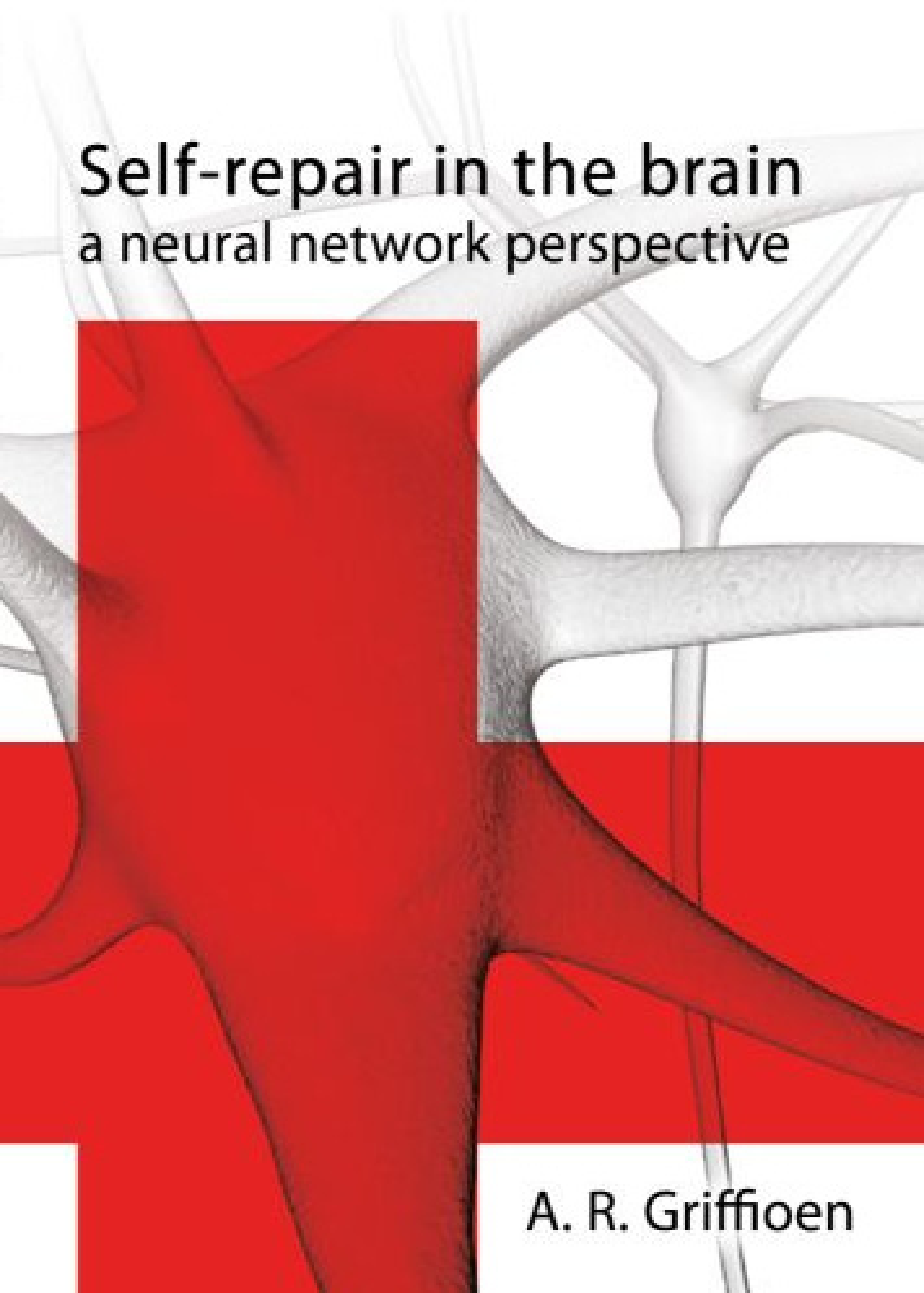
Self-repair in the brain : a neural network perspective

Griffioen, A.R.

Publication date
2008
Document Version
Final published version

Self-repair in the brain
a neural network perspective



A. R. Griffioen

Self-repair in the brain: a neural network perspective

Printed by PrintPartners Ipskamp

Lay-out cover: Marco de Stefanis

Dankwoord

Allereerst dank aan al degenen die een bijdrage hebben geleverd aan mijn proefschrift. Dit zijn natuurlijk mijn promotoren Jaap Murre, Jeroen Raaijmakers en de overige leden van de commissie. Jaap voor het beschikbaar stellen van mijn positie, al zijn aanwijzingen en correcties. Jeroen voor zijn goede suggesties om de hoofdstukken tot een geheel te smeden. Met Tony heb ik samen gewerkt. Naast zijn vriendschap heb ik ook nog veel van hem opgestoken. Ik heb inzicht verkregen in de kunst van het mathematisch modelleren en de kracht en grenzen hiervan. Ook dank ik mijn mijn stagiares Wouter en Leendert voor hun bijdrage.

Verder gaat mijn dank uit naar al degene die mij hebben geholpen bij het schrijven. Allereerst Ina die een aantal hoofdstukken voornamelijk op het engels zou corrigeren. Uiteindelijk heeft ze ook de inhoud flink aangescherpt. Dit maakte mij duidelijk van wie mijn vriendin haar scherpte heeft geërfd. Verder hebben mij geholpen met het verbeteren van de tekst Wery, Raoul en Marije.

Dan dank voor alle gezelligheid die ik aan de UvA heb genoten. Hiervoor waren o.a. mijn kamergenoten verantwoordelijk. In chronologische volgorde waren dit eerst Martijn M., Paul, Robert B. en de binnenwaaiende Martijn B.. Later werden dit Frits, Janneke, Pedro en Steve. Met Martijn M. heb ik nog een cursus ontwikkeld en deze een aantal jaar gegeven. Met hem samenwerken is niet zo'n kunst, want als je niet uitkeek hoefde je niets te doen. Naast mijn kamergenoten ben ik verder enige tijd intensief optrokken met Hilde, Ingrid, Lourens, Lucia, Niels, Raoul en Wery. Bedankt voor die mooie tijd! Voor alle overige UvA gerelateerde activiteiten zoals EPOS uitjes, hacky sack, klim- en skatepartijen bedank ik: Bjørn, Denny, Durk, Erik-Jan, Han, Hedderik, Ingmar, Jennifer, Maarten, Maartje, Michiel, Noortje, Pauline en Sander.

Dan mijn familie, ik had Ina al bedankt voor al haar corrigeerwerk (maar ze is natuurlijk ook een fantatische schoonmoeder). Mijn beider ouders voor al het oppas-werk, in het bijzonder mijn moeder die heel wat papa-dagen in oma-dagen heeft veranderd, zodat ik aan mijn proefschrift kon werken. En natuurlijk Marije mijn supervrouw, die naast haar eigen promotie, baan en kinderen toch ook nog tijd had voor haar 'mannetje'.

Amsterdam, april 2008
Robert Griffioen

Self-repair in the brain: a neural network perspective

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 5 juni 2008, te 10.00 uur

door

Albert Rob Griffioen
geboren te Djakarta, Indonesië

PROMOTIECOMMISSIE

Promotores:	Prof. dr J.M.J. Murre
	Prof. dr J.G.W. Raaijmakers
Overige leden:	Dr. C. van Heugten
	Prof. dr. V.A.F. Lamme
	Prof. dr. H. van der Maas
	Prof. dr. E.O. Postma
	Dr. M.E.J. Raijmakers
	Prof. dr. B. Schmand

Universiteit van Amsterdam
Faculteit der Maatschappij- en Gedragswetenschappen

1. INTRODUCTION **9**

1.1 INTRODUCTION: SELF-REPAIR AS MAINTENANCE THROUGH REDUNDANCY 9
1.2 THE SELF-REPAIR MODEL 10
1.2.1 Modeling the triage of recovery 10
1.2.2 Autonomous self-repair in connectionist models versus autonomous self-repair in the brain 13
1.3 CHAPTER OVERVIEW 14
1.3.1 Goals of the thesis 14
1.3.2 Detailed chapter overview 15

2 **19**

THE SELF-REPAIRING BRAIN: A SYNTHESIS **19**

2.1 INTRODUCTION 20
2.2 THE ADVANTAGE OF SELF-REPAIR 20
2.3 NEUROBIOLOGICAL EVIDENCE OF SELF-REPAIR 21
2.3.1 Neurobiological evidence of memory redundancy 21
2.3.2 Neurobiological evidence of memory maintenance 24
2.4 BEHAVIORAL CORRELATES OF SELF-REPAIR 26
2.4.1 The use-it-or-lose-it principle 27
2.5.2 The serial lesion effect 30
2.5 SELF-REPAIR IN CONNECTIONIST SYSTEMS 33
2.5.1 Redundancy in a connectionist network 33
2.5.2 Maintenance in connectionists networks 37
2.5.3 Demonstration of self-repair in a connectionist network 39
2.6 DISCUSSION 40

3 **43**

SELF-REPAIRING NEURAL NETWORKS **43**

3.1 SELF-REPAIR AS MAINTENANCE OF REDUNDANCY 44
3.2 SELF-REPAIR AND RANDOM GRAPH THEORY 45
3.3 SELF-REPAIR IN HOPFIELD NETWORKS 51
3.4 SELF-REPAIR IN THE 'CORTEX' PART OF THE TRACELINK MODEL 55
3.5 DISCUSSION 60
APPENDIX A: ACTIVATION RULES OF THE TRACELINK MODEL 62
Activation rule 62
Threshold control 62

4 **65**

ANALYSIS OF RANDOM CUED SELF-REPAIR IN FEEDFORWARD CONNECTIONIST SYSTEMS **65**

4.1 INTRODUCTION 66
4.2 THE ASSOCIATIVE MEMORY MODEL 69
4.3 RETRIEVAL UNDER LEARNING AND LESIONING 71
4.3.1 Introduction: Retrieval probability 71
4.3.2 Weight differences: the effect of learning and lesioning on system stability 72
4.3.3 The influence of the activation probability p on system stability 80
4.3.4 The effect of pattern size |P| on system stability 83
4.4 DISCUSSION 85
APPENDICES 89
Appendix A 89
Appendix B. The effect of different step sizes of learning rate and lesion size 90

5 **93**

INVESTIGATING SELF-REPAIR IN A CORTICAL NEURAL NETWORK **93**

5.1 INTRODUCTION 94
5.2 METHODS 95
5.2.1.1 The neural network 95

5.2.1.2 The neuron model ... *97*
5.2.1.3 Hebbian learning ... *98*
5.2.2 Simulation procedure ... *99*
5.2.3 Analysis of memory representations ... *100*
5.3 RESULTS ... 101
5.3.1 Demonstration of self-repair in the cortical neural network ... *101*
5.3.2 Investigating the effect of the amount of self-repair and amount of damage ... *102*
5.4 DISCUSSION ... 104
APPENDICES ... 107
Appendix A. Network connectivity ... *107*
Appendix B. Neural parameters ... *108*
Appendix C. Simulations to find the appropriate stimulus intensity ... *108*
Appendix D. The cluster algorithm ... *111*
Appendix E. Exploratory search of the amount of self-repair and amount of damage ... *112*

6 SELF-REPAIR OF NEURAL CIRCUITS DURING SLEEP ... 115

6.1 INTRODUCTION ... 116
6.2 THE SELF-REPAIR SLEEP MODEL ... 118
6.3 SELF-REPAIR SLEEP SIMULATION ... 120
6.4 MEMORY MAINTENANCE AND MEMORY CONSOLIDATION ... 122
6.4.1 Theories of memory maintenance ... *123*
6.4.2 Theories of memory consolidation ... *125*
6.5 REDUNDANCY ... 126
6.6 SELF-REPAIR: MAINTENANCE OF REDUNDANCY ... 128
6.6.1 Self-repair and memory maintenance ... *128*
6.6.2 Self-repair and memory consolidation ... *130*
6.6.3 Experimental data of memory maintenance and memory consolidation supporting self-repair ... *131*
6.7 DISCUSSION ... 134
APPENDICES ... 138
Appendix A. Connectivity of the model ... *138*
Appendix B. The neural model and plasticity rule ... *139*

7 DISCUSSION ... **143**

7.1 INTRODUCTION ... 143
7.2 SMALL DAMAGE: THE NECESSITY OF SELF-REPAIR ... 144
7.3 TESTING THE SELF-REPAIR HYPOTHESIS ... 146
7.4 APPLICATION OF MODELS OF SELF-REPAIR: MODELS OF BRAIN RECOVERY ... 148
7.5 FUTURE RESEARCH ... 151

REFERENCES ... **155**

DUTCH SUMMARY ... **167**

1. Introduction

1.1 Introduction: self-repair as maintenance through redundancy

This thesis focuses on self-repair as maintenance through redundancy and investigates the hypothesis that this type of self-repair is a property of neural networks of the brain. Self-repair-through-maintenance is ubiquitous in nature. It can be found in nearly all organisms on many different levels ranging from amoebe to primates. Indeed, the very fundament of genetic evolution, the double helix structure, allows DNA to repair itself (Ayala & Kigger, 1982). A striking example that nicely illustrates the power of self-repair is the *Dienococcus radiodurans*, a radiation-resistant bacterium that is able to survive under conditions of starvation and oxidative stress. Its DNA self-repair and genetic redundancy enable this organism to withstand severe ionizing and ultraviolet irradiation effects (White *et al.*, 1999). A more common example of self-repair at the molecular level is presented by the human skin, which repairs itself in a process of replacement that goes on continuously, even in the absence of trauma, although there is considerable fine-tuning depending on wear and tear. Skin also repairs areas of lost tissue.

There is ample evidence of recovery after brain injury (Bach-y-Rita, 1990; Cotman & Nieto-Sampedro, 1982; Elbert & Rockstroh, 2004; Kolb, 1995; Kolb *et al.*, 1987; Kolb *et al.*, 1997; Marshall, 1984; I. H. Robertson & Murre, 1999; Taub *et al.*, 2002) suggesting a self-repair capacity of the brain. Based on such data Robertson & Murre (I. H. Robertson & Murre, 1999) distinguishes a triage of recovery: autonomous recovery, guided recovery, and compensatory recovery. Each type of recovery is associated with a type of lesion: autonomous recovery with a mild lesion, guided recovery with a moderate lesion, and compensatory recovery with a severe lesion. A mild lesion will recover spontaneously and specific external stimulation or a rehabilitation program is unnecessary. With a moderate lesion, representations are potentially reusable and restitution may be possible given appropriate type, timing, and frequency of stimulation. A severe lesion will not recover and only compensation by other brain areas is possible.

To explain the data of the different types of recovery Robertson & Murre (I. H. Robertson & Murre, 1999) proposed a model of self-repair by maintenance through redundancy. With a mild lesion, sufficient redundancy is still present and the brain can recover with the available information present in the damaged neural circuits. In case of a moderate lesion, redundancy in the neural circuits is insufficient and a rehabilitation program,

which can provide the missing information, is needed to restore the damaged neural circuits. In case of a severe lesion, there is insufficient redundancy left and no rehabilitation program can restore the damaged neural circuits. If the wrong type of stimulation or rehabilitation program is applied to a lesion it may result in maladaptive repair. In this case erroneous information is fed into the repair process and wrong connections are enhanced, which may lead to brain disorders like phantom limb disorder (Elbert & Rockstroh, 2004; I. H. Robertson & Murre, 1999). One goal of this thesis is to develop a model that is able to model the triage of recovery. The development of this model will contribute to the other goal of this thesis, namely to demonstrate that self-repair is possible in the brain. These goals will be further explained in Section 1.3. In the next section, we will present the connectionist model that can model the triage of recovery. This model is based upon the idea of maintenance through redundancy. We call this the self-repair model.

1.2 The self-repair model

1.2.1 Modeling the triage of recovery

In Chapter Two, we will construct a connectionist model of self-repair by maintenance through redundancy based upon empirical data. In this model redundancy resides in its connectivity. The capability of repair is provided by plasticity mechanisms that restore connectivity. Self-repair is the continuous repair over time. The self-repair model consists in addition to repetitive repair of repetitive lesions. To simplify the model, we regard it as consisting of consecutive lesion-repair cycles similar to the procedure of serial lesion experiments (see Chapter Two). In a cycle repair or a lesion can be omitted, which results in different frequencies of self-repair and lesions over time.

A lesion is modeled by adding a (negative) number to the connections. Self-repair is modeled by feeding the network with a stimulus and applying a given plasticity mechanism to the connections, which also results in a change of the connections. The triage of recovery can be modeled by having different types of lesions, different types of self-repair, and manipulating frequency of damage and self-repair. Below we will discuss in more detail how lesions and self-repair are modeled. One has to keep in mind, however, that the way they can be modeled depends on the type of connectionist network. More neuro-biological detailed models usually have more options and parameters to model them.

To model a particular type of lesion one can vary parameters of lesion size and lesion distribution. The size of a lesion can be for instance a number drawn from a uniform

distribution between a minimum and a maximum number. Another possibility is to put some of the connections to zero. In this thesis, putting the weights of a particular neuron to zero means that the synapses are lost, but because we simulate synaptic re-growth their values can increase again. The lesion distribution determines which portion of connections will be hit by a lesion. For instance, damage will hit an area of the network containing a specific set of memory representations.

Self-repair has as variables the stimulus type and type of plasticity mechanism, where each variable can have several parameters. The most important parameter for stimulus type is how similar or dissimilar the stimuli for self-repair are from stimuli that can retrieve stored memories. In a simple connectionist network, like the Hopfield network (Hopfield, 1982) that will be discussed below, the difference between two stimuli can be measured with the Hamming distance.

In Chapter Two, we will discuss neuro-biological data of plasticity. The Hebbian learning mechanism is derived from this data. To illustrate repair with a plasticity mechanism, we will discuss repair with the Hebbian learning mechanism in a simple feed-forward network and extend it to a Hopfield network (Hopfield, 1982).

In neural network learning, one distinguishes between supervised learning and unsupervised learning. For a simple feed-forward network, supervised learning with Hebbian learning is as follows. An input stimulus is administered to the network by activating the nodes representing the input stimulus in the input layer. The output stimulus is administered by activating the nodes representing the output pattern in the output layer. This is shown in Figure 1.1. The two patterns are associated with each other by Hebbian learning. The Hebbian learning rule strengthens connections of active nodes and decreases the strength of connections of which one node is active and the other is inactive. In the Hopfield network, the input layer is the output layer and in case of auto-associative learning, the input stimulus is similar to the output pattern.

