The consolidation of learning during sleep: comparing the pseudorehearsal and unlearning accounts

A. Robins*, S. McCallum

Computer Science Department, University of Otago, PO Box 56, Dunedin, New Zealand

Received 22 December 1998; accepted 2 April 1999

Abstract

We suggest that any brain-like (artificial neural network based) learning system will need a sleep-like mechanism for consolidating newly learned information if it wishes to cope with the sequential/ongoing learning of significantly new information. We summarise and explore two possible candidates for a computational account of this consolidation process in Hopfield type networks. The “pseudorehearsal” method is based on the relearning of randomly selected attractors in the network as the new information is added from some second system. This process is supposed to reinforce old information within the network and protect it from the disruption caused by learning new inputs. The “unlearning” method is based on the unlearning of randomly selected attractors in the network after new information has already been learned. This process is supposed to locate and remove the unwanted associations between information that obscure the learned inputs. We suggest that as a computational model of sleep consolidation, the pseudorehearsal approach is better supported by the psychological, evolutionary, and neurophysiological data (in particular accounting for the role of the hippocampus in consolidation). © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Consolidation; Sleep; Catastrophic forgetting; Pseudorehearsal; Unlearning

1. Introduction

Most artificial neural network (ANN) learning algorithms are based on “concurrent” learning, in other words the whole population of training items is presented and trained as a single, complete entity. Training is then finished and no further information is learned by the network. In contrast, human learning is an ongoing, lifelong process. It is “sequential” in that we can learn new information at any time and integrate it into what we already know.

Most ANN learning algorithms are not capable of sequential learning because of a fundamental underlying problem: when new information is learned by a network it significantly disrupts or even eliminates information that has been previously learned. This problem has been identified in many forms in literature, it is commonly called the “catastrophic forgetting”, “catastrophic interference”, or “serial learning” problem.

In this paper, we suggest that the brain is capable of sequential learning because it has solved the catastrophic forgetting problem by consolidating newly learned information during sleep, a process which has been widely explored and supported in literature. We suggest that in order to likewise be capable of sequential learning, ANN based (“brain-like”) learning systems will also need a specific process (analogous to sleep consolidation) which serves to integrate new and old information in the network. We are interested in exploring possible computational models of this process using Hopfield type ANNs. Hopfield networks are dynamical systems where learning creates attractors in a multi-dimensional state space. The dynamical properties of these networks, their ability to operate as associative and content-addressable memories, and their ability to represent and “reason” about multiple “soft constraints” make them an attractive framework for modelling cognition (Rumelhart, 1989; Crick & Mitchison, 1995).

The main focus of this paper is to describe two possible solutions to the catastrophic forgetting problem in Hopfield type networks and compare them as candidates for a computational model of the process of sleep consolidation. The two methods explored are the “pseudorehearsal” method (Robins, 1995; Robins & McCallum, 1998) and the “unlearning” method (Hopfield, Feinstein & Palmer, 1983; Crick & Mitchison, 1983). Pseudorehearsal was originally developed as a solution to the catastrophic forgetting problem in MLP (multi-layer perceptron) type networks by Robins (1995), and was extended to Hopfield type networks.
networks by Robins and McCallum (1998). Although not directly connected to the catastrophic forgetting literature, the unlearning method addresses the same issues of capacity and serial/sequential learning tasks in Hopfield networks.

We begin by describing the catastrophic forgetting problem in Section 2. We briefly review possible solutions in MLP type networks before moving on to describe and evaluate the pseudorehearsal and unlearning methods in Hopfield type networks. In Section 3, we describe the consolidation of learning which is thought to occur during sleep. We show how we believe that the catastrophic forgetting problem in ANNs motivates such a consolidation process in the brain, and compare the pseudorehearsal and unlearning methods as candidates for a computational account of the process. We suggest that the pseudorehearsal account is better supported by the psychological and neurological data, and in Section 4 we briefly review evolutionary data which also seems to be consistent with this claim.

2. Catastrophic forgetting, causes and solutions

Here, we describe the catastrophic forgetting problem. Catastrophic forgetting has been studied mostly in the context of MLP type architectures. We briefly review possible solutions in MLP networks based on reducing representational overlap, rehearsal, and pseudorehearsal. We then summarise possible solutions in Hopfield type networks, reviewing the use of rehearsal and pseudorehearsal in this type of network, and introducing the unlearning method. Finally, we compare the pseudorehearsal and unlearning methods in Hopfield type networks.

2.1. Catastrophic forgetting

The task of sequential learning highlights the “stability/plasticity dilemma” (Grossberg, 1987). Ideally the representations developed by a learning system should be stable enough to preserve important information over time, but plastic enough to incorporate new information when necessary. The use of variable connection weights as a medium for encoding information leads most ANNs to err on the side of excessive plasticity—new learning changes the weights, thus causing previously learned inputs to generate the wrong outputs. Grossberg suggests the analogy of a human trained to recognise the word “cat”, and subsequently to recognise the word “table”, being then unable to recognise “cat”. In short, unmanaged plasticity is the root of the catastrophic forgetting problem.

While stability/plasticity issues are very general, the term
“catastrophic forgetting” has tended to be associated with “static” networks (rather than “dynamic” nets which add or remove nodes) employing supervised learning. A number of recent studies have used MLP type networks (typically employing back-propagation) to highlight the problem of catastrophic forgetting and explore various issues—these include: McCloskey and Cohen (1989), Hetherington and Seidenberg (1989), Ratcliff (1990), Lewandowsky (1991), Murre (1992a), French (1992, 1994, 1997), McRae and Hetherington (1993), Lewandowsky and Li (1995), Sharkey and Sharkey (1995), Robins (1995, 1996), Frean and Robins (1998). Typical illustrations of the problem use a network to learn a “base” population of items (input/output vector pairs). Subsequently, a new item or population of items is learned. The effect of this new learning on the old items (base population) can be illustrated by re-testing the base population. The error of the originally learned base population of items typically increases “catastrophically” after learning even a single new item.

