier. In short, from the shared starting point of an investigation of the consequences and management of catastrophic forgetting, the account proposed by Robins (1995) and developed in this paper, and that of McClelland *et al.* (1995), both explore the nature of a learning mechanism which is assumed to occur (at least partly and possibly predominantly) during sleep.[6] Although the two accounts emphasize different aspects of this mechanism, we suggest that they are complementary aspects of a coherent picture.

Both accounts propose that mixing new and old items during training is the basis of an appropriate solution to the catastrophic forgetting problem. While the exploration of the possible nature of such a mechanism is the central thrust of our work on rehearsal and pseudorehearsal, the proposal is not developed by McClelland *et al.* beyond its statement in these very general terms and the coining of the description 'interleaved learning'. Their implementation of interleaved learning is likewise intended as a very general illustration. A standard backpropagation network is used to train a base population for 500 epochs, new items are added to the population, and training of all items is continued for a further 100 epochs. While this method is effective at learning new items without catastrophic forgetting, it provides no theoretical or practical advances over existing ANN learning methods, and was not proposed or intended as a biologically plausible mechanism.

Effective and biologically plausible interleaved learning is, however, in principle central to the McClelland *et al.* account:

> The observation that interleaved learning allows new knowledge to be gradually incorporated into a structured system lies at the heart of our proposals concerning the role of the hippocampus in learning and memory (McClelland *et al.*, 1995, p. 434).

We suggest that the rehearsal and pseudorehearsal methods that we have proposed, and in particular the general conclusions that approximations of old information can be used in place of the old information itself and that rehearsal needs to be broad but may be shallow, go some way towards outlining the possible nature of a biologically plausible mechanism.

In contrast, the need for separate learning systems, certainly implied in a biological context by the drawing of a distinction between new and old information and the use of a rehearsal type mechanism, is not developed at all in our previous work. This is the topic which McClelland *et al.* explore in depth. They combine a wide range of neurological and behavioural evidence and ANN models, to develop both a consistent and well-motivated account of the hippocampal and neocortical learning systems and a quantitative model of the learning process.

In short, the account proposed by McClelland *et al.* describes the systems necessary for integrating new and old information (complementary fast and slow learning mechanisms), describes the process in general terms (interleaving new and old information in learning) and anchors the necessary system in a well-motivated biological framework (the hippocampus and the neocortex). We have explored the possible nature of the learning process in more detailed terms (proposing a broad-based rehearsal of approximations of old information) and emphasized the link to the biological mechanisms of learning during sleep (including the random stimulation of the cortex).

These accounts complement each other in the exploration of the possible nature of the biological learning mechanisms which manage catastrophic forgetting in the successful long-term consolidation of information. Arising from this synthesis we propose the following tentative model of learning and consolidation in humans. Information from the environment is checked against predictions based on long-term memory. If there is a good match then we have already learned this information, so no learning is required. If there is an approximate match then this is new information which is consistent with what we already know and is safe to integrate directly with our long-term memory. Neocortical learning of this item can begin during waking. If there is a poor match then this is new information which is very different from what we already know. As it would disrupt the associations in long-term memory, such information is learned during waking by the fast hippocampal system. During sleep the neocortical long-term memory base gets randomly, broadly and transiently stimulated, creating a 'pseudoitem effect', an approximation of old information which can be used in learning. At the same time the hippocampus reinstates the information that it is holding into the cortex for learning. Learning occurs in the cortex involving both the rehearsed approximations of the old information and the hippocampally mediated new information. This integrates the new information into the context of the old information without disrupting it.

This tentative model ignores a wealth of information and competing hypotheses about the nature of learning and the function of the hippocampus. It does, however, combine robust consequences of ANN learning with significant supporting psychological and neurological evidence. To the extent that it is true, it is no doubt only a part of a complex picture, but it may be a useful characterization of at least a significant part of that picture, and it may be a useful framework for further work.