In the Hopfield network unsupervised learning is as follows. An input stimulus is administered to the network by activating the nodes representing the input stimulus in the input layer. The activation caused by the stimulus is propagated through the network to the output layer that settles after some time. The association between input stimulus and output pattern in which the network settles can then be learned in a similar way as described above with a Hebbian learning rule. When a Hopfield network is initially empty, i.e. the weights of connections are zero, the network will always settle into an output pattern with no activated neurons at all.

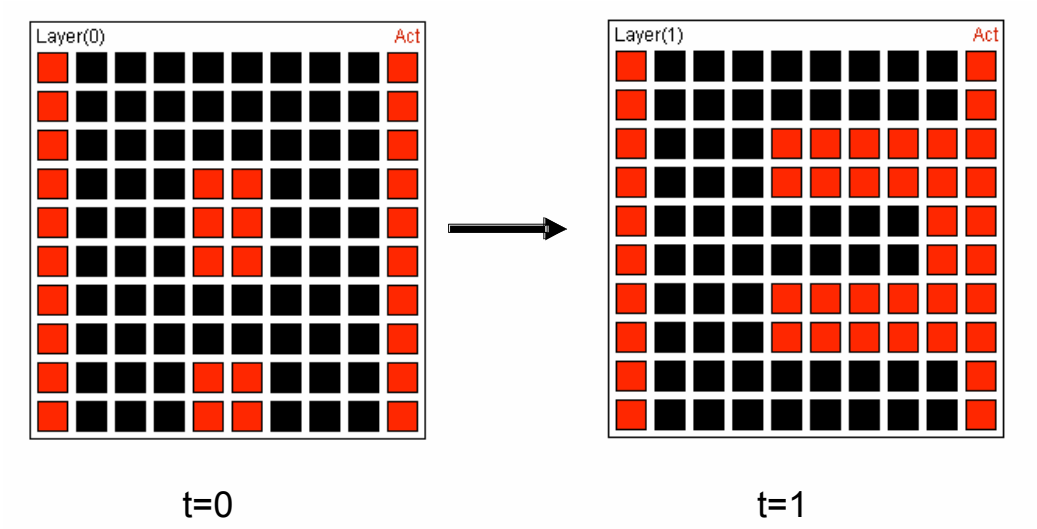


Figure 1.1. An example is shown where an input pattern, a letter A, is associated to an output pattern, which is a letter E. At time step *t* the input pattern A is presented to the input-layer. Then at time step *t+1* the output pattern E is activated in the output-layer.

One way to solve this problem is to initialize the connections with random values before learning such that an input pattern will be associated with some random output pattern. Another way to solve the problem is that after an input stimulus a random output pattern is constructed by activating nodes of the output layer randomly according to a given stochastic process. In this way, input stimuli will also be associated with a random output pattern.

In the Hopfield network repair by relearning is carried out by the unsupervised learning procedure. Given that there are learned or stored patterns and the weights are damaged, first the stimulus used for storing the pattern is administered to the network. Then the network iterates for a while to settle or converge to an output pattern. An example of pattern convergence is shown in Figure 1.2. If the weights are not critically damaged, the network will still be able to completely retrieve the correct

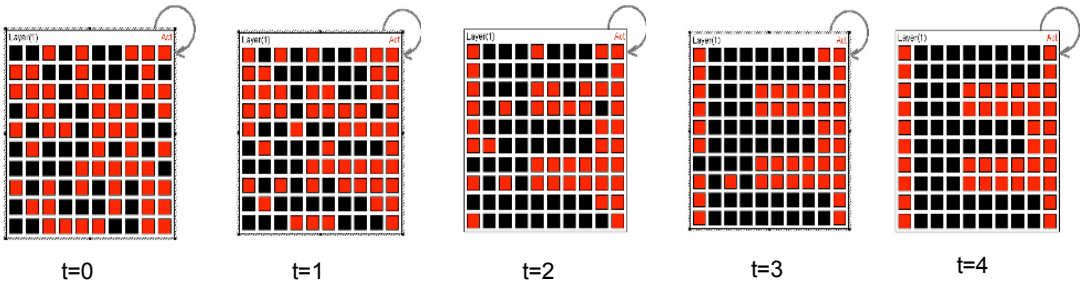


Figure 1.2. The figure shows how a random pattern converges to a given pattern. The pattern to which it converges is the letter E.

output pattern. The network will apply the Hebbian learning rule and the weights will be strengthened again.

Given a certain type of connectionist model, we can model the triage of recovery with the self-repair model as follows. Severe lesions are modeled by lesions that cannot be repaired by any self-repair scheme. The two most important parameters to model guided and autonomous recovery are the relative frequency of lesions compared with repair and the self-repair type. Relative frequency can be regulated by the number of times lesions or repair take place within one lesion-repair cycle (that could be zero). It is possible to simulate a recovery period with rehabilitation by alternating a self-repair type modeling guided recovery with a self-repair type modeling autonomous recovery and omitting lesions at all. The guided recovery uses stimuli that are better able to retrieve memory representations than the stimuli of autonomous recovery. Also different types of plasticity rules can be used to model the different types of self-repair. The exact way how a different type of recovery has to be modeled should depend on the available empirical data. The different types of recovery can be used to classify different types of self-repair in the brain: guided recovery is guided self-repair and autonomous recovery is autonomous self-repair.

1.2.2 Autonomous self-repair in connectionist models versus autonomous self-repair in the brain

In the context of connectionist models, we can also distinguish different types of self-repair on the basis of the stimulus used for repair. In case of guided or supervised self-repair, a stimulus strongly associated with a stored pattern is used. This can for instance be a stimulus used during the training phase or a prototype of the training stimuli. Since with this type of self-repair we can determine what stimulus to provide to the network and how many times to administer it, we have control over which memory representation will be repaired and the degree of repair. In case of autonomous self-repair, a randomly generated cue is used to select a stored memory representation. Since with this type of self-repair a stochastic process is determining what stimulus is chosen and how many times it is chosen, we have no control over which memory representation will be repaired and the amount of times it will be repaired. Autonomous self-repair, therefore, is more difficult to operate reliably than guided self-repair.

There may be differences between autonomous self-repair in the brain and autonomous self-repair in artificial neural networks as discussed above. Autonomous self-

repair in artificial neural networks is driven by randomly generated stimuli. In this case, any similarity between these stimuli and stimuli associated with stored patterns is a coincidence. Conversely, autonomous self-repair in the brain is probably triggered by stimuli that are more strongly associated with the stored patterns than are randomly generated patterns. That is because we assume that the brain is adapted to its environment. This implies that stimuli used for repair resemble the stimuli used for building the brain during evolution and during lifetime. Since autonomous self-repair of artificial neural networks is the most difficult form of self-repair to achieve in artificial neural networks, it represents a worst case scenario for autonomous self-repair in the brain. Self-repair in the brain probably resembles more the guided self-repair strategy of artificial neural networks. Thus, implementing autonomous self-repair in artificial neural networks will show that the most difficult type of self-repair is theoretically possible, making its existence in the brain more likely. Another reason to focus on autonomous self-repair in connectionist models is that if we can succeed modelling autonomous self-repair, it is likely that we can also succeed in modelling other types of self-repair. For these reasons in this thesis the stress is on autonomous self-repair. If we speak about autonomous self-repair, we mean self-repair with random stimuli. In the rest of the thesis, if we talk about autonomous self-repair in the brain that could be without random stimuli, we will explicitly mention this.

1.3 Chapter overview

1.3.1 Goals of the thesis

This thesis aims to show that self-repair is a possible process taking place in the brain. Another goal is to lay the foundation for models of brain recovery after damage. To achieve the first goal we will construct a self-repair model based on neurobiological and behavioral empirical data (Chapter Two). We, furthermore, will show that self-repair can work by first demonstrating an ‘easy’ type of self-repair in a simple connectionist model, which is consistent with the model of Chapter Two, and later will demonstrate autonomous self-repair in that same model (Chapter Three). To make it more likely that self-repair can take place in the brain, we will derive from the model of Chapter Three a connectionist model that is more neurobiological detailed in which we will also demonstrate autonomous self-repair (Chapter Five). Thus, the strategy to prove that self-repair can work in the brain is to take the most difficult type of self-repair, namely autonomous self-repair. Remember from Section 1.3 that by demonstrating it in an artificial model we show that autonomous self-repair in the brain is

also more likely possible. Since autonomous self-repair will be the main focus of this thesis, Chapter Four will be devoted to derive some general characteristics of autonomous self-repair in connectionist systems of different complexity with a mathematical model. The goal of Chapter Six attempts to relate autonomous self-repair in connectionist systems with autonomous self-repair in the brain with random stimuli. Thus, autonomous self-repair in the brain is here equivalent to autonomous self-repair in connectionist systems. With these chapters and the argument of Section 1.3 that if autonomous self-repair can be modeled we are also able to model other types of self-repair, we will also lay the foundation for a model of brain recovery that is rich enough to model different data of brain recovery and rehabilitation.

Why do we want to demonstrate self-repair in different connectionist models? In other words, why is making one model based on the empirical data not sufficient? The reason is that connectionist models have properties about which the experimental data of Chapter Two does not say much. All models of this thesis with which we will investigate self-repair are connectionist models. In general, this class of models is famous for their neural plausibility compared to other type of memory and cognitive models. This means that many properties of connectionist models are consistent with properties of the brain. If, however, we go into the details there are differences concerning similarity with the brain between the different types of connectionist models. For example, in the before discussed Hopfield there is no distinction between different types of neurons. The more neurobiological detailed models have more properties and parameters than the less detailed models. They, therefore, have a richer repertoire of neural behavior than the less detailed model. It, furthermore, is more or less assumed that the more biological detailed models are more similar to the brain. As we will also see in this thesis is that the more neurobiological detailed model allows implementing aspects of self-repair more realistically (Chapter Seven).

Besides the question whether it is still a model, it is, however, practical impossible to model all details of the brain, because of the enormous computational cost. In general, the common practice is that a model is a result of a tradeoff between the particular research question and computational cost. This thesis is no exception to the common practice. The connectionist models of this thesis have a simulation run time such that it is possible to search for parameter values and replicate results.

1.3.2 Detailed chapter overview

In Chapter Two, we will address the question whether memory representations in the brain are endowed with a self-repair capacity. In order to answer this question, we will present a

theory of maintenance of redundancy and demonstrate how it can extend the lifetime of memory representations, enormously. We will review neurobiological data of redundancy and plasticity. With the data, we will argue that the brain possesses redundancy on different levels of the brain (synaptic, cellular, and neural circuit level) and maintenance is carried out by plasticity mechanisms in the brain. We will, furthermore, review behavioral data of the use-it-or-lose-it principle and the serial lesion effect supporting redundancy, maintenance, or maintenance of redundancy. Based on the data we will build a model of self-repair. The model's procedure of lesions followed by repair is similar to the serial lesion effect. The intensity of self-repair is dependent on external stimuli similar to the 'use-it-or-lose-it' principle. Redundancy is present as some form of abundance of connections, while plasticity mechanisms, derived from the plasticity data (Section 2.3.2) maintains redundancy at some minimal, safe level. Moreover, in a connectionist model implementing the self-repair model, we will demonstrate that self-repair is able to extend memory lifetime. Finally, we will show how the ideas of the self-repair theory can be applied to normal aging.

Chapter Three is a first exploration of self-repair with mathematical and connectionist models. We will investigate redundancy, which in neural networks is present in the connections, with random graph theory. Then we will address the question whether self-repair is possible in connectionist models at all and what types of self-repair are possible. Concretely, the latter two research questions imply that we will study guided and autonomous self-repair in the classical connectionist Hopfield model (Hopfield, 1982). In a soft k-winner-take-all network with stochastic neurons, we will also demonstrate self-repair and explore it further. Finally, we will discuss the necessity of self-repair in the brain and why we chose to model self-repair by changes in connectivity.

The main topic of Chapter Four is to investigate random cued self-repair or autonomous self-repair with an analytical model. The model allows us to express memory retrieval in terms of probability. The aim is to derive results that can be applied to the more complex simulation models of the other chapters and the brain. The results will come from the following research questions: What will be the effects on system stability (1) of weight differences due to learning alone, (2) of weight differences because of learning and lesions together, (3) the activation probability, and (4) pattern size? System stability is expressed in the retrieval probabilities of the weakest and strongest memory representation. The first retrieval probability indicates the risk of a system of losing a memory representation. The second retrieval probability gives information about a possible runaway memory representation. Using the probabilities of the weakest and strongest memory representations as

a measure, we will show under which conditions the most difficult type of self-repair, autonomous self-repair, is possible. Since the brain fulfils the important condition of comprising many patterns, we will argue that autonomous self-repair is feasible in the brain.

In Chapter Five, we will study autonomous self-repair in a more neurobiological plausible network. Questions addressed in this chapter are whether autonomous self-repair is possible in such a network and how to achieve it? The main difference with the models of the previous chapters is that this model possesses spatio-temporal properties that will allow us to relate data generated by the model to experimental data. The model is for instance more neurobiological detailed in network structure and neuron activation rule. It might represent for instance a small part of the primary somato-sensory cortex (SI), in which each neural assembly represents a finger of one hand.

In chapter Six, we will investigate whether self-repair by maintenance of redundancy takes place during sleep. In particular we will address the question whether autonomous self-repair with random cueing takes place during sleep. An indication of random brain activation is the unstructured order of dreams. Furthermore, internal random activation would probably also not be desirable during daytime, because it would interfere with the external stimuli. We will investigate the research question of self-repair during sleep by demonstrating in a neurobiological plausible model of sleep that self-repair works and extends the lifetime of a memory. The investigation will, furthermore, consists of a review of models and data of processes of memory maintenance and consolidation taking place during sleep. This review will show that processes during sleep may be able to carry out self-repair. It will, furthermore, explain that theories and models of these processes imply that the brain possesses redundancy as is proposed by self-repair. It will, finally, show that the processes of memory maintenance and memory consolidation are similar to the algorithmic procedure of self-repair. In particular, many theories of these sleep processes use random cueing as a way of activation.

In the last chapter, we will discuss the main results of this thesis. Among others, we will discuss to what extent we have proven the hypothesis that self-repair by maintenance through redundancy takes place in the brain. In the context of the hypothesis, we will discuss why there is a need for self-repair and present an experiment to test the hypothesis. We will, furthermore, explain how models of this thesis can easily be applied to modeling recovery from brain damage. The type of damage of these models of recovery will not be small and diffuse as will be investigated in this thesis, but large and localized. In addition to model recovery from brain damage, we will discuss other future research with the models of this thesis.

2

The self-repairing brain: a synthesis

Abstract

In this chapter, we will argue that memory representations in the brain are endowed with a self-repair capacity. We will present a theory of maintenance of redundancy and demonstrate how it can extend the lifetime of memory representations. We will review neurobiological data of redundancy and plasticity. We will argue that the brain possesses redundancy on different levels of the brain (synaptic, cellular, and neural circuit level) and argue that maintenance may be carried out by plasticity mechanisms in the brain. We will, furthermore, review behavioral data of the use-it-or-lose-it principle and the serial lesion effect supporting redundancy, maintenance, or maintenance of redundancy. Based on the data we will build a model of self-repair in which we demonstrate that self-repair is able to extend memory lifetime. Finally, we will show how the ideas of the self-repair model can be applied to normal aging.

2.1 Introduction

In this chapter we will address the research question whether memory representations in the brain have a self-repair capacity by maintenance of redundancy. We address this question in the following way: (1) we will show the advantage of having a self-repair process in a thought experiment, (2) we will review neurobiological data that support self-repair, (3) we will review behavioral data that support the self-repair hypothesis, and, (4) derived from the biological data we will build a connectionist model in which we demonstrate self-repair.

All types of self-repair are necessarily based on redundancy of information in a system: without redundancy, loss of information would be irreversible. By keeping redundancy at or above some minimal level, the chance of irreversible information loss due to damage is kept very low. In a thought experiment, we will show that self-repair may cause an enormous increase in memory lifetime compared with unrepaired neural circuits. This thought experiment, furthermore, illustrates our approach to self-repair by the maintenance of redundancy (Section 2.2). After having established that self-repair in memory systems can have a definite advantage, we ask whether self-repair does indeed take place in the brain. We will address this question by reviewing neurobiological evidence for brain redundancy and review plasticity processes of the brain that may carry out maintenance (Section 2.3). We will, furthermore, review behavioral data that support the self-repair ideas of redundancy and maintenance of redundancy (Section 2.4). From the neurobiological data we derive a particular connectionist model of self-repair. In this model, we will demonstrate self-repair. We conclude this chapter with a discussion how the connectionist model of self-repair can be used to model normal aging.

2.2 The advantage of Self-repair

A thought experiment illustrates the limitations of redundancy and the great gains of self-repair. Let us suppose that a security agent has to guard an important document. Suppose furthermore, that the chance that the document is destroyed by the enemy is 0.5 in a month. The expected lifetime of the document, or in other words the time before the document is destroyed, would then be 1/0.5 or 2 months.

Let us introduce redundancy in our thought experiment. The security agent copies his document 9 times and gives it to 9 other security agents. Instead of having the information once, in which case there would be no redundancy, we now have a ten-fold redundancy. If we still suppose that the chance of losing a copy is 0.5 in a month, it can easily be demonstrated

that the expected lifetime of the document, or in other words the time until the last of all copies is lost, is approximately 4 months. The ten-fold increase in redundancy increases the expected lifetime by a factor 2.

Now, let us introduce self-repair, a process that maintains redundancy. The ten-fold redundancy is still present because we again have ten agents who each possess one document. Repair is modelled by a monthly meeting of the ten agents. During this meeting, an agent who still has his document makes copies of it and gives one to each agent who has lost his document. At the end of the month, each agent has once again one document. This repair process continues until one month, by chance, all copies are found to be lost. With a 50% loss rate per copy, the monthly survival probability of the document's contents is $1 - 0.5^{10} \cong 0.999$ and its expected lifetime is 85 years (1023 months). Without the monthly 'repair' session, we have seen that the expected lifetime is only 4 months. Consequently a ten-fold redundancy gives only a small increase in expected lifetime, while a repair process increases the lifetime more than five hundred times.

This example shows that self-repair by maintenance of redundancy can increase the lifetime expectancy of memories enormously. In the thought experiment, memory is represented by a document and redundancy is present in the form of simple copies of this document. Memory in the brain is most likely to be organized as a connectionist system rather than a system where redundancy is present in the form of exact copies (we will address this topic the next section). In Section 2.5, we will further explain how memory may be organized in a network or connectionist system and where redundancy in such a system resides. We will explain how it is redundant and how self-repair can be implemented. We will, also, illustrate self-repair in a specific connectionist system: the Hopfield network.