Although, usually explored using MLP networks, catastrophic forgetting can be just as easily demonstrated using Hopfield type networks, as is the case in Robins and McCallum (1998). Fig. 1 summarises data which illustrates catastrophic forgetting as described in that paper. The network is trained on a base population of 44 items so that every item is stable. If the network is then trained in a sequence:† given access to the previously learned base population items (input/output pairs) consisting of the new item and some items selected from the base population

2.2. Reducing overlap in MLP networks

French (1992) suggests that the extent to which catastrophic forgetting occurs is largely a consequence of the overlap of distributed representations, and that the effect can be reduced by reducing this overlap. Several studies have explored mechanisms for reducing representational overlap and their impact on catastrophic forgetting. The novelty rule (Kortge, 1990), activation sharpening (French, 1992), context biasing (French, 1994), and techniques developed by Murre (1992b), and McRae and Hetherington (1993) all fall within this general framework. These methods focus on increasing the separation (orthogonality) of hidden unit representations developed by the network, typically by creating “sparser” representations (hidden unit patterns with a smaller number of active units).

These methods do not, in general, prevent catastrophic forgetting from occurring, although they do ameliorate its effects to the extent that they allow the base population to be subsequently retrained to the criterion more quickly than is the case in an unmodified network. The novelty rule (Kortge, 1990) has been shown to prevent catastrophic forgetting, but can only be used with autoencoder (autoassociative) networks. Catastrophic forgetting may also be prevented if it is possible to pre-train the network on a population which is “relevant” to the base population and new items (simulating prior knowledge of a task domain) as explored by Sharkey and Sharkey (1995), McRae and Hetherington (1993). In McRae and Hetherington’s simulations this pre-training naturally reduced the overlap of hidden unit representations in subsequent learning.

2.3. Rehearsal and pseudorehearsal in MLP networks

A second general approach to preventing catastrophic forgetting involves “rehearsing” the base population by retraining some base population items as the new items are trained. Assuming that we learn new items one by one in a sequence:

• given access to the previously learned base population items (input/output pairs)
• for each new item to be learned

• construct a training population for the network consisting of the new item and some items selected from the base population
• learn the training population in the usual way (for the given network).

Ideally, this allows the new items to be incorporated into the structure of the base population instead of just overwriting it. If all old (base population) items are rehearsed, this
simply amounts to retraining the entire base population as new items are introduced (as is the case in, for example, the “interleaved learning” proposed by McClelland, McNaughton and O’Reilly (1995)). In general, however, it is possible to make do with rehearsing significantly less than the whole of the base population. Rehearsal was first explored in the context of catastrophic forgetting by Hetherington and Seidenberg (1989) and Ratcliff (1990), and a range of rehearsal methods have been explored by Murre (1992a); Robins (1995).

Rehearsal is an effective solution as long as the base population items are actually available for relearning. It may be, however, that the old items have been lost, or it is not practical for some reason to store them. In any case, retaining old items for rehearsal in memory seems somewhat artificial, as it requires that they be available on demand from some other sources, which would seem to make the memory redundant! It is possible to achieve the benefits of rehearsal, however, even when there is no access to the base population. In other words, we can do rehearsal even when we do not have the old items to rehearse. This “pseudorehearsal” mechanism, introduced by Robins (1995), is based on the use of artificially constructed populations of “pseudoitems” instead of the actual old items.

In MLP type networks, each pseudoitem (input/output pair) is constructed by generating a new input vector at random, and passing it forward through a network in the standard way. Whatever output vector this input generates, it becomes the associated target output. A population of pseudoitems constructed in this way can be used instead of the actual base population items in any rehearsal regime:

- for each new item to be learned
  - construct a population of pseudoitems by creating random inputs and passing them forward through the network to generate their associated outputs
  - construct a training population for the network consisting of the new item and some items selected from the population of pseudoitems
  - learn the training population in the usual way (for the given network).

In other words, rather than explicitly storing all learned items for later rehearsal, pseudorehearsal approximates this information whenever it is needed.

Using MLP type networks, this method has been shown to be effective over a range of different populations: autoassociative and heteroassociative randomly constructed data sets by Robins (1995), Ans and Rousset (1997); autoassociative learning of the Iris data set by Robins (1996); a classification task using the Mushroom data set by French (1997); and a structured “task domain” by Silver and Mercer (1998)\(^1\). In short, pseudorehearsal is an effective solution to the catastrophic forgetting problem that enables sequential learning in an ANN.

It is important to note that pseudorehearsal does not preserve the weights of a network, these can change dramatically during learning. The method works directly at the level of the function embodied by the trained network. Rehearsing pseudoitems (points randomly chosen from the base population function) ensures that the general shape of the old population function is preserved, and changes necessary to accommodate the new item are restricted to being local to the region of the new item. This localisation of changes to the function is the essence of the pseudorehearsal method in MLP type networks (Robins, 1995; Frean & Robins, 1998). As a starting point to a formal analysis of pseudorehearsal, Frean and Robins (1998) present an analysis of the method in linear networks.

2.4. Rehearsal and pseudorehearsal in Hopfield type nets

The principles of rehearsal and pseudorehearsal can also be applied to Hopfield type networks, although the actual pseudorehearsal mechanism turns out to be somewhat different in practice, and the distinction between the two starts to blur somewhat. In this section, we briefly summarise the use of rehearsal and pseudorehearsal in the Hopfield type network introduced in Section 2.1. These topics were first explored in Robins and McCallum (1998) and some of the results presented there are summarised in Fig. 1.

In general terms, the rehearsal process is the same as for MLP networks:

- given access to the previously learned base population items (inputs to be autoassociated)
- for each new item to be learned
  - construct a training population for the network consisting of the new item and some items selected from the base population
  - learn the training population in the usual way (for the given network).

If all old base population items are relearned along with each new item, then performance on the base population remains perfect (the “Rehearse 100%” condition in Fig. 1). The degree of retention of the base population is directly proportional to the percentage of the population that is rehearsed. Relearning a randomly selected 50% of base population items along with each new item still provides some protection for the base population during the learning of the new items (see the “Rehearse 50%” condition).