7. Summary

To summarize, a broad range of ANN models suffer from the catastrophic forgetting problem which prevents the long-term consolidation of information in the network. We have outlined solutions to the catastrophic problem based on rehearsal mechanisms (using old information) and pseudorehearsal mechanisms (using approximations of old information), which may offer a framework for modelling long-term consolidation in ANNs. We have also related the pseudorehearsal mechanism to the sleep consolidation hypothesis (one of many hypotheses about the function of sleep), which proposes that newly acquired information is integrated into existing long-term memory during sleep. We suggest that the catastrophic forgetting problem in ANNs is a problem that has actually occurred in the evolution of the mammalian brain, and that the pseudorehearsal solution to the problem in ANNs is functionally equivalent to the sleep consolidation solution adopted by the brain. This provides an evolutionary motivation for sleep consolidation, converging evidence for the veracity of sleep consolidation hypothesis, and predictions about the details of the process. McClelland *et al.* (1995) also explore the consequences of the catastrophic forgetting problem for learning in organisms. From this same starting point, they reach conclusions which are different in emphasis from our own, but complementary aspects of a coherent underlying picture of the nature of a biologically plausible learning mechanism.

We suggest, then, that pseudorehearsal mechanisms may provide a useful framework for modelling long-term learning and the consolidation of information in ANNs. Furthermore, we claim that this approach can be strongly related to a biologically plausible account of the consolidation of information during sleep.

Acknowledgements

This paper was written while I was a visiting research associate at Carnegie Mellon University. My thanks to the Psychology Department and my host Professor James

McClelland for supporting this visit. Thanks also to James McClelland and members of the PDP Research Group for very useful comments on drafts of this paper.

Notes

1. Several previous explorations of mechanisms for reducing catastrophic forgetting have focused on reducing representational overlap, particularly in the hidden unit representations developed by the network. The novelty rule (Kortge, 1990), activation sharpening (French, 1992), techniques developed by Murre (1992) and McRae and Hetherington (1993), and context biasing (French, 1994), all fall within this general framework. The most direct of these methods does not prevent a base population from being disrupted by intervening items, but they do protect the base population to the extent that they allow it to be subsequently retrained to criterion more quickly than is the case in a standard backpropagation network.
2. Each interventing item was trained (i.e. weights were adjusted) at least once. Many of the intervening items required no further training to satisfy the criterion, but many did require several further training presentations.
3. The ratio of old items in the rehearsal buffer to the size of the base population is an important factor. The performance of all regimes can be arbitrarily improved by increasing the size of the rehearsal buffer, but the use of too much of the base population in rehearsal reduces the utility and plausibility of the rehearsal process. (Trivially the rehearsal buffer can be made so large that the whole base population is relearned for every new item.) For the sake of consistency, in this paper we continue to use the baseline established in Robins (1995), setting the size of the rehearsal buffer to include a number of old items equal to roughly 15% of the size of the base population. This figure appears to provide an acceptable trade-off beween performance and the amount of rehearsal required.
4. Note that the same chain of reasoning leads McClelland *et al.* (1995) to a different (complementary) conclusion, as discussed in Section 6.
5. Which they called by an alternative name, catastrophic interference.
6. From our perspective we would argue that such learning occurs predominantly during sleep (cf. the other 'off-line' mechanisms proposed by McClelland *et al.*), particularly if, as seems likely, it is disruptive to ongoing cognition.

References

Burgess, N., Shapiro, J.L. & Moore, M.A. (1991) Neural network models of list learning. *Network*, **2,** 399–422.

Carpenter, G.A. & Grossberg, S. (1988) The ART of adaptive pattern recognition by a self-organising neural network. *Computer*, **21,** 77–88.

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

Empson, J. (1993) *Sleep and Dreaming*, 2nd edition. Hemel Hempstead, UK: Harvester Wheatsheaf.

French, R.M. (1992) Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, **4,** 365–377.

French, R.M. (1994) Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. *Proceedings of the 16th Annual Cognitive Society Conference.* Hillsdale, NJ: Lawrence Erlbaum, pp. 335–340.

Greenberg, R. & Pearlman, C.A. (1974) Cutting the REM nerve, An approach to the adaptive role of REM sleep. *Perspectives in Biology and Medicine*, **17,** 513–521.

Grossberg, S. (1987) Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, **11,** 23–63.

Globus, G.G. (1993) Connectionism and Sleep. In A. Moffitt & M. Kramer (Eds), *The Functions of Dreaming.* Albany, NY: State University of New York Press.