2.3 Neurobiological evidence of self-repair

In this section, we will review some neurobiological evidence for self-repair. More specifically, we will address questions concerning the nature of memory redundancy and the maintenance of this redundancy in the neural brain. What is it that provides memory encoded in neurons and synapses with the property of redundancy (Section 2.3.1)? Furthermore, what possible mechanism carries out the maintenance or repair in the brain (Section 2.3.2)?

2.3.1 Neurobiological evidence of memory redundancy

Redundancy of memory in the brain is probably not present in the form of exact copies. We rather expect it to be encoded in the myriad of neurons and their connections. In this section,

we will give three different examples of redundancy found at separate levels of the brain: i.e. at the synaptic, the cellular, and the neural systems level.

Redundancy at the synaptic level may be present in the form of spare or silent synapses. Behavioural evidence of spare synapses becoming operational after damage is derived from lesion studies showing a rapid reorganization in the receptive field of cells after a lesion. These changes are so rapid that it is unlikely they are caused by structural growth like synaptogenesis (S. R. Butler, 1988). It therefore is suggested that these rapid changes are brought about by the unmasking of normally inhibited synapses or by changes in the efficiency of synapses. Evidence supporting the idea of the release of inhibition has been found by Jacobs and Donoghue (Jacobs & Donoghue, 1991). Evidence supporting the idea of efficiency changes has been found in different parts of the developmental and adult mammalian brain by Liao et al. (Liao *et al.*, 1999) and Reid et al. (Reid *et al.*, 2004). The efficiency changed synapses that lack glutamate AMPA-receptors. Without these the cell is more difficult to activate, because glutamate is one of the most important excitatory neurotransmitters in the brain. They are 'silent' in the sense that they are not as active as normal neurons with AMPA-receptors. They can be turned into normal functioning cells by long-term potentiation (LTP), which inserts AMPA-receptors. LTP is the most likely neurobiological correlate for learning as we will explain in the next section. It should be noted that we know no experiments showing that 'AMPA' silent synapses take care of the rapid reorganization after brain damage. It, does, however support our idea that repair can be carried out by normal learning mechanism such as LTP.

Redundancy is present at the cellular level in the form of neurogenesis. For reasons unknown, the brain does not generate new neurons on a large scale. One of the reasons may be the risk of memory instability as for example is shown by a study of Parent et al. (Parent *et al.*, 1997). At a small scale, however, there is evidence of newly generated neurons during the entire lifetime. Research has shown that neurogenesis takes place in the adult brain in the olfactory bulb (Altman, 1969; Luskin, 1993) and dentate gyrus (Altman & Das, 1965; Eriksson *et al.*, 1998). For the dentate gyrus it is known that neurons are produced at a constant rate and that some of these newly generated neurons survive while others die. The number of survivors increases after a learning task (Gould *et al.*, 1999) or after damage to the hippocampus (Liu *et al.*, 1998). The latter suggests that newly generated neurons are involved in restitution after compensation for damage. Though redundancy in the form of spare neurons takes place at a small scale, it can have a functionally large impact given the pivotal role of the hippocampus (dentate gyrus) in learning and memory.

Two different types of redundancy are to be found at the neural systems level. A first type of redundancy is in the form of a backup system or spare parts. A second type of redundancy is the network redundancy that may be present in neural circuits as will be explained in the Section 2.5. A clear example of the first type of redundancy is the control of the respiratory system (Kavanau, 1997). If the normal neural circuit of the respiratory system fails, another (spare) circuit immediately takes over. By a neural circuit we mean a network of synaptically connected neurons or set of neurons that are functionally connected, in such way that they reliable become activated when a particular memory is retrieved (I. H. Robertson & Murre, 1999). Redundancy at the neural system level is probably more frequently due to network redundancy in neural circuits. A function or memory trace in neural circuits may be distributed over many parts of the brain (Braitenberg & Schüz, 1991). Though neuroscience research often tends to connect a particular function to a particular brain region, there are several arguments against strict localization of functions. For one thing, even if functions are localized in one region it can be the case that the functions are distributed across that particular region. Such population coding has been found in the motor cortex (MI) (Sanes & Donoghue, 2000). Secondly, the meaning or function of a region or of single neuron is determined by its connecting parts (Friston, 2002). Function in this case has the meaning of a cortical area specialized in some aspects of visual or motor processing. This function can only exert its meaning in concert with its connected parts. These particular cells or regions are part of a memory trace and can only be activated if that memory trace is activated. For example, the motor cortex is activated along with regions of language and vision when one is writing a sentence. Cells and regions are not part of a single memory trace, but of many memory traces: a motor area involved in writing can also be involved in playing the piano. Memory of the brain and in particular the cortex seems to be distributed over the cortex with cortical regions involved in several memory traces (Haxby *et al.*, 2001). Friston (Friston, 2002) has called this property of memory functional integration. Others have noted this property of memory before and gave it different names as multiplexing (Bach-y-Rita, 1990) and superimposedness (Crick & Mitchison, 1983).

If a particular cortical area happens to be part of one or several memory traces, information also resides in other cortical areas connected to it. For example, a connected cortical area gives a particular activity pattern as input. In case of damage this input can be used as information to rebuild the cortical area. Suggestive evidence for the rebuilding of one cortical area through another cortical area is an experiment by Ramirez et al. (Ramirez *et al.*, 1999). Their experiment shows that previous injured behaviour, caused by a lesion to a lateral

hippocampal area, is reinstated by newly sprouted synapses in the damaged hippocampal area with connections of the contra-lateral homologue of the damaged hippocampal area.

We addressed redundancy in the brain at the synaptic, neuronal, and neural systems level. Redundancy in the brain means that information is present to reinstate or substitute the damaged parts. This is information for (re)building damaged parts, but also the correct architecture. This is the reason why replacing damaged cells by stem cells is a only one part of the solution (Kempermann & Gage, 1999; Magavi *et al.*, 2000; Weiss, 1999). Surrounding cells have to give the right signals to make the correct connections with the new cells. Under normal circumstances information in neural circuits may guide repair processes at the lower synaptic and cellular level. For example the circuit level may provide information how many silent synapses must become active or determine the number of newly generated neurons in the dentate gyrus. Finding out how these processes are regulated can give us insight into neural repair after damage.

2.3.2 Neurobiological evidence of memory maintenance

It is known that after severe brain damage structural changes within the damaged and connected areas may compensate for the damage (Bach-y-Rita, 1990; Jones, 2000; I. H. Robertson & Murre, 1999). In this section, we will present neurobiological data of brain plasticity that may take place in the normal adult brain.

A first type of brain plasticity is long-term potentiation (LTP). This is the enhancement of synaptic responses resulting from brief, repetitive activation of an excitatory monosynaptic pathway by high frequency responses of electrical pulses (Bliss & Lomo, 1973). LTP can change the cell's morphology (a neural circuit) in the following ways. It can lead to synaptogenesis in different ways: by changing the number of synapses (Toni *et al.*, 1999), by remodelling existing synapses (Geinisman, 2000), or by activating silent synapses (Isaac & Mayes, 1999; Voronin & Cherubini, 2004). Similar to LTP, long-term depotentiation (LTD) is the most likely neurobiological candidate for the weakening of memory traces (Artola & Singer, 1993; Dudek & Bear, 1992).

The exact role of LTP and LTD in learning is still uncertain. In a review about the role of LTP in learning, Martin et al. (Martin *et al.*, 2000) give 4 criteria that LTP has to meet to be necessary and sufficient for learning, namely: detectability, mimicry, anterograde alteration, and retrograde alteration. Detectability means that if an animal displays memory of some previous experience, a change in synaptic efficacy should be detectable in the nervous system. Mimicry means that if a same spatial synaptically weight pattern is artificially

induced, the animal should display 'apparent' memory of some not occurred 'past' experience. Anterograde alteration means that the prevention of a weight change during a learning experience should impair the animal's memory of that experience. Retrograde alteration means that interventions altering the spatial distribution of synaptic weight induced by prior learning experience should alter the animal's memory of that experience. Most studies are about the detectability of which the study of Rioult-Pedotti et al. (Rioult-Pedotti *et al.*, 2000) is a clear example. They show that behavioral forelimb motor skill learning strengthens the horizontal connections of MI opposite to the trained forelimb. After saturating the connections they artificially induce LTD in the connections. Unfortunately, they did not try to bring the weights of the connections to baseline and test whether the animals were unable to carry out the task (retrograde alteration). One of the conclusions of Martin et al. (Martin et al., 2000) is that until now there is a wealth of evidence supporting the necessity of LTP in learning, but not its sufficiency, because there are no experiments carried out in the hippocampus, amygdala, or the cortex that meet all 4 criteria. LTP thus plays a critical role in learning, but it may not be the complete story.

Suggestive experimental evidence that Hebbian learning is involved in brain repair is found in similar morphological and chemical changes observed after LTP and brain damage. An example of a similar morphological change are the spine modifications after LTP (Star *et al.*, 2002; Trachtenberg *et al.*, 2002) and damage (Kolb & Gibb, 1993; Villablanca *et al.*, 1998). An example of similar chemical change is the brain-derived neurotrophic factor (BDNF) found after LTP (Kovalchuk *et al.*, 2002; McAllister *et al.*, 1999) and repair after damage (Ikeda *et al.*, 2003; Kleim *et al.*, 2003; Tropea *et al.*, 2003).

A second type of brain plasticity that may be involved in brain repair is homeostasis. It is plasticity involved in the formation, maintenance, and proper functioning of neural circuits. It keeps cells, in particular central neurons, within working range for effective information transfer in a changing input environment. In other words, it addresses problems and questions of how cortical neurons can be prevented from falling silent or from saturation if the average input falls too low or rises too high. There is empirical evidence of homeostasis forms of synaptic plasticity (Turrigiano *et al.*, 1998; Turrigiano & Nelson, 2000, 2004) and a homeostasis form of neuron intrinsic excitability (Desai *et al.*, 1999; W. Zhang & Linden, 2003). Homeostasis mechanisms of synaptic plasticity adjust all of a neuron's synaptic weights up or down. This synaptic scaling can be due to pre- and postsynaptic changes. For example, total synaptic strength can be regulated by presynaptic transmitter release or postsynaptic receptor clustering. Homeostasis of intrinsic neuron plasticity can adjust the

intrinsic excitability of the whole neuron by changing the axo-somatic ion channels. It can also adjust the excitability of a specific dendritic module by a local dendritic change in voltage-gated channels (W. Zhang & Linden, 2003). Intrinsic neuron plasticity can be induced by similar stimulus patterns as LTP and it is often co-expressed with LTP or LTD.

Until now there exists no experimental proof of the involvement of homeostasis mechanisms in repair of damage. One reason is that research focusses on changes in the normal functioning brain. Researchers in the field of homeostasis plasticity did, however, make a relevant remark on the relation between homeostasis mechanisms and brain repair. Turrigiano and Nelson (Turrigiano & Nelson, 2004) suggested that the type of mechanism depends on the type of perturbation. It seems illogical to assume that mechanisms set in motion by perturbations due to learning would not be activated by perturbations due to damage. Although different mechanisms may be used for different perturbations, this does not refute the notion that damage up to some limit can be repaired by the brain.

In Section 2.5.2 we will explain that LTP and LTD form the so-called Hebbian learning mechanism and that homeostasis plasticity represents homeostasis plasticity mechanisms. We will, furthermore, argue how these mechanisms carry out self-repair at the neural circuit level in a connectionist network.

2.4 Behavioral correlates of self-repair

In this section we will address behavioral evidence of self-repair. We will discuss behavioral data clustered around two concepts or theories that are used to explain the behavioral data, that is the use-it-or-lose-it principle and the serial lesion effect. We will review some experiments that support the self-repair ideas of redundancy and maintenance of redundancy. We will moreover argue that the use-it-or-lose-it principle represents maintenance and the 'use-it-or-lose-it' theory represents maintenance of redundancy. We will furthermore review data that connect neural structures and processes with behavioural data. We will end each section with concluding remarks about each concept and self-repair.

A caveat is that the behavioral data can only provide support for the self-repair theory. It can never give a definite proof for the self-repair theory. The brain comprises billions of neurons and a number of synapses that is a multiple of the number of neurons. To couple behavior to the possible temporal spatial patterns of neurons and synapses is a very difficult task. We have discussed in Section 4.1, that although the cortex can be divided into more or less functionally distinct regions (Kandel *et al.*, 1991), the strict localization of a complete memory function in a particular region is very likely to be an overstatement. All functional

regions discovered by neuroscience participate in many memory traces. Even if we identify the functions of all these regions, the great challenge remains to find out how the functional integration of the regions operates (Friston, 2002). Furthermore, we still lack the tools to measure the spatial temporal patterns of the brain precisely. As a consequence, one cannot determine the exact relationship between neural and synaptic activity on the one hand and behavior on the other hand. Due to the lack of knowledge of this relationship, one can only establish a correlation between neural activity and behavioral data. Therefore, behavioral data can only support a neural theory and cannot decide between subtlety different (competing) neural theories.

2.4.1 The use-it-or-lose-it principle

The use-it-or-lose-it principle states that brain use extends its lifetime. It is assumed that loss will occur automatically due to deterioration processes like aging, injury (including self-inflicted injuries like alcohol intake), and disease. A definition for behavioral data is given by Salthouse et al. (Salthouse *et al.*, 2002), which reads as follows: use-it-or-lose-it is used to explain behavioral data showing that age and injury-related effects on measures of cognitive performance can be moderated by the individuals' lifestyle, and particularly by the amount of cognitive stimulation individuals receive in their daily lives.

Longitudinal behavioral data of humans show that intellectual leisure activities (attending lectures, classes, and playing desk games of skill), social leisure (attendance of dances, sport events, and visiting a public house or pub), and physical activities (aerobics, walking, jogging, and gardening) correlate with slower cognitive decline in healthy elderly (Elwood *et al.*, 1999; Gold *et al.*, 1995; Scarmeas & Stern, 2003; R. S. Wilson *et al.*, 2003). With regard to the effect of cognitive stimulation after injury, it is known for recovery after alcohol use that it is influenced by environmental factors such as physical exercise and cognitive stimulation (Goldman, 1990). Recovery seems to be accelerated if newly sober subjects are asked to "use their heads" at a level that is equal or slightly beyond, their current level of functioning.

The data above suggest an influence of cognitive stimulating activities on cognitive functioning. In 1999 Hultsch et al. (Hultsch *et al.*, 1999) found that the 'use it or lose it' data can also be explained by the opposite causal relation, namely an influence of cognitive functioning on activities. They also showed this opposite causal relationship between activities and cognition for the data of Gold et al. (Gold et al., 1995). Since that moment there has been a discussion about the causal direction of activity and cognitive functioning.

Differences of interpretation of results were attributed to study sample, selection of indicators of study sample, and methodological issues. We will, therefore, discuss recent research to counteract the idea of an uniquely one-way causal relationship of cognitive functioning on activities. We should bear in mind, however, that it is hard to give definite conclusions based on longitudinal studies only. We will say more about this at the end of section.

The LASA-study of Aartsen et al. (Aartsen *et al.*, 2002) is an example of a study that stated to have found an only one-way causal relationship of cognitive functioning on activities. This one-way relationship was, however, only found if corrected for an unmeasured variable of social economic status. When no correction was performed for this variable and was not left out of the analysis, the causal relationship went both ways. A weak point is that the variable of social economic status cannot be left out of the analysis, because it is represented among others by sports, which as we will explain later in this section, is an important cognitive stimulating activity for the use-it-or-lose-it effect. Moreover, a later study of Bosma et al. (Bosma *et al.*, 2002) corrected for social economic status and still found the two-way relationship between cognitive stimulating activities and cognitive functioning. They, furthermore, found that mental-workload, and thus cognitive stimulating activity, protects against cognitive impairment (Bosma *et al.*, 2003). These data thus suggest a reciprocal interaction between activities and cognitive performance in healthy elderly. Simply stated, for a healthy brain it is easier to stay healthy.

What is the effect of use on the brain? Human data exist that indicate that brain use has an effect on brain anatomy. Research on taxi-drivers (Maguire *et al.*, 2000); musicians (Münte *et al.*, 2002; Schneider *et al.*, 2002) (Bengtsson *et al.*, 2005), Braille-readers (Hamilton & Pascual-Leone, 1998; Sterr, 1998), and jugglers (Draganski *et al.*, 2004) show an effect of brain use on the adult's brain anatomy. Some of this research suggests a relationship between anatomy and use on a time scale of years. For example, the taxi-driver experiment shows that taxi-drivers have an enlarged hippocampus. This enlargement correlates with the amount of time spent as a taxi-driver. It is, however, suggested that it is brought about in a period of two years, because this is the amount of time taxi-driver training usually requires. Whereas the taxi-driver experiment suggests a relation between use and brain anatomy over a time scale of years, the experiment with jugglers demonstrates a change over a time span of mere months. In this experiment novel juggle learners showed an increase of the visual area for movement detection (V5) in only two months. All these experiments demonstrate that the brain can change even in adulthood.

In addition to the human data, there are animal data describing the effect of brain use on the brain's anatomical structure. The research in the previous section with rats in enriched environments demonstrates the effect of use and of specific stimuli on brain anatomy, e.g. an effect on dendritic branching and synapse density (Greenough *et al.*, 2002; Kolb, 1999). Other environment rich studies of Saito et al. (Saito *et al.*, 1994) and Nakamura et al. (Nakamura *et al.*, 1999) show that an enriched environment restores the decrease of synaptophysin contents that takes place due to normal aging. Rat studies, furthermore, show that the decrease of neurogenesis in the aging rat (Kuhn *et al.*, 1996) can be counteracted by environmental enrichment (Kempermann *et al.*, 1998). These last two type of experiments show not only an effect of brain use on neural processes and structure, but also indicate that neural markers of aging can be counteracted by environmental enrichment.

Longitudinal research in general and specifically towards use-it-or-lose is very complex; it has many variables. The debate about the causal relationship between cognitive stimulating activities and cognitive functioning illustrates this. Because there are so many variables in which experiments can differ like the determination of the amount of cognitive stimulation, the validity of self-reports, and the test-group sample, it is hard to draw definite conclusions. Only a perfect experiment can resolve this. In the perfect experiment, one could randomly assign individuals to different cognitive levels and assess their complete cognitive abilities over time (Salthouse et al., 2002). This perfect experiment is impossible because of fundamental, practical, and ethical reasons. Definite conclusions are therefore unattainable. Fortunately, there are other data available that can shed some light on the issue. There is the neuro-biological data described in the paragraph above.