Assuming that we do not have access to a separate store of the old items to rehearse, we can successfully adapt pseudorehearsal to Hopfield type networks, and this provides significant protection from catastrophic forgetting. For Hopfield type networks, each pseudoitem (input to be autoassociated) is generated by creating a random input and iterating it through the network until a stable state/attractor is reached. This final stable state is the pseudoitem. A

\(^1\)The Iris and Mushroom data sets are described in Murphy and Aha (1994).
population of pseudoitems created in this way (but excluding duplicate items) can be used instead of the actual base population items in any rehearsal regime:

- construct a population of pseudoitems by creating random inputs and iterating each one through the network until a stable state is reached
- for each new item to be learned
  - construct a training population for the network consisting of the new item and some items selected from the population of pseudoitems
  - learn the training population in the usual way (for the given network).

Fig. 1 (“Pseudorehearse” condition) shows the effect of learning a population of 256 pseudoitems along with each new item learned. In terms of protecting the base population, this pseudorehearsal regime is highly effective, resulting in less than five unstable base population items by the time the twentieth new item has been learned.

The generation of pseudoitems is similar in spirit in both MLP and Hopfield type networks; in both cases a randomly generated “map” of the behaviour of the network is used to preserve the actual old items during further learning. In MLP networks this map (the population of pseudoitems) is the random sample of the function learned by the network. Hopfield networks, however, are best interpreted not as function approximators, but as dynamical systems where training manipulates an energy surface to create attractors for stored items. A large number of attractors exist in a trained network, including the actual learned items, and also “spurious” attractors which do not correspond to learned items but may be generated by interactions between them. In this case, the map is the randomly chosen attractors.

Pseudoitem populations generated in this way in Hopfield type networks will naturally contain duplicates of base population items, as these constitute significant attractors in the state space. In the “Pseudorehearsal” condition described above, the populations of 256 pseudoitems contained on an average duplicates of 80% of the base population. Pseudorehearsal is still slightly effective, however, even when these base population duplicates are specifically excluded from the pseudoitem population (the “Pseudorehearse * ” condition of Fig. 1). In other words, in the state space created by learning the base population, preserving attractors that do not correspond to base population item attractors (attractors usually described as “spurious” attractors or “crosstalk”) is slightly effective in preserving the base population item attractors themselves.

However, there is no need and no principled way (if we are assuming no access to the base population) of ruling out these base population duplicates, and it is of course highly beneficial to have them in the pseudoitem population. By “pseudorehearsal” in Hopfield type networks, we then mean the relearning of internally generated, randomly selected stable states (pseudoitems) rather than inputs supplied by the environment, where these pseudoitems may in fact duplicate items on which the network has actually been trained.

Note, that with some additional overhead we can generate pseudoitems which are better than random in the sense that they are closer in structure to the old items learned by a network. If some second network (possibly an unsupervised network extracting principal components) is used to learn the general structure of the base population then this network could be used to “generate” pseudoitems of a similar structure.

2.5. Unlearning in Hopfield networks

Although, not directly connected to the catastrophic forgetting literature, a mechanism called “unlearning” addresses the same issues and has been used to increase the capacity of Hopfield networks in serial/sequential learning tasks. Like pseudorehearsal, unlearning is based on randomly selecting attractors in a trained network. Having located them, however, the way they are used in unlearning is completely opposite to the pseudorehearsal approach.

The unlearning method was first proposed by Hopfield et al. (1983); Crick and Mitchison (1983). The model has been extended, particularly to the case of serial/sequential learning of interest in this paper, and explored by other authors (see e.g. Wimbauer, Klemmer and Van Hemmen (1994), Christos (1996), Van Hemmen (1997)). It is a powerful and interesting approach that has attracted a lot of attention and interest.

Unlearning is based on the standard Hopfield network using a variant of the basic Hebbian learning where pairs of units that are both in the same state increase their connection strength, and pairs of units that are in different states decrease the strength of their connections. We will refer to this type of learning as Hopfield learning. The networks also generally use symmetric weights and an activation function that returns $-1/ + 1$. The process of unlearning as originally proposed by Hopfield et al. (1983) is to:

1. start the network in a random initial state
2. relax the initial pattern to a stable state
3. unlearn the stable state using Hopfield learning with a small $-\epsilon$ so that $W_{ij} = W_{ij} - \epsilon N(\xi, \xi)$ (where $N$ is the number of units, $\xi$ the activation of unit $i$ and $\epsilon$ the learning constant).

---

4 Note that here we illustrate the case of generating a single population of pseudoitems (following most of the results presented by Robins and McCallum (1998)) instead of a new population of pseudoitems for each new item (as was the case in the MLP networks described above). Robins and McCallum (1998) discuss and explore both approaches in the context of Hopfield type networks.

5 We are grateful to Jerome Friedman (personal communication) for this suggestion.
Table 1
Percentage retrieval for a Hopfield network with 100 neurons having learned patterns until it is overloaded. Second part shows results with unlearning applied after each learned item (unlearning 50 dreams with \(-0.01\) as learning constant). Sampled with 2000 random samples after every presented pattern.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Total number of sorted patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>26.25</td>
</tr>
<tr>
<td>2</td>
<td>27.15</td>
</tr>
<tr>
<td>3</td>
<td>29.85</td>
</tr>
<tr>
<td>4</td>
<td>16.95</td>
</tr>
<tr>
<td>5</td>
<td>9.25</td>
</tr>
<tr>
<td>6</td>
<td>7.8</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>3.15</td>
</tr>
<tr>
<td>9</td>
<td>1.45</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
</tr>
<tr>
<td>11</td>
<td>0.6</td>
</tr>
<tr>
<td>12</td>
<td>0.8</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>0.1</td>
</tr>
<tr>
<td>15</td>
<td>0.2</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0.1</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Spurious</td>
<td>16.2</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>26.15</td>
</tr>
<tr>
<td>2</td>
<td>27.55</td>
</tr>
<tr>
<td>3</td>
<td>26.95</td>
</tr>
<tr>
<td>4</td>
<td>20.35</td>
</tr>
<tr>
<td>5</td>
<td>25.9</td>
</tr>
<tr>
<td>6</td>
<td>19.2</td>
</tr>
<tr>
<td>7</td>
<td>25.15</td>
</tr>
<tr>
<td>8</td>
<td>34.3</td>
</tr>
<tr>
<td>9</td>
<td>17.1</td>
</tr>
<tr>
<td>10</td>
<td>23.3</td>
</tr>
<tr>
<td>11</td>
<td>24.95</td>
</tr>
<tr>
<td>12</td>
<td>31.7</td>
</tr>
<tr>
<td>13</td>
<td>5.5</td>
</tr>
<tr>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>26.55</td>
</tr>
<tr>
<td>16</td>
<td>14.2</td>
</tr>
<tr>
<td>17</td>
<td>17.9</td>
</tr>
<tr>
<td>18</td>
<td>15.75</td>
</tr>
<tr>
<td>19</td>
<td>21.6</td>
</tr>
<tr>
<td>20</td>
<td>13.1</td>
</tr>
<tr>
<td>21</td>
<td>14.25</td>
</tr>
<tr>
<td>22</td>
<td>10.05</td>
</tr>
</tbody>
</table>