Hetherington, P.A. & Seidenberg, M.S. (1989) Is there 'catastrophic interference' in connectionist networks? *Proceedings of the 11th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 26–33.

Hinton, G.E. & Sejnowski, T.J. (1986) Learning and relearning in Boltzmann machines. In D.E. Rumelhart, J.L. McClelland & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* Cambridge, MA: MIT Press.

Hobson, A.J., Mamelak, A.N. & Sutton, J.P. (1992) Models wanted: Must fit dimensions of sleep and dreaming. In J.E. Moody & R.P. Lippman (Eds), *Advances in Neural Information Processing*, Vol. 4. San Mateo, CA: Morgan Kaufmann.

Hopfield, J.J., Feinstein, D.I. & Palmer, R.G. (1993) Unlearning has a stabilising effect in collective memories. *Nature*, **304,** 158–159.

Kortge, C.A. (1990) Episodic memory in connectionist networks. *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum, pp. 764–771.

Lewandowsky, S. (1991) Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In W.E. Hockley & S. Lewandowsky (Eds), *Relating Theory and Data: Essays on Human Memory in Honour of Bennet B. Murdok.* Hillsdale, NJ: Lawrence Erlbaum.

Lewandowsky, S. & Li, S. (1995) Catastrophic interference in neural networks: Causes, solutions, and data. In F.N. Dempster & C. Brainerd (Eds), *New Perspectives on Interference and Inhibition in Cognition.* New York: Academic Press.

McClelland, J.L., McNaughton, B.L. & O'Rielly, R.C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the success and failures of connectionist models of learning and memory. *Psychological Review*, **102,** 419–457.

McCloskey, M. & Cohen, N.J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 23. New York: Academic Press, pp. 109–164.

McRae, K. & Hetherington, P.A. (1993) Catastrophic interference is eliminated in pretrained networks. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum, pp. 723–728.

Moffitt, A. & Kramer, M. (Eds) (1993) *The Functions of Dreaming.* Albany, NY: State University of New York Press.

Murphy, P.M. & Aha, D.W. (1994) UCI Repository of Machine Learning Databases [http://www.ics.uci.edu/~ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Murre, J.M.J. (1992) The effects of pattern presentation on interference in backpropagation networks. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum, pp. 54–59.

Nadal, J.P., Toulouse, G., Changeux, J.P. & Dehaene, S. (1986) Networks of formal neurons and memory palimpsets. *Europhysics Letters*, **1,** 535–542.

Ratcliff, R. (1990) Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, **97,** 285–308.

Robins, A.V. (1995) Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science*, **7,** 123–146.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* Cambridge, MA: MIT Press.

Sano, K., Mayangi, Y., Sekino, H., Ogashiwa, M. & Ishijima, B. (1970) Results of stimulation and destruction of the posterior hypothalamus in man. *Journal of Neurosurgery*, **33,** 689–707.

Sharkey, N.E. & Sharkey, A.J.C. (1994a) Understanding catastrophic interference in neural nets. *Technical Report CS-94-4*, Department of Computer Science, University of Sheffield, UK.

Sharkey, N.E. & Sharkey, A.J.C. (1994b) Interference and discrimination in neural net memory. In J. Levy, D. Bairaktaris, J. Bullinaria & P. Cairns (Eds), *Connectionist Models of Memory and Language.* London: UCL Press.

Smith, C. (1993) REM sleep and learning: Some recent findings, In A. Moffit & M. Kramer (Eds), *The Functions of Dreaming.* Albany, NY: State University of New York Press.

Stewart, M. & Fox, S.E. (1991) Hippocampal theta activity in monkeys. *Brain Research*, **538,** 59–63.

Sutton, J.P., Mamelak, A.N. & Hobson, J.A. (1992) Modeling states of waking and sleeping. *Psychiatric Annals*, **23,** 137–143.

Wilson, M.A. & McNaughton, B.L. (1994) Reactivation of hippocampal ensemble memories during sleep. *Science*, **265,** 676–679.

Winson, J. (1972) Interspecies differences in the occurrence of theta. *Behavioural Biology*, **7,** 479–487.

Winson, J. (1990) The meaning of dreams. *Scientific American*, November 1990, 42–48.