The neuro-biological data of use-it-or-lose-it show the effect of brain use on quantity of connectivity and structure. From the point-of-view of the neurobiological data, plasticity mediates the effect of use. This explains why not only cognitive activities like playing chess have a positive effect on mental health, but also why the less cognitive activities like physical excercise (Colcombe S, 2003; Colcombe *et al.*, 2004; R. D. Hill, 1993) have an effect on plasticity (Cotman & Berchtold, 2002). Plasticity may thus be the mechanism mediating use by restoring connectivity. Although mental health is not equivalent to connectivity, it is implicitly assumed that mental health correlates with a certain degree of connectivity. For example, the shrinking of dendritic arbors in the adult brain is regarded as a sign of aging (Uylings *et al.*, 1999) and injury (Jones, 2000; I. H. Robertson & Murre, 1999). Hereby we provide plasticity as a mechanism for the use-it-or-lose-it effect, the absence of which was

given as a reason to distrust the use-it-or-lose-it theory by researchers in the longitudinal research field (Salthouse et al., 2002).

Summarizing we conclude that, given the behavioural use-it-or-lose-it dataand the neuro-biological data of the effect of use on the brain, there is sufficient evidence for the use-it-or-lose-it concept and for a maintenance process driven by cognitive stimulation or use.

2.5.2 The serial lesion effect

The serial lesion effect is the finding that a series of small lesions with intermittent recovery periods will result in better final performance compared with a single large lesion, the size of which equals the cumulative size of the small lesions. Behavioral data supporting the serial lesion effect are provided by human and animal studies. In humans, the serial lesion effect has been observed in cancer patients for well over a century. Already in 1836, the French physician Dax observed that sudden damage to the left hemisphere was far more likely to produce aphasic symptoms than a slowly developing tumor. In animal experiments, the serial lesion effect has been demonstrated in various areas of the brain of both the rat (somatosensory, hippocampus, reticular formation, frontal cortex or amygdala) and the monkey (parts of visual cortex, somatosensory cortex: Brodmann 4 and 6) (Finger & Stein, 1982). The controlled animal experiments show that the serial lesion effect is influenced by many variables: i.e. age (Corwin *et al.*, 1981), inter-operative time (Gavin & Isaac, 1986), the level of sensory stimulation to which the subject is exposed during the lesion intervals including training (Corwin et al., 1981; Finger & Stein, 1982; Scheff *et al.*, 1977), the type of area damaged (Finger & Stein, 1982), and the order in which the areas are damaged (Curtis & Nonneman, 1977; Finger & Stein, 1982).

What is the relation between the serial lesion effect and the brain? A first hypothesis is that the time between the lesions is used for repair and in particular for the repair of brain connectivity. This is supported by data showing an effect of sensory stimulation and training. The control of sensory stimulation is similar to the control of the enrichment of an environment. As was reviewed above, this can have an effect on brain connectivity. Training can be regarded as a specialized or guided form of enrichment, which has been shown to have an effect on brain connectivity (Biernaski & Corbett, 2001; Nudo *et al.*, 1996; Withers & Greenough, 1989). A compelling experiment that connects the serial lesion effect to connectivity is by Ramirez et al. (Ramirez et al., 1999). In this experiment, they demonstrate that uni-lateral progressive lesions in the entorhinal cortex are accompanied by (axonal)

sprouting in the crossed tempero dentate pathway. The behavioral significance of this sprouting is underlined by two findings. The first finding is that the transection of the crossed tempero-dentate pathway reinstates the behavioral deficit. The second finding is that enhanced synaptic efficacy after sprouting precedes behavioral recovery.

A second hypothesis regarding the relationship between the serial lesion effect and the brain is the reduced deficit explanation. The reduced deficit explanation postulates that the first lesion initiates processes that reduce the impact of subsequent lesions. In this case the inter-operative time is not used for repair but for 'reducing' processes. The above-mentioned experiment of Ramirez et al. (Ramirez et al., 1999) is consistent with the reduced deficit explanation. They found no behavioural recovery after the first lesion, but recovery did occur after the second lesion. Moreover, behavior was not reinstated if there was no second lesion after the first (prime) lesion. On the other hand, an experiment by de Castro and Zrull (Castro de & Zrull, 1988), presents evidence favoring the serial recovery hypothesis. In an experiment where they first made a lesion in one hemisphere and a second lesion in the homologue area of the opposite hemisphere, they show that the secondary lesion produces the same contra-lateral (behavioral) deficits as compared to the first lesion. In case of the reduced deficit hypothesis the second lesion would have resulted in less deficit. Their experiment, furthermore, found that the deficits of the first lesion recovered, contrary to, the experiment of Ramirez et al. (Ramirez et al., 1999).

Several explanations can be offered for the varying results of Ramirez et al. (Ramirez et al., 1999) and de Castro and Zrull (Castro de & Zrull, 1988). A first explanation is that the serial lesions of Ramirez et al. involved uni-lateral lesions, while the serial lesions of de Castro and Zrull (Castro de & Zrull, 1988) involved bi-lateral lesions. Thus, the variable of a different brain area may explain the different results. A second explanation is that both the reduced deficit hypotheses and serial recovery hypotheses can be true and are not mutually exclusive as is assumed by de Castro and Zrull (Castro de & Zrull, 1988). The processes suggested in the two hypotheses may take place simultaneously.

An indication for this is that although after the first (prime) lesion of the Ramirez et al. experiment (Ramirez et al., 1999) no behavioural recovery occurs, axonal sprouting is observed. So, there may be some repair following the first lesion, but not sufficient for behavioural recovery. After the second lesion there is increased axonal sprouting compared to the sprouting response after the first lesion. This boosted reorganization after the second lesion may be a mixture of a process reducing the impact of subsequent lesions initiated after the first lesion and repair after the second lesion.

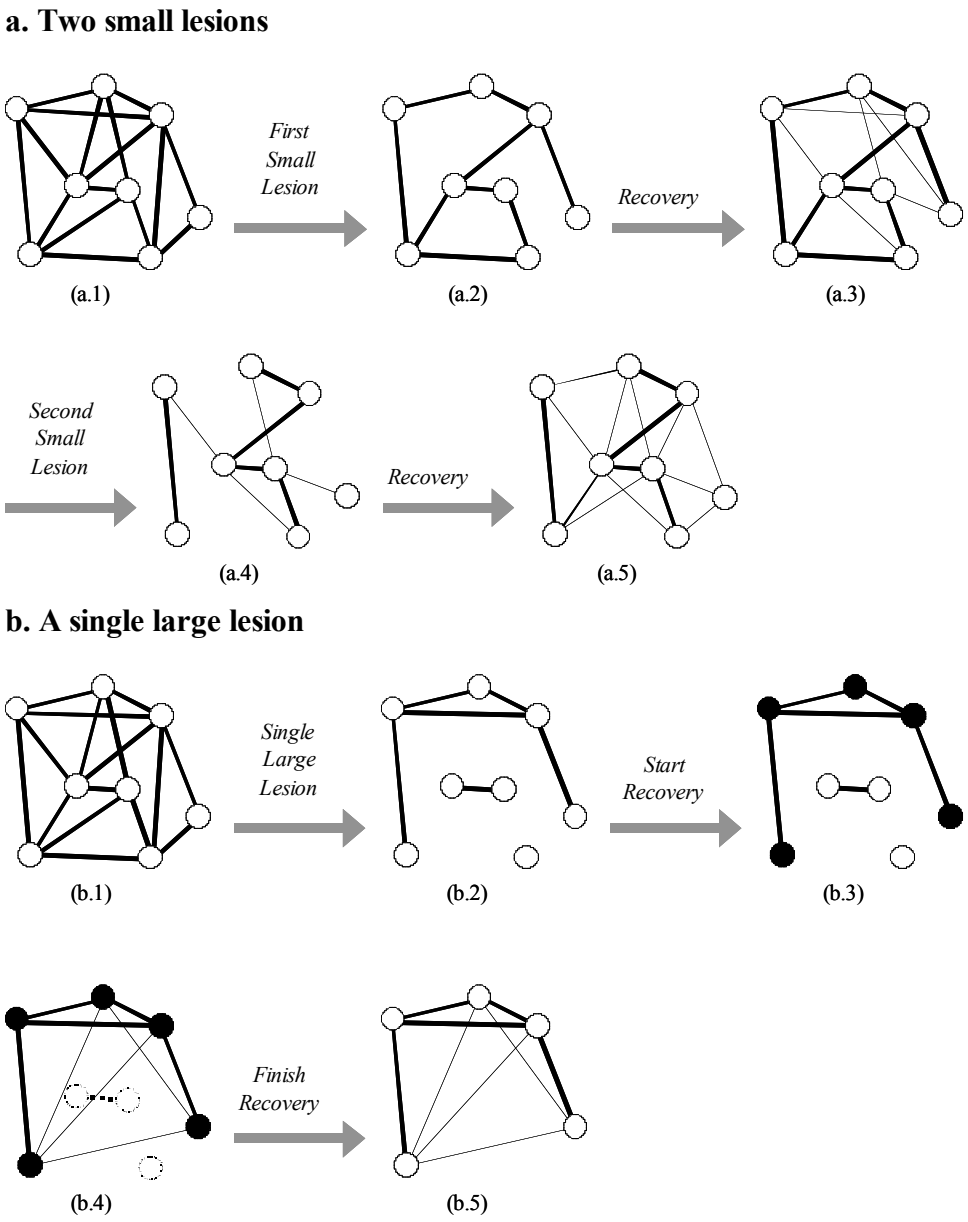


Figure 2.1. The serial lesion effect illustrates the difference between a multiple staged small lesions and a single large lesion of equivalent size. (3a) Depicts the effect of the two stage small lesion. (a.1) A well connected, intact neural circuit. (a.2) After the first diffuse lesion the same circuit is still connected but less densely. (a.3) Self-repair takes place. Because the neural circuit was still well-connected enough after the first lesion, or in other words contained enough redundancy, all nodes receive a weight update. (a.4) The renewed redundancy is enough to compensate for the second lesion after which again every node receives a weight update. (a.5) The result is a neural circuit where every node is directly or indirectly connected to any other node of the neural circuit. (3b) Depicts the effect of a one stage lesion that is equivalent in size to the two stage small lesion of Figure 2.3a. (b.1) The same intact neural circuit as in 3.a.1 (b.2) receives a large lesion. (b.3 and b.4) Self-repair takes place, but the neural circuit does not contain enough redundancy, to get completely re-connected. (b.5) The result is a degenerated neural circuit.

Additional experiments have to be carried out to elucidate the effect found by Ramirez et al. (Ramirez et al., 1999). For example, what if there are three progressive lesions instead of two? Whereas the original experiment does not unequivocally support the self-repair theory in its current form, the enhanced sprouting after the second and third lesion would constitute support for the self-repair theory.

The above data on the serial lesion effect suggest that the brain is engaged in repair after damage. As we argued, the brain possesses redundancy. If the size of damage is such that a given memory representation can only be partially retrieved, because too much information has been lost, it can also only be partially repaired by a plasticity mechanism. Therefore, serial lesions will be easier to repair: with each small lesion only little information is lost, which can easily be repaired. This difference between a single stage lesion and a serial lesion is shown in Figure 2.3a and b. The serial lesion effect shows that there is an interaction effect between the amount of redundancy left in a system and the degree to which a system can be repaired.

2.5 Self-repair in connectionist systems

2.5.1 Redundancy in a connectionist network

In Section 2.3.1 we discussed that the brain possesses redundancy on several levels of the brain, namely synaptic, neural, and neural systems level. Since a connectionist system is a network, it possesses network redundancy. We will first discuss the network redundancy in connectionist systems and discuss how the other types of brain redundancy can be modelled in these systems.

Redundancy in connectionist systems resides in the connections between nodes. Despite the loss or weakening of some connections, a memory representation may still be connected well enough to be retrieved. By retrieval of a well-connected memory representation we mean that activation of a subset of nodes will lead to the activation of the full representation. Retrieval is thus implemented as a process in which some nodes of a memory representation are activated - in the human brain analog, say, by input from another cortical circuit or sensory input - followed by a process of “spreading activation” that will activate all nodes.

In a connectionist system, redundancy can also reside in different memory traces. The nodes of the network can be part of a single memory trace, but may also be part of several memory traces. If one part of a memory trace is damaged to such an extent that it cannot be

activated or retrieved from a given group of nodes A, it is still possible that another group B can activate the damaged part. This activation process can be very complex involving many memory representations. From a connectionist point-of-view, a network is redundant as long as the nodes of the network are sufficiently connected for retrieval through spreading activation. In such pattern completion, nodes may be activated via nodes of its own memory representation or via other associated memory representations.

A node in a connectionist network can represent a concept of varying levels of abstraction. For example, in the classical network by Rumelhart and McClelland (Rumelhart & McClelland, 1981) a node represents a simple feature ranging from one letter to a whole word. The same difference in level of abstraction of representation applies to a group of nodes. It can stand for anything from the representation of a word meaning, the association between two words, to an entire cognitive function such as the ability to recognize faces or to direct attention to a position in the visual field. This rather loose definition is deliberate as we propose that the principles of self-repair as set forth here may apply to many levels of representation, from a single episodic memory to a complete cognitive function.

The exact degree of redundancy is dependent on the type of neural network. To give an idea about the influence of this, we will give some examples of aspects of memory coding determining redundancy. Coding here refers to the way memories are represented in connectionist systems or in artificial neural networks. For a more thorough overview of factors determining neural network redundancy see for example (Cowan, 1995). For a detailed analysis of redundancy in different types of networks see for example G. Bolt (Bolt, 1991) and Tchernev et al. (Tchernev *et al.*, 2005).

A first aspect of coding determining redundancy is whether binary or real valued coding is used. This depends on the threshold function that determines when nodes or neurons of the connectionist system are activated. The activation function of a neuron determines the output on the basis of the sum of incoming activations. To illustrate how the activation function can influence redundancy we will discuss two commonly used activation functions, the sigmoid function and the binary threshold function. With the sigmoid function the output of a neuron can be any real value between zero and one. Coding following a sigmoid function yields relatively specific values, for example 0.3455. If retrieval is conceptualized as the reproduction of this exact value with little margin for error, one can imagine that the network has little redundancy. With the removal of one input connection, the neuron generates a (slightly) different output.

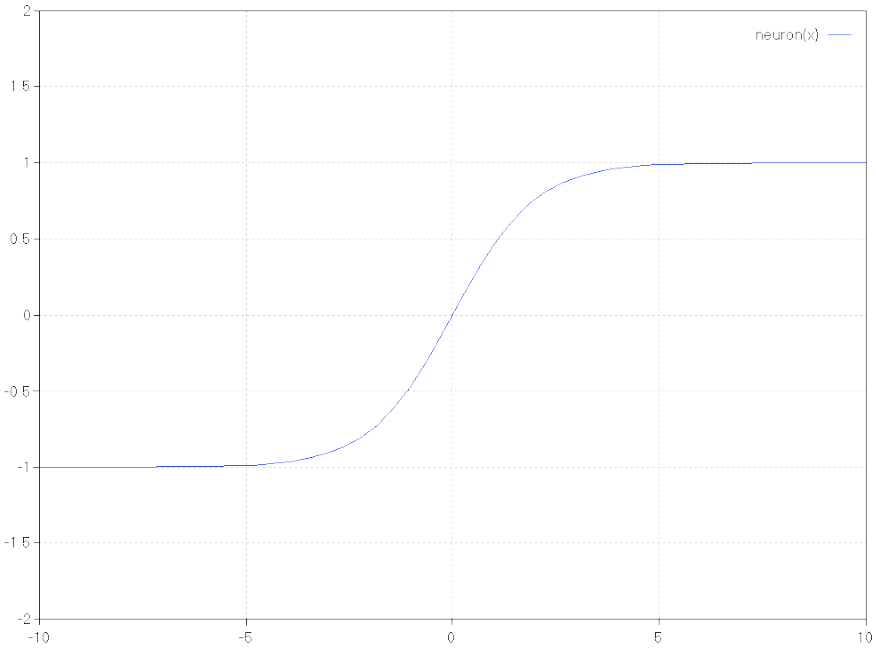


Figure 2.2. This figure shows an example of a sigmoid activation function, where the x-axis represents the input of the neuron and the y-axis represents the output of the neuron. The maximum and minimum output, one and minus one, are approached when the input of the neuron reaches ten and minus ten. If the input of the neuron is zero the output also is zero.

With the binary threshold function the output of neurons is either zero or one. If one connection is lost, the threshold may still be exceeded and the output remains unchanged.

A second aspect of coding that determines redundancy is the degree of distributeness of memory representations. A simple example suffices to illustrate this. Suppose a memory is not distributed over different neurons and that one neuron represents a memory. If this neuron is connected to five input neurons that are needed to activate the output neuron, then a neuron will have five connections. If all of these are lost, the neuron cannot be activated and subsequently the memory is lost. Alternatively, suppose that a memory representation is distributed over five neurons. These five (output) neurons are connected to five input neurons. If five connections are lost, depending on which connections are lost, all 5 neurons may still be activated. The distributed nature of memory representations provides a connectionist system with the property of graceful degradation (Anderson, 1983), where with an increasing amount of damage neural network performance will slowly decrease instead of failing catastrophically, as is the case in digital or symbolic systems.

A third aspect of coding determining the degree of redundancy is whether memory representations are orthogonal or non-orthogonal (overlapping). If memory representations are

overlapping there may be interference caused by co-activation of other patterns during retrieval. It is possible that a memory representation is only partially activated or that parts of other memory representation are activated. In the latter case we speak of retrieval of spurious patterns. With orthogonal memory representations the effect of interference of spurious patterns is smaller. Intuitively, this also makes it clear why sparse coding, where a memory is stored in a small fraction of the number of network neurons, leads to better retrieval performance (Amari, 1989), because the overlap between memory representations is smaller. Although connections between memory representations can lead to inference, it can also be profitable with regard to redundancy: other memory representations can contribute to activate a memory representation. We will investigate this property thoroughly in Chapter Two, where we will see that there is a tradeoff between interference effects and memory retrieval via other memory representations.

A fourth aspect of coding determining the degree of redundancy is the temporal component. The two main types are temporal coding and rate coding (for more details see (Gerstner & Kistler, 2002; Rieke *et al.*, 1999)). With temporal coding, the precise timing of spikes from a number of input neurons activates the output neuron. With rate coding, the temporal average over the spikes from the pre-synaptic neurons determines the activation of the post-synaptic neuron. If connections are lost and temporal coding is used it is more difficult to attain the required number of simultaneous spikes than to attain an average number of spikes over a given period. In other words, temporal coding may result in less redundant network behavior than rate coding.