Spurious 19.35 25.15 23.6 25.65 25.15 21.95 34.3 32.35 38.95 37.55 63.7 56.4 51.15 67.75 64.2 71.3 69.5 77.45 75.6 80.15
The cycle 1–3 is referred to as a “dream”. This dream cycle is repeated many times with a small negative learning constant (from which unlearning takes its name). The amount of unlearning needs to vary depending on the correlation of the patterns. The suggested value for 50% active random patterns is about 50 dream cycles for each pattern that has been learned, with a $-0.01$ learning constant. The rationale for unlearning is that it finds and removes the spurious attractors and associations between learned states in the network, and by removing these recovers the true learned states.

The method is generally effective, significantly extending the capacity of a Hopfield network in serial/sequential learning situations (see e.g. Christos (1996), Van Hemmen (1997)). Results from our own simulations replicating the approach of Christos (1996) are shown in Table 1. After learning, the network is sampled with 2000 random inputs, which are each relaxed to a stable state. The measure of interest is the percentage of these samples that relaxes to a learned pattern (denoted in the leftmost column) versus an unlearned/spurious state (the bottom row). Without unlearning, the percentage of learned patterns/attractors retrieved decreases quickly as more and more patterns are learned. The network will almost always relax to unlearned/spurious states. With unlearning, the percentage of learned patterns retrieved is much higher, and the percentage of spurious states found correspondingly lower.

There are two distinct approaches as to when unlearning is applied. Van Hemmen (1997) lets the network learn a large number of patterns before the start of unlearning, whereas Christos (1996) performs unlearning after only a few patterns have been learned. The different techniques end up with very different results. The Van Hemmen procedure fills the network with patterns until it is overloaded and the only attractors are spurious ones. Unlearning then operates on the system removing these large spurious attractors, which are assumed to be correlations of the learned patterns. As the “dream” cycles continue, the original patterns become stable. The basins of attraction of these stable states are relatively small. Van Hemmen states that during the random sampling of the network, the attractors found will almost never be desired/previously learned states of the network.

Christos (1996) uses a different approach, applying unlearning to a network earlier while it is still able to recover some of the desired patterns. The stable states that are unlearned include some of the original patterns. With this approach, it is often the case that a recently learned pattern dominates the sample space. By unlearning these large attractors, other stored patterns become stable again, and start to recover during sampling. The capacity of this approach does not appear to be as high as that of Van Hemmen’s method, and the right level of unlearning must be used. Our results confirm Christos’ conclusion (Christos, 1996) that this type of unlearning is only able to store about the same number of patterns as weight decay over long periods.

2.6. Evaluating the pseudorehearsal and unlearning methods

It is difficult to directly compare the pseudorehearsal and unlearning methods as they have been presented to date, as the two approaches have been explored using different architectures and learning algorithms, and different ways of measuring and presenting forgetting and retention. We shall be addressing this issue in future studies. There are many open questions regarding these apparently contradictory methods, including the effect of the storage capacity of the networks, the role of noise in training and its effect on the kinds of attractors created, the type of learning algorithm, and the nature of the learning task.

While comparing the approaches conceptually, however, one of the first points that emerges is that each is addressing a subtly different problem. Using the delta rule based learning, as in both Hopfield type network simulations and in MLP networks, catastrophic forgetting is caused by the catastrophic disruption of old patterns that results when learning a new pattern disrupts the delicate balance of weights in the network (as catastrophic disruption increases, the network tends to map every input on to an essentially random output). Using the Hebbian based learning, catastrophic forgetting is caused by the catastrophic correlation between patterns that accumulates as learning new patterns continuously increases the weights in the network (as catastrophic correlation increases, the network tends to map every input to a state where all units in the network are in the same state). Pseudorehearsal addresses the catastrophic disruption problem by constraining the weights derived by new learning to produce behaviour similar to that of the original weights. Unlearning addresses the catastrophic correlation problem by explicitly reducing weights after new learning so as to reduce the correlations between patterns. Each approach, however, has some problematic features.

The error correcting delta rule based learning used in our pseudorehearsal simulations is not at all plausible in a biological context. Pseudorehearsal only works in a network where rehearsing the original items helps storage. Simple Hebbian variants do not have this property, rehearsal simply increases all weights resulting in catastrophic correlation. Most Hebbian algorithms in practice incorporate some mechanisms (such as weight decay, clipping, or normalisation) to deal with this unbounded weight increase.

While pseudorehearsal obviates the need for storing the base population separately for access during rehearsal, it does have its own, albeit more transient, storage requirement. In order to be effective, the randomly generated pseudotexts have to be stored in some way so that they can be relearned with the new items. This temporary storage of a
large population is not desirable, and we suggested (Robins, 1997) that it would be more biologically plausible to use a system which effectively uses two weights per connection, an "old" weight used to generate pseudoitems as they are needed and a "new" weight which is adapted during new learning. This proposal echoes that of Hinton and Plaut (1987).

In the initial exploration of pseudorehearsal in Hopfield type networks presented in Robins and McCallum (1998) and summarised in this paper, serial learning is not yet as effective as it is when applying pseudorehearsal in MLP networks. In MLP networks, both the base population and the new items can be well retained. In our Hopfield type simulations, the base population is well retained but it is difficult to maintain good retention of new items learned early in the sequence. This limitation is explored further in Robins and McCallum (1998) and we hope to address it in future research.

The unlearning method is not without its own less desirable features. The network architecture and the learning algorithm are as far removed from real biology as the networks used to simulate pseudorehearsal. Crick and Mitchison (1995, pp. 150–151) discuss some of these unrealistic features (such as the use of symmetric connections). To this list we would add the note that the current simulations of unlearning are restricted to cases where the items (patterns of activation) to be stored by the network contain almost exactly 50% “active” elements. Any deviation from this restriction leads to the difficult task of predicting a specific unlearning constant for each input based on a measure of its correlation with other learned patterns. A further general point about unlearning is that it does not actually maintain or strengthen learned information at all, it simply helps to “restore” it by weakening other information which has been more recently superimposed. In this case, damage done to old information (the basins of attraction) accumulates and that the old information is eventually lost permanently. We now turn to some other more specific issues.

Using Van Hemmen’s method there are too many stable states. Van Hemmen (1997) states that during random sampling the network never relaxes to a learned/desired pattern. While unlearning moves the attractors around “The huge number of attractors does not decrease” (Van Hemmen, 1997; p. V9). This leaves the energy surface with thousands of attractors other than the ones originally learned. Van Hemmen escapes the criticism of “a needle in a haystack” by an interesting side effect of synchronous updating, namely that random inputs often relax into 2-cycle states. These states become very common after unlearning, and Van Hemmen claims that they make up 100% of the randomly sampled patterns. The only patterns that do not relax to this type of state after unlearning are the most recent block of the learned patterns that remain stable in the conventional sense. The 2-cycle artefact is unlikely to be biologically plausible, therefore we feel that this form of unlearning still has the problem of overly numerous stable states. From our exploration of this issue using asynchronous update, Table 2 shows that the number of distinct stable states found from 2000 random samples. The values are for before unlearning where the network has learned 40 patterns using the Hopfield learning algorithm, and after unlearning 2800 dream states with a −0.01 learning constant. Note, that every random input found a distinct stable state after unlearning. It also shows the average distance between the initial random state and the relaxed stable state. This average of 35.70 gives an estimate as to the total number of stable states as around one million. When only 40 of these are the desired patterns, there is a major problem of differentiating the real from the spurious states. The requirement that the network is overloaded and unstable before the unlearning process takes place makes this approach difficult to use in an online context and biologically implausible. With this condition, the only time the network is useful is immediately after an unlearning cycle.

The method proposed by Christos (1996) is to perform unlearning while the network is still able to recover some of the original patterns (the results shown in Table 3 can be compared with Van Hemmen’s method in Table 2). There are also some problems with this procedure. Most significantly this method results in the frequent unlearning of states that have just been learned. This results in changes which are functionally very similar to simple weight decay, as is also the case (albeit less obviously) in Van Hemmen’s method. Tables 2 and 3 show the average of the absolute value of the weights before and after unlearning. If unlearning is at least partially doing a decay-like change the question is what happens when this is removed? We have explored this issue in our own simulations by allowing unlearning to rearrange weights, while normalising them so that on an average they do not decay. Using both Van Hemmen and Christos type unlearning, the network became unable to learn new patterns. Thus one of the significant aspects of how unlearning helps serial learning is just a weight decay effect.

The amount of unlearning required is a delicate matter for

### Table 2

<table>
<thead>
<tr>
<th>Distinct stable</th>
<th>Distance</th>
<th>Weight average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>444.40</td>
<td>42.59</td>
</tr>
<tr>
<td>After</td>
<td>2000.00</td>
<td>35.70</td>
</tr>
</tbody>
</table>

Table 2 Hopfield learning of 40 patterns followed by dreaming phase of 2800 cycles with unlearning of −0.01

### Table 3

<table>
<thead>
<tr>
<th>Distinct stable</th>
<th>Distance</th>
<th>Weight average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>122.43</td>
<td>41.83</td>
</tr>
<tr>
<td>After</td>
<td>610.45</td>
<td>41.87</td>
</tr>
</tbody>
</table>

Table 3 Hopfield learning of one pattern followed by dreaming phase of 50 cycles with unlearning of −0.01
3. ANN accounts of consolidation during sleep

Both the pseudorehearsal (Robins, 1996) and the unlearning (Crick & Mitchison, 1983, 1995) methods of managing catastrophic forgetting/sequential learning in ANNs have been linked to a popular hypothesis about the function of sleep, namely that newly learned information is consolidated into long term memory during sleep. In this section, we describe this sleep consolidation hypothesis and the evidence that supports it, show how we believe the catastrophic forgetting problem in ANNs motivates a need for a consolidation process in the brain, and compare the pseudorehearsal and unlearning methods as candidates for a computational account of the process.

3.1. The consolidation of learning during sleep

There are many theories on the function of sleep, ranging from the physiological to the psychoanalytic (see e.g. Moffitt & Kramer, 1993; Empson, 1993; for overviews of research on sleep and dreaming). While sleep may serve many functions we are interested in exploring one that is a simplification of the complex interactions that occur in the brain, and to a certain extent it is necessary to abstract away the detailed mechanics of each method and examine the general principles involved.

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 permanent 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 et al., 1993, p. 356).

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 later.

Although sleep consolidation has been most often linked to REM sleep, some studies have emphasised the role of another sleep stage, SWS (slow wave sleep). Winson (1990), for example, reviews studies relating to: the hippocampal theta rhythm that occurs during the apprehension of significant or changing environmental information and also during REM sleep (suggesting a link between waking and sleeping learning mechanisms); place field cells that fire maximally as an animal explores a new environment and become active again during sleep; 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. Hennevin et al. (1995) divide their review into sleep deprivation studies and other studies. While the evidence from deprivation studies is complex, REM sleep deprivation often results in impaired memory retention performance. Of the other studies reviewed a consistent observation is that after animals have been exposed to training, they spend increased time in subsequent REM sleep. Smith (1996) reviews similar material, and also has evidence showing that the neural changes usually observed following exposure to enriched environments are disrupted by REM sleep deprivation. Smith (1996) summarises results from a number of studies relating to REM sleep and learning, concluding with the following summary points:

- REM sleep increases or increases in number of REMs appear after substantial learning has taken place.
- These REM sleep/REMs increases appear to persist for many hours or days after the end of acquisition.
- 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 permanent learning-memory deficits. These vulnerable time periods have been named REM windows.
- 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 et al., 1993, p. 356).