We will now discuss the different aspects of coding with respect to the brain. The first aspect of coding determining redundancy related to binary versus real-valued coding. It is more or less agreed that real neurons generate spikes, that is, they fire in an all-or-non fashion and thus have a binary threshold function (Rieke et al., 1999). The second aspect of coding determining redundancy was the degree of distributivity of memory representations. Single neurons code for something quite simple, for example a neuron in the primary auditory cortex responds to a tone of a certain frequency. Most neurons, however, are not activated alone, but together with other neurons (deCharms & Zador, 2000). For instance, complex objects in macaques are represented by combinations of feature columns, each column containing numerous neurons (Tsunoda *et al.*, 2001). The third aspect of coding determining the degree of redundancy was whether memory representations are orthogonal or non-orthogonal (overlapping). Although the memory representations for very simple features like a tone might not be overlapping, the memory representations for more complex tasks overlap. The

previous example of macaques showed that complex objects are represented by different feature columns. It is possible that one of those feature columns is involved in several objects. It has been shown for cortical regions that they are involved in several memory traces (Haxby et al., 2001). The fourth aspect of coding determining the degree of redundancy related to temporal components of the neural code. It is agreed upon that both rate coding and temporal are used by the brain (deCharms and Zador (deCharms & Zador, 2000) and Rieke et al. (Rieke et al., 1999)). A clear example is encoding of tactile information, where three different cortical areas can use both coding strategies to represent the location of the tactile stimulus (Nicolelis *et al.*, 1998).

In the rest of this thesis we mainly investigate network redundancy. Synaptic redundancy can be modeled by for instance increasing the weight of a connection. Redundancy by neuron replacement could also be modeled by for instance allowing connections always to grow back even if all weights of a neuron are zero. However, questions like how new neurons are created and how they are directed to the right place are not addressed by this thesis.

2.5.2 Maintenance in connectionists networks

The above exposition shows how connectionist systems may possess redundancy. Redundancy in a connectionist system implies that a damaged memory representation can still be retrieved. What could carry out the repair or maintenance (of redundancy) in the connectionist system? In this section we will present two mechanisms that are derived from the plasticity data of Section 2.3.2. We will argue that these mechanisms, supposed to be involved in neural circuit changes in the intact brain, may repair minor damage and in that way can maintain redundancy.

A first type of mechanism that may be involved in brain repair is the Hebbian learning mechanism. It was introduced by Hebb in his seminal work ‘The organization of behavior’ (Hebb, 1949). Hebb's theory states: *“Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability.... When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased*“(p. 62). Hebbian learning is supported by data of long-term potentiation. It resembles Hebbian learning, because its induction requires a simultaneously pre-synaptic neurotransmitter release and post-synaptic depolarisation. The original Hebb learning rule of strengthening of

connections was later extended by weakening of connections (Koch, 1999). Theoretically connection weakening was addressed by for example Sejnowski (T.J. Sejnowski, 1977) and Palm (Palm, 1982; T.J. Sejnowski, 1977). Hebbian weakening is supported by the data of long-term depression mentioned in Section 2.3.2.

Hebbian learning restores connectivity within a memory representation, because it strengthens the connections of two simultaneously activated nodes. This process of weight strengthening can enhance redundancy and undo possible damage. Figure 2.3 summarizes the main principles of how Hebbian learning can repair damage. To summarize it shortly, the intact neural circuit in 1a loses or has weakened connections as depicted in 2.3b. The lost or weakened connections are regained through a process of activation with Hebbian learning as depicted in 2.3c-2.3e.

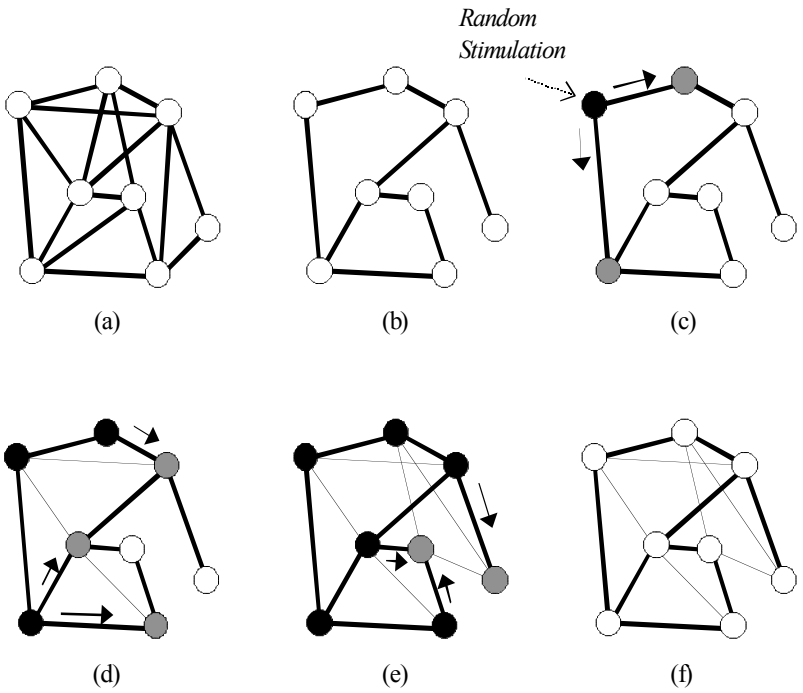


Figure 2.3. Schematic illustration of automous reconnection through maintenance of redundancy. The circles represent neural groups, while the lines indicate the tracts in the neural circuit. Activated neural groups are shown as black, filled circles. (a) A well connected, intact neural circuit. (b) After diffuse lesioning the same circuit is still connected but less densely. Self-repair takes place (c) by the activation of neural groups because of an external cue. The activation spreads over the neural circuit (d) and (e) while a Hebbian learning process forms connections, not necessarily the same as the original ones. (f) After this repair stage, the circuit is again well connected and the resultant circuit is now less vulnarable to further lesioning, compared to its pre-repair state (b).

The second mechanism is the homeostasis mechanism that is based on the homeostasis data of Section 2.3.2. This mechanism can be modeled by normalizing the

weights of all incoming or outgoing connections of a node. This is the way how it will be implemented in this thesis. A more sophisticated homeostasis mechanism was investigated by Horn et al. (Horn *et al.*, 1998a, 1998b). This mechanism uses random activation to determine the strength of the basin of attraction of a memory representation. On the basis of this strength the weights of the memory representations are downscaled. This keeps the neuronal firing rate stable irrespective of external input changes just as was found by Turrigiano et al. (Turrigiano et al., 1998). The mechanism was tested by adding noise to the weights. Given that the size of the errors remained within a certain order of magnitude the mechanism was able to maintain its memory and undo itself from errors. If we interpret the errors as lesions this simulation shows that a homeostasis mechanism can counteract damage.

2.5.3 Demonstration of self-repair in a connectionist network

The self-repair model consists of multiple lesion-repair cycles similar to the serial lesion effect, where a lesion is followed by a period in which repair could take place. We take it as an assumption that repair takes place after a lesion. Repair in this thesis is self-repair that is modelled by a three-step process in which (1) neurons are activated, (2) activation is allowed to spread to connected neurons, and (3) connections are updated. The intensity of self-repair as is modeled here depends on the stimuli: The more stimuli the more self-repair, which is consistent with the 'use-it-or-lose-it' principle. The update of connections in this thesis is through the Hebbian learning mechanism possible combined with a homeostasis mechanism in the form of simple normalization.

We illustrate and demonstrate the effect of self-repair in the Hopfield model (Hopfield, 1982), a well-investigated connectionist model. The network was initially trained with five randomly generated patterns at the first time step. Then at each following time step a repair epoch and a test determining how well stored patterns are still present take place, where a repair epoch consists of a lesion-repair cycle. Figure 2.4 shows how the accumulated noise rapidly degrades the performance of the non-repaired network, whereas the repaired network preserves its memory representation**.** More details of this simulation can be found in the next chapter.

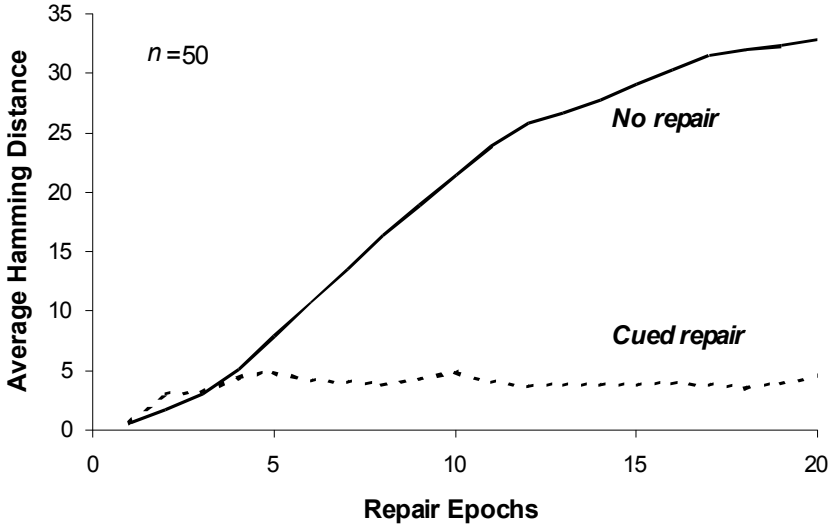


Figure 2.4. Simulation of self-repair with a Hopfield network using slightly distorted cues during repair. The curves are based on 50 replications. See text for an explanation.

2.6 Discussion

In this chapter, we introduced the idea of maintenance of redundancy and demonstrated how this can greatly extend memory lifetime. We reviewed neurobiological data of redundancy and plasticity. We, furthermore, reviewed behavioral data of the use-it-or-lose-it principle supporting the idea of maintenance and the serial lesion effect, which supports our postulate of maintenance of redundancy. Derived from the data we will build a model of self-repair. The procedure of lesions followed by repair is similar to the serial lesion effect. The intensity of self-repair is dependent on external stimuli similar to the 'use-it-or-lose-it' principle. Redundancy is present as some form of abundance of connections, while plasticity mechanisms, derived from the plasticity data (Section 2.3.2) maintains redundancy at some minimal, safe level. Moreover, in this model we demonstrated that self-repair is able to extend memory lifetime. In the remainder of this section, we will explain how the ideas of self-repair can be applied to brain recovery after damage and normal aging.

The principles of self-repair can also be applied to the aging of a normal, intact brain. The aging brain is subjected to constant damage. One of the causes of aging may be oxidative processes (Barja, 2004; Harman, 1956) that directly affect genes responsible for learning and memory (Lu *et al.*, 2004) and may lead to gray or white matter lesions (Resnick *et al.*, 2003;

Salat *et al.*, 1999). Self-repair during aging resembles the serial lesion effect with the difference that damage is very small; this makes it possible that autonomous recovery with complete neural restitution can take place. The degree of possible restitution depends on lesion size and the amount of repair. The amount of repair is determined by the amount and type of activities one is engaged in, which drive the plasticity mechanisms for maintenance. Similar to brain recovery after damage, repair during the aging process has its limits. If the (accumulated) damage has become too large and redundancy has been eaten away, maladaptive repair can take place. An example may be hyperintensities (DeCarli *et al.*, 1995; Yikoski *et al.*, 1995), which are accumulations of white matter following ischemic insults. Because lesions of the insults are too extensive, the information left after damage is insufficient for autonomous self-repair during aging and resulted therefore in maladaptive repair and aberrant wiring.

One of the implications of self-repair in normal aging is that plasticity mechanisms taking place in it are able to carry out self-repair. The simulations of the previous section and of Horn et al. (Horn et al., 1998a, 1998b) show that these mechanisms in principle are able to do it. The idea is not new, it has been suggested before that neural plasticity mechanisms in the normal vertebrate brain are able to repair damage up to a limit (Bailey & Kandel, 1993; Cotman & Nieto-Sampedro, 1982; Kolb, 1999; Xerri *et al.*, 1998). Cotman & Nieto-Sampedro (Cotman & Nieto-Sampedro, 1982) for example argued that reactive growth and synapse renewal are an extension of the normal operation and maintenance of brain circuits. The two types of mechanisms we discussed are both found in a normal functioning brain. Is it possible that these mechanisms do extend the normal operation and maintenance of brain circuits? The question is whether these particular mechanisms can repair damage to some extent. It is not so easy to verify this hypothesis empirically in a normal brain. How do we detect small changes, the diffuse lesions and subsequent repair, in brain connectivity? Until now there are no methods to measure such changes. So far our knowledge of the behavioural effect of damage to the connectivity and subsequent repair of connectivity comes from experiments with larger damage (for example see Kolb (Kolb, 1995) and Ramirez et al. (Ramirez et al., 1999)). Unfortunately, these experiments do not show (or did not investigate) the involvement of plasticity mechanism of the normal functioning brain. Evidence of the involvement of these mechanisms is mostly indirect and we cannot, therefore, draw definite conclusions.

The exact involvement of plasticity mechanisms discussed in this chapter in self-repair remains to be established and further research has to be carried out to unravel the cellular and

synaptic mechanisms underlying the formation and maintenance of neural circuitry in the normal functioning brain. It is possible that minor damage triggers a different set of mechanisms than normal changes in the normal adult brain. This, however, would not disprove the main point of this chapter, namely that memory in the brain possesses redundancy and that this redundancy in memory is somehow maintained.

3

Self-repairing neural networks

Abstract

This chapter is a first exploration of self-repair with mathematical and connectionist models. We investigate redundancy, which in neural networks is present in the connections, with random graph theory. Then we address the question whether self-repair is possible in connectionist models at all and what types of self-repair are possible. Concretely, the latter two research questions imply that we will study guided and autonomous self-repair in the classical connectionist Hopfield model (Hopfield, 1982). In a soft *k*-winner-take-all network with stochastic neurons we will also demonstrate self-repair and explore it further in this model. Investigations in these models show that with continuous lesioning and repair, the self-repair process must be constrained, otherwise runaway processes lead to a degenerate representation where a single pattern overtakes all resources. Several such constraints are proposed and implemented. Finally, we discuss some issues regarding self-repair in the brain.

3.1 Self-repair as maintenance of redundancy

This chapter explores how neural networks can repair themselves after a certain percentage of their connection weights has been removed or perturbed by addition of a noise term. Our approach to self-repair is based on maintenance of redundancy that remains after such diffuse lesions. The main question we will address in this chapter is whether self-repair is possible in connectionist networks at all.

In neural networks, it is not immediately clear how much redundancy is present, because each neuron may support a different part of a representation and we cannot say that the neurons themselves are somehow copies of each other. The redundancy resides in the connections of the network. Self-repair is modelled by a three-step process in which (1) nodes are activated, (2) activation is allowed to spread over the rest of the connected nodes, and (3) connections are added between activated units or weights on existing connections are updated. This process is illustrated in Figure 3.1. Our repair mechanism uses the redundancy found in most types of attractor networks. One of the consequences of this redundancy is that if all synapses receive a small random perturbation, even an incomplete cue may still allow perfect recall of the original pattern. As we shall demonstrate, the success of the self-repair mechanism is strongly dependent on near-perfect retrieval.

We will first investigate some of the characteristics of self-repair, mainly redundancy, using some of the results of the theory of random graphs. This is followed a brief analysis of self-repair in Hopfield networks (Hopfield, 1982). After having proven that self-repair can work in theory in Hopfield networks, we will investigate it further with simulation studies to find out the details and to find out whether self-repair with randomly generated cues (autonomous self-repair) is possible. It will be shown that autonomous self-repair can only work with an adapted Hopfield model. To attain a more realistic model of memory in this model, we will investigate overlapping patterns. We will further explore self-repair in a more complex network, namely in the neocortex part of a model of long-term memory (Meeter & Murre, 2005; Murre, 1996). In this model, we will also demonstrate self-repair and investigate the effect of stimulus type and learning rule that are two important parameters of self-repair. In the last section, we will discuss the necessity of self-repair in the brain and why we chose to model self-repair by changes in connectivity.

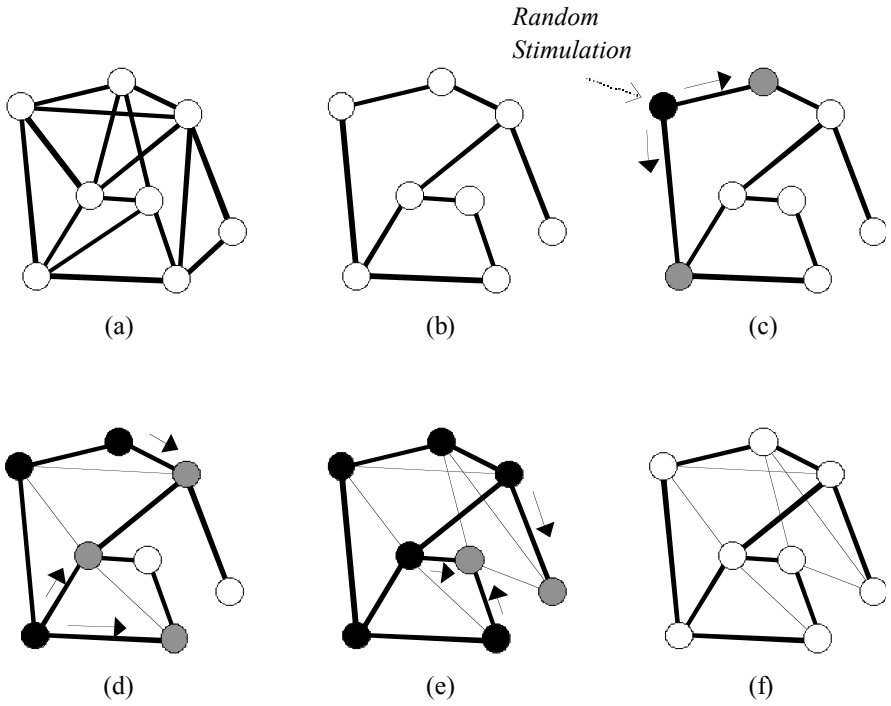


Figure 3.1. Schematic illustration of automous reconnection through maintenance of redundancy. The circles represent neural groups, while the lines indicate the tracts in the neural circuit. Activated neural groups are shown as black, filled circles. (a) A well connected, intact neural circuit. (b) After diffuse lesioning the same circuit is still connected but less densely. (c) Some neural groups become activated through an external cue. (d) Activation spreads through the circuit, while a Hebbian learning process forms connections, not necessarily the same as the original ones. (e) After this repair stage, the circuit is again well connected. (f) The resultant circuit is now less vulnarable to further lesioning, compared to its pre-repair state (b).