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 later.
environment were more likely to fire together during SWS. Further evidence on the role of SWS was reviewed by Kavanau (1997).

In addition to these studies of the process, one neural structure in particular is frequently identified as having a central role in sleep consolidation. The hippocampus and associated structures of the limbic system are implicated in many cognitive functions, particularly memory processes and spatial cognition. Evidence supporting the central role of the hippocampus in sleep consolidation is described by Winson (1990), Wilson and McNaughton (1994), McClelland et al. (1995), Kavanau (1997), Buzaski (1998). Wilson and McNaughton (1994), Kavanau (1997), Buzaski (1998) all conclude that information appears to be “played back” or transferred from the hippocampus to the neocortex during sleep:

...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)

It should also be noted that the hippocampal theta rhythm which occurs during sleep appears to facilitate in the hippocampus and other structures the LTP (Long Term Potentiation) mechanism assumed to underly learning (Winson, 1990; Kavanau, 1997).

### 3.2. Catastrophic forgetting and consolidation

Sleep is a complex phenomenon likely to have multiple functions. Accepting that the consolidation of newly learned information is one of these functions, some obvious questions arise. Why is a special process required to consolidate new information (as opposed to just integrating new information directly into long term memory as it is learned), and why does this process occur during an “off line” state? These questions are not often addressed in the consolidation literature. We have suggested however (Robins, 1996), that the catastrophic forgetting problem provides a compelling answer to the first of these questions and a reasonable hypothesis about the second. An evolutionary perspective also sheds light on both issues (see Section 4).

In ANNs, the catastrophic forgetting problem is a basic consequence of the plasticity of the representing medium, connection weights. As such it is a broad and pervasive problem, requiring (in most ANN types) a special mechanism such as pseudorehearsal or unlearning to manage the integration of new learning. As neural representations are also thought to be mediated by plastic connection weights (synapses), it seems probable that a version of the catastrophic forgetting problem has also been encountered during the evolution of the brain (see e.g. Winson (1990) and Kavanau (1997)). If this is the case then the brain must have evolved its own mechanism to manage the integration of new learning.

We suggest that in mammals sleep consolidation is exactly this mechanism, or in other words that the reason for the sleep consolidation process is to manage the catastrophic forgetting problem in the brain. We further suggested (Robins, 1996) that the pseudorehearsal mechanism provided a good account of the computational process of sleep consolidation. The major focus of the current paper is to compare (in the context of Hopfield type networks) this pseudorehearsal account of consolidation to the unlearning method that has also been linked to sleep (Crick & Mitchison, 1983, 1995).

McClelland et al. (1995) follow a similar chain of reasoning to a related conclusion. Having also considered the implications of the catastrophic forgetting problem in ANNs as a general problem for learning in brains, the authors conclude that “complementary” learning systems are required. McClelland et al. propose that the hippocampal system allows for the rapid learning of new information, and that this new information is then “reinstated” by the hippocampus in the neocortex and slowly integrated with old information/long term memory (using “interleaved learning” i.e. full rehearsal of the old information). This account obviously fits well with the evidence suggesting a transfer of information from the hippocampus to the neocortex during sleep (as discussed in Section 3.1). Although they do not make a strong connection with the sleep literature, the authors note that the integration process they propose is assumed to occur “in off-line situations including active rehearsal, reminiscence, and other inactive states including sleep” (McClelland et al., 1995; p. 424).

### 3.3. The pseudorehearsal account of sleep consolidation

As noted above we have already claimed that the pseudorehearsal mechanism provides a plausible account of the computational process of consolidation during sleep (Robins, 1996). This claim was based on MLP (back-propagation) based simulations. We suggest, however, that the account generalises easily to the Hopfield type networks of interest in the current paper (catastrophic forgetting and pseudorehearsal are robust and general processes which have been replicated using a variety of network architectures and training populations). Beyond their common function of integrating new and old information, there are a number of parallels between pseudorehearsal and sleep consolidation.

Pseudorehearsal approximates old information by using randomly constructed inputs to sample the behaviour of a network. While it is not currently possible to confirm in detail that sleep consolidation is using this mechanism, there is evidence that is at least suggestive of such a process. Moffitt and Kramer (1993; pp. 3–4) cite the apparent
randomness of dreaming at the physiological level. 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. Kavanau (1997) notes that PGO spikes in animals exhibit parallels with intense dreaming.

Pseudorehearsal uses the approximations of old information as a context within which new information must be integrated during learning. If the random stimulation of the neocortex during sleep is evoking old information, then the evidence relating to the sleep consolidation hypothesis and particularly the transfer of information from the hippocampus to the neocortex supports the proposal that new information is being integrated into this context.

Catastrophic forgetting is worst when new inputs/output mappings are significantly different from those already learned by the network. Catastrophic forgetting does not occur in MLP networks (we have not yet confirmed this in Hopfield type networks) when new information is added which is consistent with the information already learned by the network (Robins, 1996). Similarly, sleep consolidation does not appear to be necessary for new learning which is easy or consistent with an animals’ existing knowledge (see the fourth point made by Smith (1993) noted in Section 3.1).

3.4. The unlearning account of sleep consolidation

Since the unlearning method was first described (Hopfield et al., 1983; Crick & Mitchison, 1983) it has been suggested to be a process which occurs during sleep (Crick & Mitchison, 1983). For most of the subsequent work on the specific properties of unlearning in Hopfield networks as reviewed above, the connection to sleep is often stated (and the unlearning episodes are referred to as dreams) but seldom pursued in great depth. Crick and Mitchison (1995) provide a detailed discussion of the unlearning (which they call “reverse learning”) account in the context of the sleep literature, but they abstract unlearning away from the detailed mechanisms of artificial networks to a much more general set of principles.

Crick and Mitchison (1995) suggest that the unlearning account is attractive because it is based on very general principles of brain function. The brain appears to be an associative memory system that can be at least approximated by simple associative ANNs. Such networks store information in a way which is robust, distributed, superimposed, and content-addressable. Overloading such a network often produces outputs, which are a combination of stored states. The authors focus on REM sleep in particular and suggest that unlearning during REM sleep “will reduce somewhat the mixed outputs produced by overlapping memories, while leaving intact the unmixed memories which the net was supposed to store” (Crick & Mitchison, 1995; p. 149). As in the pseudorehearsal account, the random PGO inputs from the brain stem are identified as the source of random input.

Crick and Mitchison suggest that the function of REM sleep is more likely related to the host of unremembered dreams than it is to the few remembered ones, and that REM dreams could be called “bizarre intrusions”. In discussing the alternative account of Davis (1985), who (much like the pseudorehearsal proposal) suggests that memories are relearned during sleep (so that synapses that are weakening due to metabolic turnover can be renewed) Crick and Mitchison are adamant that bizarre REM dreams (random attractors) would be a “very odd choice to reinforce”. It is the bizarre content of these dreams and their proposed relationship to the undesirable overlapping attractor states of the network memories that motivates the unlearning account. Such dreams/states are undesirable and should not be reinforced, but removed from the memory/network.

Crick and Mitchison admit that the most powerful criticism of the unlearning account is that deprivation of REM sleep does not produce the effects that the theory would predict. Oddly, however, they note (p. 152) that “Because our brain is so very complicated it is not clearly exactly what these predictions should be...”, before moving on to briefly considering an increase in imagination or fantasy as a possible consequence. We return to this topic below.

Brown (1993) also explores the unlearning account as a possible function of sleep. He suggests that if the theory is correct then it might lead to an understanding of why lithium salts terminate manic episodes, and that abnormalities of unlearning might account for some aspects of schizophrenia, mania, and depression.

3.5. Comparing the pseudorehearsal and unlearning accounts

Current ANNs are in general very gross abstractions of actual brain function. As Crick and Mitchison (1995; pp. 151–152) also note, many of the specific assumptions made by ANN models are directly at odds with observed facts about the brain. As described in Section 2.6, both the pseudorehearsal and unlearning mechanisms as they are currently explored in our ANN models make their fair share of biologically implausible assumptions. To a certain extent however, as with most ANN modelling, it is necessary to put aside the specifics of the artificial models and focus on the more general principles. Bearing this in mind, we now compare pseudorehearsal and unlearning as possible accounts of the computational process underlying the consolidation of learning during sleep.

We turn at first to Crick and Mitchison (1995) discussion of the relearning of random attractors (bizarre REM dreams in their terms). As noted above, they regard random attractors as an odd candidate for relearning or strengthening, whereas this is exactly what is predicted by the
Section 3.4 showed that relearning actually does any damage, particularly if any given spurious attractor was only “found” and relearned infrequently (as would be the case in networks with a large number of shallow spurious attractors). Indeed, the pseudorehearsal method described in Section 3.4 showed that relearning only such spurious attractors in fact protects previously learned attractors during further learning, presumably by maintaining learned attractors as “overlap” between the spurious attractors which were being rehearsed.

We suggest that the opposite of Crick and Mitchison’s claim is true, that unlearning randomly chosen attractors is the most unlikely mechanism. While the predictions of the different specific approaches to unlearning differ (Section 2.5), it is clear that much of unlearning in general must be directed at the actual learned states that the network is supposed to retain. Indeed, in Christos’ approach (Christos, 1996), by far the majority of unlearning episodes are directed at unlearning exactly the most recently learned item (which is the strongest attractor in the network), while Van Hemmen (1997) requires the effective obliteration of recall in the network as part of the training process. The continual unlearning of recently learned information hardly seems like an efficient or parsimonious way to organise a learning system, especially when it seems in some cases that the advantages which are conferred by unlearning are almost indistinguishable from the consequences of the much simpler process of weight decay. Unlearning also requires a new mode of synaptic modification (currently unspecified but completely the reverse of the usual Hebbian strengthening), whereas the sleeping hippocampal theta rhythm appears to facilitate standard LTP (Winson, 1990; Kavanau, 1997). In contrast, the relearning of information proposed by the pseudorehearsal approach requires no new synaptic mechanisms, reinforces rather than weakens the desired information, is consistent with LTP facilitation during sleep, and is consistent with the explicit rehearsal of target information sometimes observed as a conscious mnemonic strategy.

Some of the strongest evidence for the pseudorehearsal account comes from the data relating to the role of the hippocampus in consolidation. McClelland et al. propose that the hippocampus serves as a fast temporary memory system, and that the new information stored there is later transferred to long term memory/the neocortex. This is consistent with the observations regarding the function of the hippocampus during sleep (Winson, 1990; Wilson & McNaughton, 1994; McClelland et al., 1995; Kavanau, 1997; Buzaski, 1998), particularly the apparent transfer of information from the hippocampus to the neocortex as described in Section 3.1. Pseudorehearsal presupposes and requires exactly this kind of complementary learning system, new information has to be stored somewhere separate until it can be integrated with long term memory during sleep. The unlearning account does not require such machinery. New information is stored directly in the long term memory (the same network) and the purpose of the unlearning consolidation process is to “clean up” or maintain this single memory system. This leaves the well established role of the hippocampus during sleep unexplained and unmotivated in the unlearning account; what would be the point of such a complex and costly system?

While Crick and Mitchison were wary of making psychological predictions about the psychological consequences of unlearning based consolidation, some general predictions seem to follow quite directly. An unlearning account appears to predict that the most difficult information to learn will be information that is structured or related, as this will create the largest number of undesirable overlapping attractors. Information that is unrelated will be easiest to learn as this will create distant attractors, which do not interfere with each other. In contrast, the robust and general finding in the psychological literature is that structured information that is similar to what is already known is easiest to learn and retain. The predictions of the pseudorehearsal account are consistent with this finding. Unlearning also appears to predict that the deprivation or disruption of consolidation will result in the loss (“overwriting”) of older learned information, while the most recently learned information is retained. In contrast, the psychological evidence (such as the deprivation studies noted above) shows that disrupting consolidation does not affect long term memory, but damages the retention of new learning. Assuming that pseudorehearsal occurs in the context of the kind of complementary learning systems described above, then the predictions of the pseudorehearsal account are again consistent with the psychological findings.