3.2 Self-repair and random graph theory

Random graphs have been analyzed extensively in the past decades (Bollobás, 1985). We will here use this framework to explore some of the limits of self-repair. An intact memory representation in a neural network, considered in isolation and ignoring overlap with other patterns, can be viewed as a connected random graph (a graph is called connected if path exists between each pair of nodes in the graph). Random deletion of connections (diffuse 'lesions') may cause it to be no longer connected. Repair can counteract such critical lowering of the connectivity. If some nodes are activated (e.g., by a random 'cue'), an entire representation will be activated through spreading activation. Through Hebbian learning new connections can then be added between activated nodes. This suggests that an analysis in terms of random graphs may be helpful in understanding self-repair in neural networks, which is what we shall explore in the next section.

In neural networks, most researchers speak of nodes or artificial neurons with connections between them. Using a different terminology, graph theorists define a graph G as a set of vertices V and a set of edges E that connect some or all of the vertices. In one basic model, a random graph G_p is defined as a set of vertices $V = \{1, 2, \ldots, n\}$ in which the *edges* are chosen independently with probability f, $0 < f < 1$ (Bollobás, 1985, p.32). In the preface of his book from 1985, reviewing the theory of random graph theory since its beginning in the 1950s by Erdös and Rényi (Erdös & Rényi, 1959), Bollobás remarks that

> *"It is often helpful to imagine a random graph as a living organism which evolves with time. It is born with as a set of n isolated vertices and develops by successively acquiring edges at random. Our main aim is to determine at what stage of the evolution a particular property of the graph is likely to arise." (p.ix).*

For the purposes of the present chapter, we take the inverse perspective and consider a random graph as the brain of a living organism that incurs diffuse lesions through synaptic turnover, aging, disease, and trauma. In this process, it loses connections at random and we are interested to know at what stage of neural decline certain properties are likely to vanish.

Random graph theory has been applied usefully to other topics in neural networks, such as the theory of cell assemblies (Palm, 1982), the analysis of learning procedures in neural networks (Feldman, 2000), the aging brain (Cerella & Hale, 1994), models of synfire chains and the structure of the cortex (Bienenstock, 1995). The notion of self-repair in the brain is derived in part from the field of reliability engineering: Earlier work on this related to recovery from brain damage, considers failure rates of parallel and serial subsystems in the brain under various conditions (Glassman, 1987). This work, however, does not pursue subsystems that are connected in complex ways (i.e., not either strictly serial or parallel), because such connection patterns give rise to very complicated mathematics. We approximate this more general case with random graph theory. Our approach differs from an earlier model by (Petsche & Dickinson, 1990). They propose an intricate neural network based on so called trellis codes (a form of redundant coding), which is not only fault-tolerant but also able to repair itself. Their model hinges on using a very specific (highly non-random) neural network architecture that, while being very effective, is not biologically plausible.

In our analysis, a neural network representation is equated with a random graph of n nodes. There is a connection between any two nodes with a probability f, which we shall call the connectivity factor. In a fully connected pattern, $f = 1.0$. A pattern can be completed (via spreading activation) with certainty from any possible sub-pattern, but only if the latter is

connected to the rest of the graph. This implies that a path of connected nodes must exist between any two nodes in the graph. In a non-connected graph, activation of these isolated nodes (or sub-graphs) cannot spread to the other nodes and pattern completion can, therefore, not occur in this case. If a graph is connected, however, pattern completion—and hence self-repair—can always take place. It is, therefore, important to determine graph connectivity

Connectivity of undirected random graphs has been well researched and is discussed extensively by Bollobás (1985). The probability of being connected can be approximated for large random graphs and calculated exactly for small graphs. In general, larger graphs have a higher probability of being connected compared to smaller graphs, if they have the same f value. For example, for a small graph with 100 nodes and a connectivity factor $f = 0.10$, the probability of being connected, p, is nearly 1.0. If f drops to 0.038, p shows a 90% drop to 0.10. A very small graph ($n = 10$) with a connectivity factor of $f = 0.163$ still has a near-zero probability of being connected. We have to raise f to 0.234 to obtain a p of 0.437 (see Bollobás, 1985, p. 399ff). As long as a graph is connected, additional connections can be formed through a Hebbian 'learning' process, adding connections in a random fashion. This increases the connectivity value f to a safer value. The probability p of a random graph being connected is expressed by the following formula, which is taken from Theorem VII.3, (Bollobás, 1985), p.150). It is valid for graphs with a large number of nodes, n:

$$p = e^{-e^{-(fn-\log n)}} \tag{1}$$

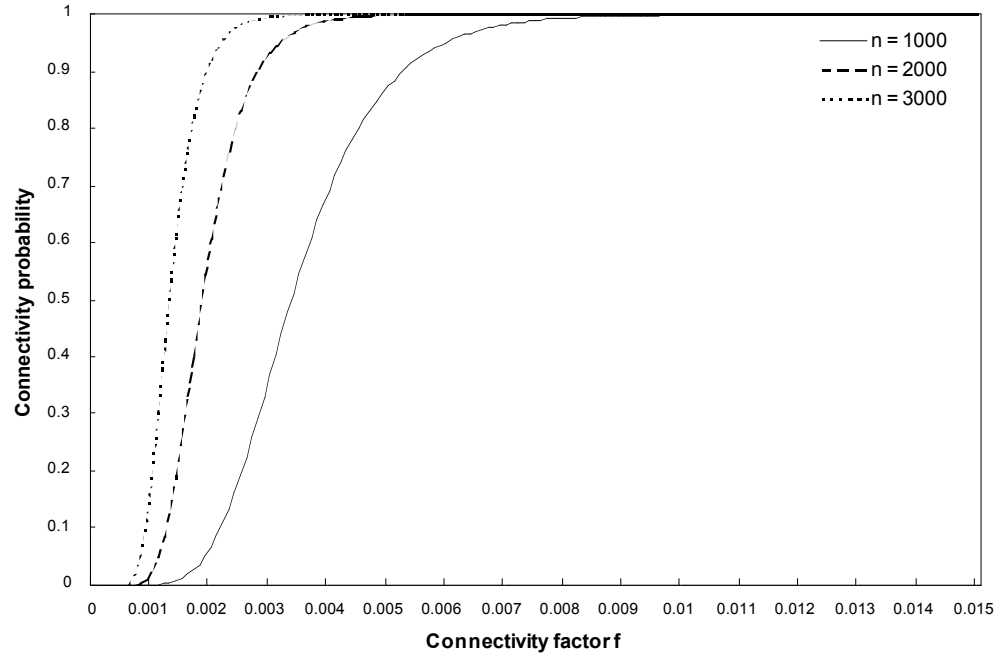
In Figure 3.2a, we have plotted the connection probability p for values of f in the range 0.0 to 0.015 and for graphs of different size n. We can observe the following properties for random graphs. For a given connectivity value f:

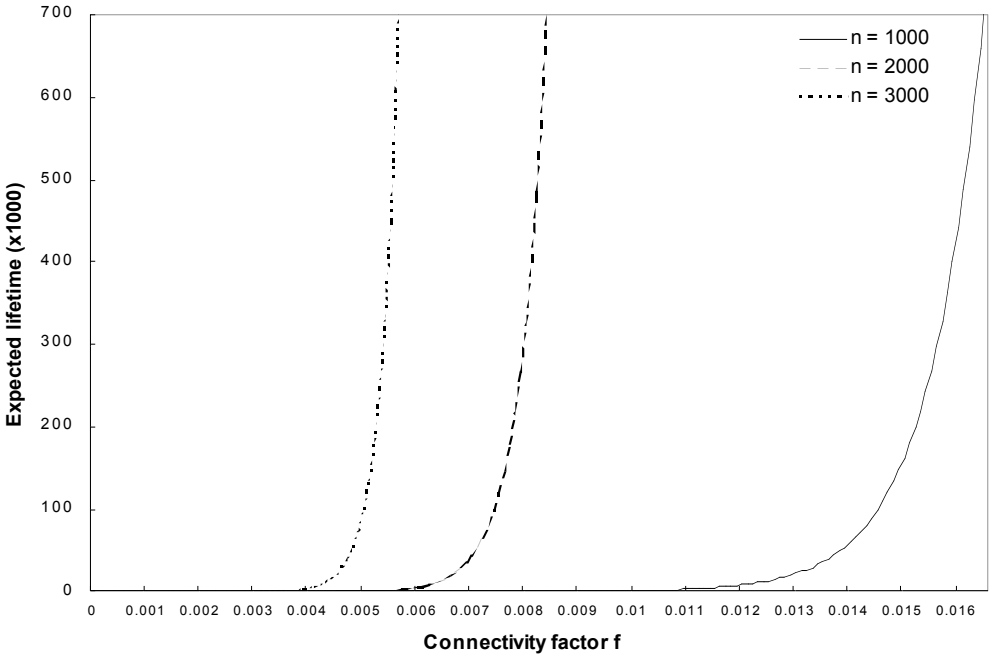
(i) The probability of connectivity is much higher for a large graph. For example, for $f = 0.005$, increasing the size of the graph from $n = 1000$ to 3000 raises p from near-zero to near-one.

(ii) The average slope of large graphs is much steeper than that of smaller graphs.

These general findings also extend too much smaller graphs as was illustrated above with graphs of sizes $n = 10$ and $n = 100$.

We now extend the example of the security agents of Chapter Two (Section 2.2) to the present discussion of random graphs. Suppose that we have a large but weakly coupled neural network representation with n nodes with a certain low connectivity value. Suppose, furthermore, that at each time interval the network is lesioned randomly so that at the end of the interval the connectivity factor is f.





(b)

Figure 3.2. Illustration of some analytical graph theoretical results. (a) Probability of a graph being connected as a function of the probability *f* that a connection exists between any two given nodes. Connectivity probability has been plotted for graphs of sizes 1000, 2000, and 3000. (b) Expected lifetime in lesion-repair cycles where *f* is the probability *f* that a connection exists between any two given nodes after the lesion has been applied.

We could, for example, start each interval with a connectivity factor of 0.10 and then lesion it with 90% to arrive at a final connectivity factor $f = 0.01$. Suppose, furthermore, that after lesioning the connectivity will be restored (repaired) randomly to its original value (i.e., to $f = 0.10$ in the example here), but only if such repair is still possible. The latter is the case, only if the graph is still connected after lesioning. We will use a very simple approach to repair, here, where we: (1) activate one randomly selected node, (2) have activation spread to other nodes to which a path exists, and (3) apply Hebbian stochastic learning through randomly adding connections between activated nodes (if a connection already exist, no new ones are added).

This repair mechanism fulfils a similar function as the exchange of manuscript copies in the copy example above, although it uses a different mechanism. We are now interested in the lifetime of a network that is lesioned to a critical value f and repaired in this manner. We can approximate p, the probability of the graph being connected with in equation (1) above. The lifetime, then, has a geometrical distribution with mean $p/(1-p)$. The expected lifetime, expressed in lesion-repair intervals, has been plotted in Figure 3.2b for critical f-values in the range 0.0 to 0.02 and for graphs of various sizes. Given that we use a repair mechanism as above, we can note the following properties:

(iii) Large graphs have a much longer lifetime than small graphs.

(iv) For large graphs, the life expectancy rises extremely steeply as a function of f.

(v) The slope of the life expectancy is much larger for large graphs than for small graphs.

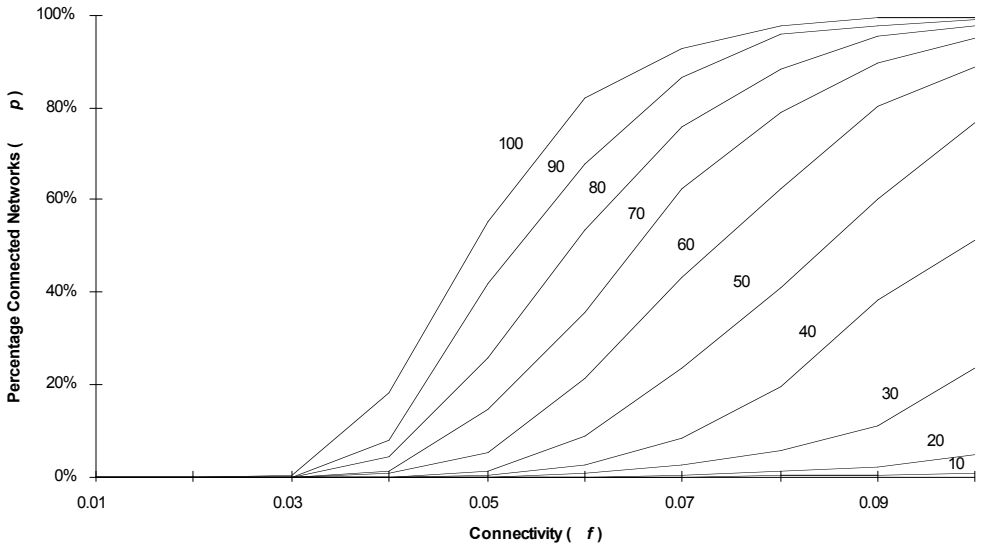
A general observation is that a process of continuous repair ensures very long lifetimes of this type of neural network representations even when they are exposed to extremely high levels of cumulative noise and damage in the form of diffuse lesions.

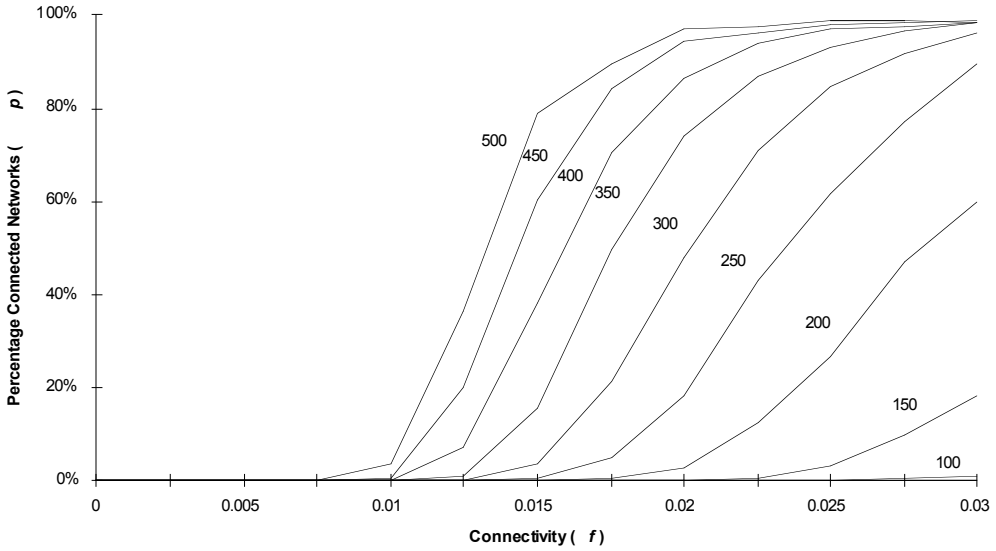
If we did not include a repair process, the expected f-value would in a few intervals drop to a level where connectivity is very unlikely (i.e., approaches zero). For example, a drop from 0.10 to $f = 0.001$ can be accomplished in a little over two intervals with 90% lesions and would almost certainly reduce connectivity to zero (see Figure 3.2a). A general conclusion can be derived from both the copy example and the graph theory. Multiple small lesions with continuous repair result in dramatically longer lifetimes compared with either multiple small lesions without repair or compared with a large lesion of the same size as the cumulative effect of the small lesions.

Connectivity probabilities only tell us something about the *extremes* of pattern completion: the case where a very small sub-pattern is still able to activate the entire representation. Generalizations to the case where we want to calculate completion of a graph from k activated nodes out of n and many other generalizations specific to neural networks are

necessary to increase the applicability of random graph theory to the theory of recovery from brain damage. Unfortunately, many of the results we are interested in have not been derived so that for now we will take recourse to a computational exploration.

Many recurrent neural networks have asymmetric connections, where for example a node *A* is connected to a node *B*, but where there is no connection from *B* to *A*. We simulated random graphs with this property (directed random graphs). The results are shown in Figure 3.3.





(b)

Figure 3.3. Simulation of connectivity in directed graphs. Plotted is the percentage of graphs that were found to be connected as a function of between-node connectivity probability *f*. Each data points is based on 1000 replications. (a) Graph sizes of 10 to 100 nodes. (b) Graph sizes of 100 to 500 nodes.

The figure shows that both increasing the connectivity and increasing the number of nodes in the network cause an increase in the graph connectedness probability p, as was described by random graph theory. There are some small differences: p values of directed random graphs are somewhat lower than those calculated by Bollobas (1985). The five properties above, however, remain valid.

We also investigated the property that under constant lesioning an undirected random graph tends to break into one very large connected graph ('giant component', (Bollobás, 1985)) with many isolated very small graphs. Our simulations observed this also for directed graphs. A consequence of this property is that above some very small number, it does not matter very much how many nodes are used to cue a damaged representation. On the one hand, randomly activating only a few nodes virtually always guarantees hitting the 'giant component'. On the other hand, one has to activate nearly all nodes in order to achieve any measure of certainty of hitting the isolated small graphs. We can, thus, conclude that the cue size is thus not a critical factor for repair when considering representations in isolation, as we do here. In case of multiple, overlapping representations, however, cue size may be important for the selection and retrieval of unique patterns.

Graph theoretical results apply mostly to sparsely activated networks, in which patterns overlap little. Below we will indeed study the repair behavior of a k-winner-take-all network model, but we will first show that self-repair can also be obtained in cases where there is very strong overlap, namely in Hopfield (1982) networks.

3.3 Self-repair in Hopfield networks

Like in most other types of networks, self-repair in Hopfield networks depends on typical connectionist features such as graceful degradation, pattern completion, distributed representations and learning capacity. These features have been studied extensively and their exact character is paradigm dependent. For example, the original paper (Hopfield, 1982) shows that approximately $0.15n$ patterns can be reliably encoded in a Hopfield network, where n is the number of nodes in the networks. If the number of patterns exceeds this number, completion to the original becomes unreliable.