Finally, both pseudorehearsal and unlearning share certain predictions. Both accounts postulate significant random and unfocused “activation” of long term memory. It could certainly be argued that this would cause severe disruption to ongoing cognition and necessarily make sleep an “off line” state. A similar argument is made from an evolutionary perspective by both Winson (1990) and Kavanau (1997). Both accounts also predict that dreams will frequently be related to recent learning and experiences.
summarise the points made and substantiated by Kavanau. Relevant to our current proposals, and here we simply memory. Among the many topics reviewed, several are comprehensive, detailed, and documented review of the factors approach. Kavanau (1997) provides a thoroughly comprehensive, detailed, and documented review of the factors

3.6. Discussion

In summary, both the pseudorehearsal and the unlearning accounts of managing catastrophic forgetting/sequential learning in ANNs have been linked to the consolidation of new information during sleep. In this section, we have compared both accounts, and claimed that the pseudorehearsal method has several advantages as a framework for modelling consolidation. As we shall discuss in Section 4, a consideration of the evolutionary evidence also appears to favour the neurological plausibility of the pseudorehearsal approach.

We suggest that the evidence reviewed above supports a model of learning and sleep consolidation that combines both functional and structural elements in a systematic and well-motivated whole. During waking cognition new information which is difficult or novel is typically stored in the hippocampus (and associated structures). During sleep, this new information is integrated in to long term memory in the neocortex. As part of this process, the neocortex is randomly stimulated and samples or approximations of old information are relearned in a pseudorehearsal like process. This process provides a buffering effect that enables the new information being replayed from the hippocampus to be integrated in to long term memory without causing the catastrophic forgetting of old information that would otherwise result.

There are a range of other suggestions linking ANNs and sleep (e.g. Hinton & Sejnowski, 1986; Gloubs, 1993; Sutton, Mamelak & Hobson, 1992; Hobson, Mamelak & Sutton, 1992). Two recent studies (NakaoHonda, Musila & Yamamoto, 1997; Horn, Levy & Ruppin, 1998) have even explored the use of random inputs/noise in ANN based models of sleep in ways which are quite different from either the pseudorehearsal or the unlearning accounts. Such alternatives are not necessarily mutually exclusive. A range of interactions is certainly possible between ANN mechanisms and the complex processes of sleep and learning.

4. An evolutionary perspective

A brief examination of the evolutionary evidence also seems to support the plausibility of the pseudorehearsal approach. Kavanau (1997) provides a thoroughly comprehensive, detailed, and documented review of the factors relating to the evolution of sleep and its relationship to memory. Among the many topics reviewed, several are relevant to our current proposals, and here we simply summarise the points made and substantiated by Kavanau.

Beyond the ability to respond to the environment with simple, fixed, reflexive actions, the origin of activity dependent synaptic plasticity was the first evolutionary adaptation of the nervous system. When such synaptic modification (strengthening) became possible, the maintenance of learned changes, particularly in circuits that were not used frequently, was probably achieved by internally mediated repetitive activation, a process known as "dynamic stabilisation". With the evolution of more complex sensory and behavioural capabilities, brain complexity increased dramatically, as did the need for storage and the maintenance of learned information. Kavanau suggests that this resulted in a conflict between the requirements of the dynamic stabilisation of memory circuits and the requirements of the central processing of sensory (particularly visual) inputs. This may have resulted in a selective pressure for lowered sensory processing states that were the evolutionary precursor of primitive sleep. This evolutionary trend can be traced further so that:

Pursuing the logical implications and consequences of long-term maintenance of synaptic efficacy by DS [dynamic stabilisation] has led to the formulation of a chain of causal evolutionary links, from the development of activity-dependent plasticity of synapses in simple metazoans to the neural adaptations represented by REM sleep in mammals and birds. (Kavanau, 1997; p. 9).

In humans:

...DS [dynamic stabilisation] largely supports the establishment and maintenance of phylogenetic and ontogenetic memory circuits in the infant child during sleep states. Circuit consolidation and maintenance appear to be accomplished by the same means in the adult. (Kavanau, 1997; p. 27).

This evolutionary perspective obviously provides strong support for part of the pseudorehearsal account, namely that the function of the random activation of memory is to strengthen or relearn information. This function can be traced from simple synaptic mechanisms in the most primitive of nervous systems in a continuous evolutionary path to the properties of sleep in birds and mammals.

Kavanau frequently mentions the replay of newly learned information from the hippocampus but does not trace the evolution of this mechanism. If dynamic stabilisation represented the evolutionary solution to the “first” problem of learning, how to maintain infrequently used learned memory circuits (and thereby led to the development of primitive sleep states), we suggest that the function of the hippocampus as a temporary store for new information is part of the evolutionary solution to the “second” problem of

---

6 Since the spontaneous activation of memory circuits maintains synaptic efficacy but does not trigger circuit functions (which are often suppressed by increased activation thresholds), this process is usually referred to as “non-utilitarian” dynamic stabilisation.
learning, how to integrate new information into old without causing catastrophic forgetting (and thereby lead to the complex processes of sleep consolidation explored in this paper). We hope to explore the evolutionary evidence relating to this proposal in future work.

5. Summary

In this paper, we have described the catastrophic forgetting problem, particularly in the context of Hopfield type networks. Catastrophic forgetting is a broad and general problem for ANNs and any learning system based on a plastic representing medium. Unless the problem is addressed or managed somehow, then it is not possible to learn in a sequential/ongoing way as newly learned information disrupts or even eliminates information previously learned by the system. We have described and compared both the pseudorehearsal and unlearning methods as possible solutions to the catastrophic forgetting problem in Hopfield type networks.

Both pseudorehearsal and unlearning have been linked to the process of the consolidation of learning during sleep. We have described the sleep consolidation process and the evidence that supports it, shown how we believe that the catastrophic forgetting problem in ANNs motivates the need for a consolidation process in the brain, and compared the pseudorehearsal and unlearning methods as candidates for a computational account of the process. We have claimed that the pseudorehearsal account is better supported by the psychological, neurological, and evolutionary data than is the unlearning account.

In short, we suggest that any brain-like (ANN based) learning system will need a sleep-like mechanism for consolidating newly learned information if it wishes to cope with the sequential/ongoing learning of significantly new information. We suggest that the pseudorehearsal-based model described in this paper is a plausible candidate for a computational account of the process of sleep consolidation.

References

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interfer-