In a Hopfield network, the weight on a connection from neuron j to i is formed by increments T^{s}_{ij}, each coding the co-occurrence of binary signals in a specific pattern S, using a Hebbian learning rule. At the beginning of a lesion-repair cycle, a small lesion is administered to each connection weight by adding a random weight perturbation e_t. This lesion is followed by a repair cycle during which we use a partial cue to recall the pattern $\tilde{S}$. We will assume

that the cue suffices to retrieve the original pattern so that we will be able to undo part of the perturbation by storing $\tilde{S}$ again:

$$\tilde{T}_{ij}^{\tilde{S}} = (2\tilde{V}_i^{\tilde{s}} - 1)(2\tilde{V}_j^{\tilde{s}} - 1) \quad (2)$$

where $\tilde{V}_i^{\tilde{S}}$ is the *i*-th element in $\tilde{S}$. This procedure is repeated for all stored patterns.

In case of one stored pattern *S* we now have for each synapse two increments plus perturbation:

$$T_{ij}^{S}(t) + \tilde{T}_{ij}^{\tilde{S}}(t) + e_t$$

Normalizing the synapse strength (in this case through division by 2) completes the repair cycle. With perfect recall we have $\tilde{S} = S$ and $T_{ij}^{S}(t) = \tilde{T}_{ij}^{\tilde{S}}(t)$, giving after $t + 1$ repair cycles

$$T_{ij}^{S}(t+1) = \left\{T_{ij}^{S}(t) + \tilde{T}_{ij}^{\tilde{S}}(t) + e_t\right\}\frac{1}{2} = T_{ij}^{S}(t) + \frac{1}{2}e_t .$$

In other words, a single repair cycle will reduce the relative effect of a perturbation by 50%. Multiple repair cycles can diminish perturbations to arbitrarily low levels, as long as perfect retrieval of the original patterns is obtained. We explored this conclusion in two simulations, first using a slightly degraded cue (within the Hopfield bounds for good retrieval) and then using random cues.

Simulation 1. Slightly distorted cues. Figure 3.4 shows the result of a simulation of self-repair in a Hopfield network with 100 nodes. For each replication, the network was trained with five randomly generated patterns. Lesion-repair cycle of a replication was executed as follows. (1) The model was lesioned (perturbed) by adding uniform noise in [-2.0,2.0] to each weight. (2) A repair trial was carried out for each of the five patterns. For each repair trial, a cue (10% distortion of an original pattern) was presented to the network. Activations were (asynchronously) updated, and weights were updated according to Equation (2). After each lesion-repair cycle, the network was tested on each stored pattern by presenting it with a cue (10% distortion of the original pattern). The resulting pattern is compared with the stored pattern and their difference is expressed in the Hamming distance. This allows us to follow the degradation of the pattern representations under continuous noise. Figure 3.4 shows how the accumulated noise rapidly degrades the performance of the non-repaired network, but the repaired network preserves its memory representation**.**

A possible criticism of the above simulation is that the repair process itself is dependent on continued access to the original patterns. Also, it is conceivable that relearning is taking place on the basis of the 90% of the patterns that is not distorted. The self-repair

method would, therefore, be more interesting, if it could work with completely random cues. From an engineering perspective the most interesting case would be repair with completely random cues that would allow the system to repair/recover itself from an arbitrary state, as is the case in self-stabilizing systems (Dolev, 2000).

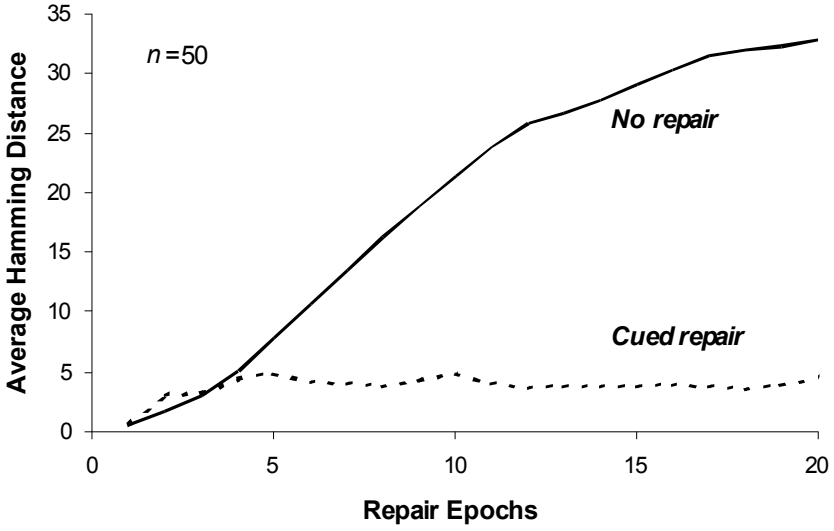


Figure 3.4. Simulation of selfrepair with a Hopfield network using slightly distorted cues during repair. The curves as based on 50 replications. See text for an explanation.

A problem is that for self-repair to take place it is imperative that: (i) Near-perfect pattern retrieval always takes place. If this is not the case, the network will learn spurious patterns and exhibit 'faulty repair'. (ii) All patterns have to be repaired approximately equally often. If these conditions are not met our initial simulations yielded the following: a few patterns become strong, which in turn may cause them to be retrieved more often. This in turn will tend to strengthen the strong pattern even further. In case of pure random cueing, this may rapidly lead to a self-reinforcing process of strengthening of a single pattern. This runaway effect may eventually cause only a single pattern to survive while all others are forgotten. Randomly driven and unconstrained consolidation and repair strategies tend to suffer from this runaway effect (Hasselmo, 1994; Meeter, 2003).

Hopfield type networks are, unfortunately, not well suited for randomly cued self-repair, because near-perfect completion can only be achieved by presenting the network with very slight distortions of the original patterns (10% or less, see Hopfield, 1982). By introducing a variant of the learning rule, however, we have found a method that allows

randomly driven self-repair. In the new learning rule, the weight is changed only if the post-synaptic neuron is 1 and furthermore the weights w_{ij} are bounded between –1 and 1:

$$\begin{aligned} T_{ij} &= \sum_s V_i^s (2V_j^s - 1) \\ w_{ij}(t+1) &= \max(\min(w_{ij}(t) + T_{ij}, 1), -1) \end{aligned} \tag{3}$$

Simulation 2. In Figure 3.5 the results of a simulation with this variant learning rule are given. This simulation uses fully random cueing. At the beginning of the simulation, a network with 100 nodes was trained on five non-overlapping patterns. Each pattern consisted of twenty activated nodes. At each time step, the network was lesioned by setting 10% of the connections to zero. During the self-repair cycle, 50% of the neurons was set to 1 randomly. Following this purely random cue, the network was allowed to settle into an attractor. Repair took place by applying Equation (3) for the thus retrieved memory. We varied the number of such repair cycles following a lesion in different versions of the simulation in order to study its effect. Following each lesion-repair cycle, the network was again tested for every pattern by presenting it with a 10% distortion of the original. The results indicate that stable self-repair can be achieved with this form of random cueing (see Figure 3.5a). We also found that doubling the number of cue-repair cycles, led to longer lifetimes. We, furthermore, tested self-repair with patterns that overlapped about 18%. Self-repair was stable as long as the lesions in each lesion-repair cycle did not exceed 1%. The cumulative effect of such small lesions disintegrates unrepaired patterns in about 350 time steps (see Figure 3.5b).

One of the functions of self-repair in the brain may be to safeguard our memory representations against perturbations. We, therefore, applied the repair process to a connectionist model of long-term memory. The model is called TraceLink and has been successfully applied to a wide range of characteristics of long-term memory and memory disorders (Bao *et al.*, 2001; Meeter & Murre, 2005; Murre, 1994, 1996, 1997; Murre *et al.*, 2001; Robbins & Everitt, 1996; Waelti *et al.*, 2001). In the next section, we will study whether within-cortex consolidation, without a contribution of the hippocampus, is feasible.

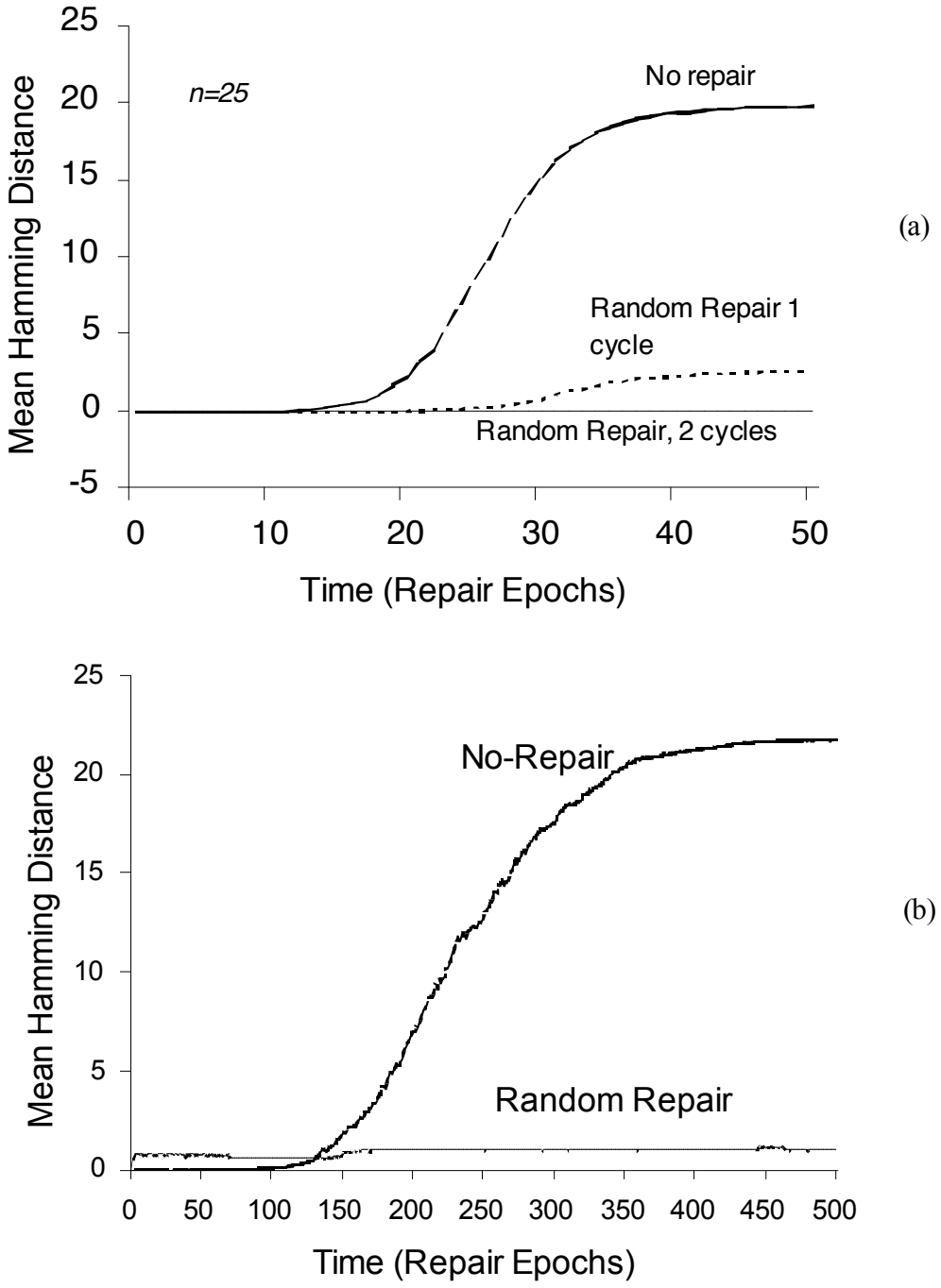


Figure 3.5. Randomly cued selfrepair in a Hopfield network (a) with non-overlapping patterns and (b) with overlapping patterns.

3.4 Self-repair in the 'cortex' part of the TraceLink model

TraceLink uses stochastic artificial neurons with synchronous update and a soft *k*-winner-take-all activation rule (see Appendix A for details of the model). A Hebbian learning rule is used. Before studying self-repair in a single simulated cortical area, we will briefly investigate

a two-area model that consists of some input area (e.g., lower visual area or somatosensory area) and a higher brain area. The lower or input area is left intact but connections to and within the higher simulated brain are lesioned. We will also assume that the input pattern remains available throughout the entire simulation. In patients with brain lesions it is often the case that lower areas of the brain are preserved and that the original stimuli (or very similar ones) are available for retraining. Indeed, during rehabilitation from brain damage, it is precisely those stimuli that are being manipulated with the goal of speeding up the recovery process (I. H. Robertson & Murre, 1999). We will use as our criterion for self-repair the extent to which the representation in the higher area remains identical to the originally established representation.

The general self-repair procedure is similar to the one followed above. We assume that initial learning has taken place, after which some lesion or perturbation is administered. Self-repair is attempted by clamping an original input pattern for a number of iterations. During this time, neurons in the damaged area fire stochastically. A threshold mechanism aims to keep the average number of activated nodes at k. Each trial consists of (a) updating all activations (synchronously), followed by (b) changing all weights using the following Hebbian learning rule:

$$\Delta w_{ij} = \begin{cases} \mu a_i a_j \text{ if } a_j = 1 \\ -\mu a_i a_j \text{ if } a_j = 0 \end{cases} \text{ with } \mu > 0 \qquad (4)$$

The learning rate is kept constant throughout this process, but the activation threshold of each non-clamped area is updated between trials. Connections can be formed between areas and within an area and are asymmetrical (i.e., it is not generally true that $w_{ij} = w_{ji}$).

Simulation 3. Illustration. The results of a single run in Figure 3.5 illustrate our approach. It shows self-repair of the trace system with a moderate lesion of 50% of the connections. Figure 3.6a shows an input area with a constantly activated pattern connected to a 'higher brain area'. The layers in the figure show the development of activations in the higher area. To facilitate visual interpretation of the simulations we have assigned a representation to the higher area that can easily be recognized (a horizontal bar). After initial learning, the input pattern is able to activate the internal representation, as shown on the far left of Figure 3.6b. Immediately after this, (moving left to right in the figure), we lesion a large proportion of the connections to the nodes of the right half of the higher area. As shown in Figure 3.6b, during iteration 1

(third from left), only the unlesioned part of the pattern remains activated. The activation level is only half of what is allowed for the higher model, so that the threshold will now start to drop gradually until about eight nodes are activated once more. During this process, each activated node may develop some connectivity to the rest of the activated pattern, as the learning parameters remain the same throughout the simulation. This gradual self-repair of the representation is apparent from iterations 1 to 149, going left to right across Figure 3.6b. Because the lesioning size was moderate and because the learning rate was sufficiently high, recovery succeeds within 150 iterations.

As a control condition, Figure 3.6c shows the same simulation but with the learning rate set to zero (i.e., no self-repair). In this case, the model remains extremely unlikely to activate the complete pattern in the right module. Instead, the left (unlesioned) part is activated plus about four random nodes in the module.

As is clear moving from left to right completion is not achieved without the learning-based self-repair process. The reason that a drop in threshold *by itself* is not sufficient for completion is that the within-area pattern interactions are not strong enough to support completion. Only when they are first strengthened can further pattern completion occur.

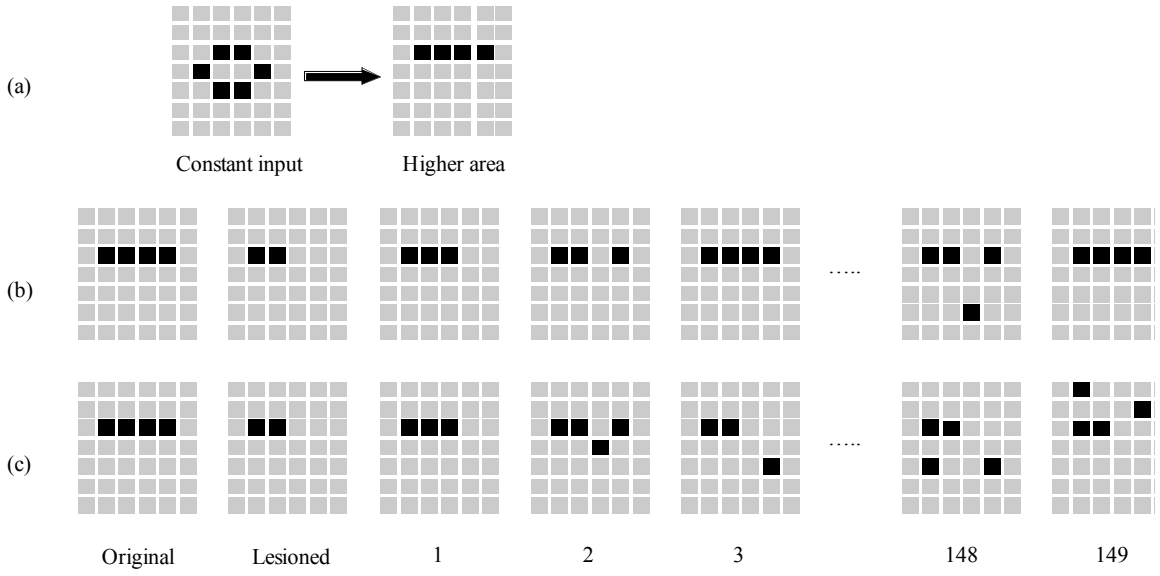


Figure 3.6. Illustration of the selfrepair mechanism in a network with stochastic nodes and soft *k*-winner-take-all. *(a)* The input pattern remains constant throughout all simulations and is fed into the a higher area. (b) Successful selfrepair after lesioning 50% of all connections to neurons in the right half of the output module. Activations are shown in the early stages (iterations 1-3) and in a later, stable stage (iterations 148-150). (c) Replication of (b), but without selfrepair. No stable completion is achieved.

Some simulation details are as follows. The network was first trained on 20 consecutive learning trials with a temporarily increased ('boosted') learning rate of 0.01. The learning rate during recovery was 0.001 in simulations Figure 3.6b and 3.6c. The temperature q was 0.3 in all cases.

Simulation 4. Repair with minimal size cues. Given the conclusions of the graph theory investigations above, we were interested in exploring whether the model could achieve self-repair by cueing a single node of the original pattern, rather then having an entire external brain area administer pattern cues. A network with 64 nodes with an initial random connectivity of 50% (i.e., 50% of possible within-network connections were set to 0) was trained with four non-overlapping patterns of 16 nodes. There were 20 lesion-repair cycles. A lesion here meant setting a fraction of the connection weights to 0 (allowing 'regrowth'). In a series of independent simulations, the lesion fraction was varied from 0 to 0.45 in steps of 0.05. The repair cycle consisted of four trials during each of which a random cue of a single node was presented to the network. Each of the four patterns received such a minimal cue. Initial T was 0.2. The network was allowed 30 iterations to converge on a specific pattern. During this process the learning rate was set to 0. After convergence (if any), for 10 trials the learning parameter was set to 0.01. The results in Figure 3.7 show successful self-repair up to

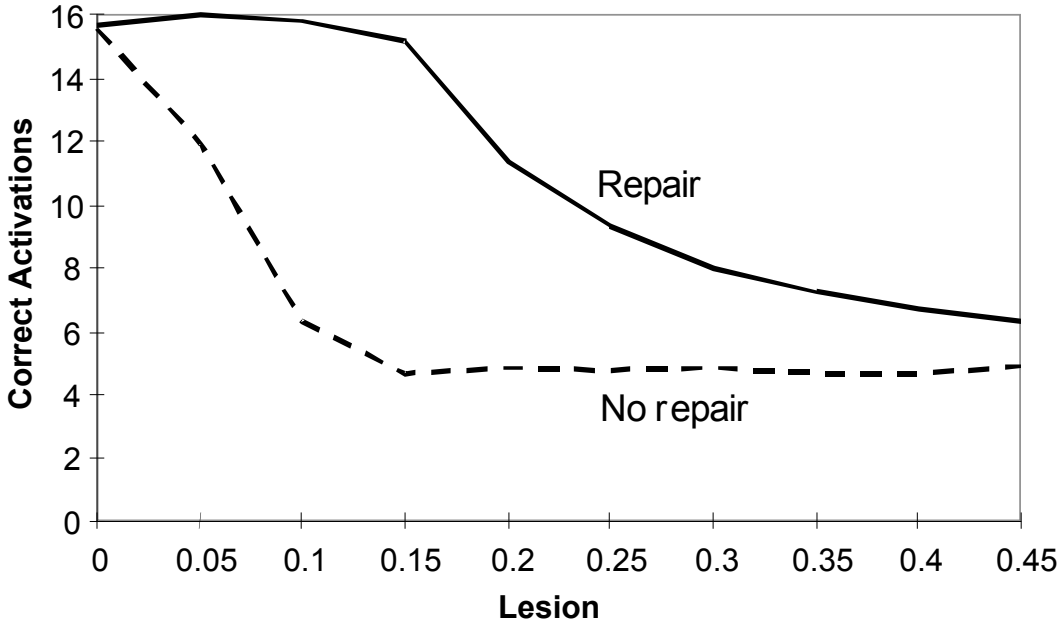


Figure 3.7. Results of a minimally cued soft k-winner-take-all network with continuous lesioning and selfrepair. The network was tested for every pattern by clamping a node that is part of a representation and let the network run/cycle until a stable attractor was reached. The number of activated nodes that are part of the target pattern is counted as correct activations. Each data point is the average of 100 replications of the entire simulation.

a lesion fraction of about 0.15. After 20 lesions of 15% without self-repair, we expect a very low residual connectivity of an estimated 2%, namely $0.50(1\text{-}0.15)^{20} \approx 0.0194$. This is far below the connectivity threshold of a graph consisting of 10 nodes (which is the size of the pattern representation). The control simulation in the figure without repair, indeed, confirms how the four patterns disintegrate completely after 20 lesions of 15%. With self-repair, however, they are still nearly completely intact.

Simulation 5. Long-term repair with a constrained learning rule. Learning rule (4) does not include any normalization. This implies that weights will increase or decrease without bound as long as learning continuous, as will be the case during repair cycles. We consider this in itself an undesirable feature, because of its clear biological implausibility. In addition, the unconstrained repair process was found to become prone to runaway processes after many repair cycles. In order to achieve long-term stability we, therefore, introduced a stop criterion for learning. If the sum of the absolute net input, summed over all neurons in the network, exceeded a preset threshold the learning rate was set to 0, otherwise, it was 0.01 as above. In these simulations, a repair cycle consisted of a single additional learning trial. In contrast to Simulation 4, cueing was fully random. Lesion fractions were varied from 0 to 0.005 in steps of 0.001. There were small differences with Simulation 4: The number of boosted learning trials was 8, initial value of T was 0.3, time allowed for convergence after repair cue was 40 iterations, time allowed for convergence after test cue was presented was 50 iterations. The repair threshold (summed absolute net input) was 100.

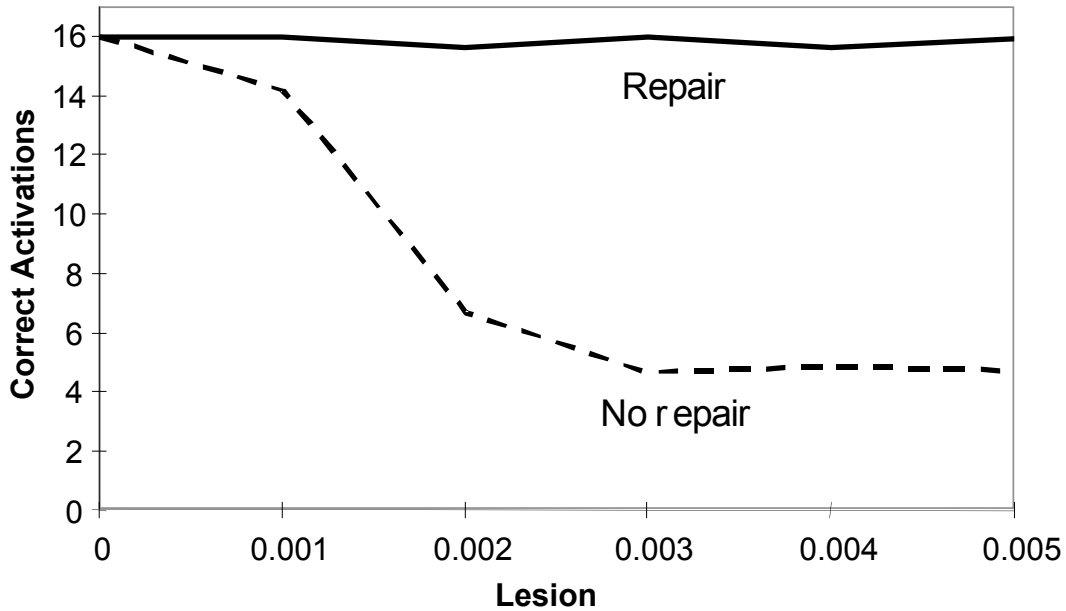


Figure 3.8. Randomly cued soft k-winner-take-all network with continuous lesioning and constrained selfrepair. Correct activations were calculated as in Figure 3.7. Each data point is the average of 10 replications of the entire simulation.

The results are shown in Figure 3.8. Even after a very long time, representations remain fully intact, while the non-repaired representations disintegrate completely with lesion fractions of 0.003 or higher. After 4000 lesion-repair cycles with a lesion size of 0.001 the process remained stable.

3.5 Discussion

This chapter was a first exploration of self-repair with mathematical and simulation models. We have shown that when memory representations are constantly lesioned diffusely, self-repair can extend their lifetime in Hopfield networks and in a soft *k*-winner-take-all network that has been used as the 'cortex' part of a model of long-term memory and amnesia. With this we have answered the main question of this chapter showing that self-repair in artificial neural networks can work. In the Hopfield model we demonstrated autonomous self-repair. Since this is a very difficult type of self-repair to model, this makes it more likely that we will be able to model other types of self-repair (Chapter One). Furthermore, for overlapping patterns we showed that perfect (autonoumous self-)repair can only be achieved by increasing the intensity of the self-repair process. This result together with the demonstration that self-repair can work in a more complex network, the TraceLink model, suggests that self-repair can work in neural networks of the brain, although they may be far more complex. Another result is that we identified an important problem for self-repair in connectionist systems: runaway repair. We will now discuss some issues concerning self-repair in the brain.

Although in the cerebral cortex and most other areas of the brain a lost neuron cannot be replaced, this does not imply that brain tissue has no possibilities for repair. If neurons themselves cannot be replaced, their dendrites, axons, and synapses still can grow longer and stronger (Bertoni-Freddari *et al.*, 1990; Bertoni-Freddari *et al.*, 1988; Buell & Coleman, 1979; DeKosky & Scheff, 1990) . The primary function of this is to counteract the effects of cumulative errors in the synapses caused by continuous neural and extra-neural noise. Because at present there is little direct evidence available, the existence and continuous operation of repair processes in the brain remains an empirically testable hypothesis.

A secondary function is to remedy, as far as possible, the effects of brain damage. A central assumption in a related paper by Robertson and Murre (I. H. Robertson & Murre, 1999) is that there exists a broad continuum from normal learning to recovery from brain-damage, and this continuum-hypothesis may also be extended to the underlying mechanisms of learning and repair. Thus, effects in long-term learning may well rely partially on some

form of neural sprouting, and various aspects of recovery from brain damage may, for example, involve synaptic strengthening of the type usually assumed in connectionist models.

A network lesion in our simulations usually consists in severing a percentage of the connections to the nodes in the area mentioned. One might object to this on the grounds that in the brain entire neurons (or neuron groups) may be lost. There are several reasons why we do not emphasize loss of entire neurons. Firstly, we are in fact already lesioning neurons: as a side effect of severing large numbers of connections, many neurons will become disconnected entirely, thus effectively silencing them permanently. Secondly, there is now strong evidence that long-term recovery is related to re-establishing of lost connections. Thirdly, the recovery of lost neurons by regaining their functionality is a trivial process as far as modeling is concerned. We do not expect that this process in itself can explain many of the aspects of recovery from brain damage, with three notable exceptions: (i) In many cases, most initial recovery will be the result of silenced neurons regaining functionality after a silent phase following immediately upon the lesion (so called *diaschisis* in neurology). (ii) Longer-term effects of recovering neurons may contribute in a nontrivial manner to the recovery from anterograde and retrograde amnesia, for example, after closed-head injury. The effects of this have been detailed in a separate paper (Meeter & Murre, 2005). (iii) Sometimes there is a non-trivial interaction between temporarily silenced neurons and the process of recovery as in the case of multimodule inhibition and excitation. In such cases, a large percentage of silenced neurons may cause an entire system to become inhibited by a contralateral inhibitory system, effectively preventing recovery of the non-silenced neurons. These simulations are summarized in Robertson and Murre (1999).

Appendix A: Activation rules of the TraceLink model

Activation rule

A node *i* has an activation a_i that can take on either of two values: 0 or 1. The probability that node *i* will 'fire' (i.e., that its activation becomes 1) increases with its net input, as follows:

$$p_i = \frac{1}{1 + \exp\left(-q^{-1} \left[\sum_{j=1}^{n} w_{ij} a_j - Inhibition \right] \right)}$$

where w_{ij} is the connection weight from node *j* to node *i*, a_j is the activation value of node *j*, and *n* is the number of nodes in the model (if there is no connection between *j* and *i*, w_{ij} is zero by default). Inhibition is discussed in the next paragraph. As in the Bolzmann Machine (Ackley, Hinton, and Sejnowski, 1985), the temperature parameter *q* controls the degree of randomness of the nodes. In addition to the *Inhibition* term, a difference with the Bolzmann machine is that we use synchronous activation update, rather than one-at-a-time of asynchronous activation updates.

Threshold control

The total number of activated nodes in a module (called *A*) is constantly monitored and firing thresholds are adjusted to ensure that this number does not wander too far from the target number *k*. The system achieves this by constantly adjusting two thresholds *T* and τ. *Inhibition* is the sum of the fast changing threshold T multiplied by the number of active nodes A, and the slow moving threshold τ: *Inhibition* = TA + τ. The control of fast inhibition, *T*, is straightforward: If the total activation at time t (A_t) is higher than *k*, *T* is increased (more inhibition), if A_t is lower it is decreased. In particular, if A_t is much larger than *k*, *T* is increased a lot; if A_t is only a bit larger, *T* is increased a little. If *A* is much larger or smaller (i.e., more than a *crit* proportion) than *k*:

$$if\ A_t > (1+crit)k$$

$$T = T + \Delta_t$$

$$if\ A_t < (1\text{-}crit)k$$

$$T = T - \Delta_t$$

else, if *A* is only slightly larger or smaller than *k*:

$$if\ A_t > k$$

$$T = T + 1/3\ \Delta_t$$

$$if\ A_t < k$$
$$T = T - 1/3\ \Delta_t$$

where *crit* is the criterion for deciding whether A_t is much larger or smaller, and Δ_t is the change made to T (*crit* = 0.20, and Δ_t = 0.01 works well for the simulations reported here). One disadvantage of this method is that *T* may change too quickly so that the module starts to oscillate violently. To prevent this, A_t is dampened by making it a moving average of the current activation and the activation of previous iterations. When A^*_t is the current level of activation, the value used to compute both the level of inhibition A_tT and the change in the parameter T is:

$$A_t = 0.5A_{t-1} + 0.5A^*_t$$

This precedes calculation of the new threshold *T*.

The slow inhibition process aims to keep the 'slow threshold' τ equal to *TA*. When the equilibrium is disturbed, for example, if the activation is diminished due to a lesion, τ slowly decreases to a new equilibrium value. The speed of this change is determined by the parameter Δ_τ. Because we envision the adjustment to be slow, Δ_τ is chosen low (0.001). The expression for calculating τ_{t+1} at t+1 is

$$\tau_{t+1} = (1-\Delta_\tau)\tau_t + \Delta_\tau TA$$

The amount of 'fast' inhibition is bounded by a minimum value T^{min} and a maximum value T^{max}. If $T < T^{min}$ it is set to T^{min}, and if $T > T^{max}$ it is set to T^{max}. Similarly, τ is also kept between upper and lower bounds: if $\tau < \tau^{min}$, τ is τ^{min}; if $\tau > \tau^{max}$, τ is τ^{max}. T^{min} and τ^{mi} were set to 0. T^{max} and τ^{max} were set to such high values that they were never reached in the simulations. Initially, τ was always set to 0.

4

Analysis of random cued self-repair in feedforward connectionist systems

Abstract

The main topic of this chapter is to investigate random cued self-repair, also called autonomous self-repair, with an analytical model. The model allows us to express memory retrieval in terms of probability providing information about system stability. System stability is expressed in the retrieval probabilities of the weakest and strongest memory representation. The first retrieval probability indicates the risk of a system of losing a memory representation. The second retrieval probability gives information about a possible runaway memory representation. We will derive results that can be applied to the more complex simulation models of the other chapters and the brain. The results concern research questions involving the effects on system stability (1) of weight differences due to learning alone (2) of weight differences because of learning and lesions together, (3) the activation probability, and (4) pattern size. We will show under which conditions the most difficult type of self-repair, autonomous self-repair, is possible. Since the brain fulfils the important condition of comprising many patterns, we argue that autonomous self-repair is feasible in the brain.

4.1 Introduction

Self-repair is the hypothesis that neural networks of the brain have a capacity of self-repair by maintaining redundancy. That is, we assume that networks of the brain have redundancy that is kept at some minimal level by learning processes, which we will describe shortly. It is inspired by the redundancy of artificial neural networks, where we postulate a repair mechanism based on cues activating the network and plasticity. Damage as well as self-repair modifies the network structure. Self-repair is effective if its modifications counteract changes caused by damage. With some types of self-repair this is not a trivial task but still feasible as we will see below.

Self-repair is a process that carries out the following algorithm over a certain time period: 1) a stimulus activates a neural network, 2) activity is allowed to spread over its nodes, 3) a learning rule updates the connections between the nodes. The first two steps determine which memories are selected. The self-repair method is mainly determined by the type of stimulus. In guided or supervised self-repair a stimulus strongly associated with a stored pattern is used, for instance, a stimulus used during the training phase or a prototype of the training stimuli. With guided self-repair we have control over which stored memory representation is repaired and the amount of time during which it is repaired. In autonomous self-repair, a randomly generated cue is used to select a stored memory representation. Since these stimuli are not associated with any stored memory representation, selection is a probability process. With this type of self-repair we neither have control over which memory representation is selected for repair nor over the amount of times they are repaired.

In this chapter, we will investigate the effect autonomous of self-repair and lesions on system stability that is expressed in terms of retrieval probability. In particular, we will investigate the effect on system stability of (1) weight differences due to learning alone, (2) weight differences due to learning and lesions, (3) of stimulus intensity, and (4) size of memory representations. Learning and the stimulus intensity are parameters of self-repair. In this research, it is assumed that learning can only increase weights as is the case in the original Hebb rule (Hebb, 1949; Koch, 1999). The way how stimulus intensity is operationalized in this model will be explained below. We will try to derive results that can be applied to the more complex simulation models of this thesis and the brain.

Autonomous self-repair, if uncontrolled, suffers from ‘runaway repair’. A similar, more familiar example of a runaway effect in the literature is runaway consolidation. Memory consolidation is the post-processing of memory traces, during which the traces may be

reactivated, analyzed and gradually incorporated into the brain's long-term memory (Maquet, 2001). Several authors have modeled consolidation as a process in which the 'to be strengthened' memories are selected randomly (Alvarez & Squire, 1994; Meeter & Murre, 2005; Murre, 1996). It is known that with this type of selection, memory consolidation is subject to runaway effects (Hemmen, 1997; Horn et al., 1998b; Meeter, 2003). In case of runaway, one memory representation becomes much stronger than the other memory representations, gradually overtaking all resources.

Runaway processes in self-organizing systems are due to competition over resources (Malsburg, 1995). In neural networks undergoing learning, the competition over resources is between the different memory representations. The weights of a memory representation are an important factor for the competitive strength of memory representations in neural networks. Weak memory representations have low valued weights compared to strong memory representations. During consolidation and autonomous self-repair, memory representations will engage in competition over activation and weight updating. The memory representation winning the activation will receive a weight update, which increases its weights and thus enhances its competitive strength. The runaway effect can be caused by an increasing weight difference between the 'runaway' memory representation and the other memory representations. Since this is a self-reinforcing process with disastrous results (wiping all memories but one), we speak of 'runaway'.

The above exposition suggests that in case of autonomous self-repair, where stored memory representation are randomly selected, the most stable memory system is a system with memory representations that are of equal strength. It, furthermore, suggests that instability in these memory systems arises because of large weight differences between memory representations that may arise with learning and damage, which may both alter the competitive strength of a memory representation. Thus, for stability in neural network systems, weight differences seem very important and have to be kept as small as possible, unless some other stabilizing mechanism is operating. To investigate the effect of weight differences on system stability, we will study an associative memory model analytically by expressing retrieval in terms of probabilities. For instance, by definition a weak pattern has a small retrieval probability. In particular we will investigate the retrieval probability of the weakest and strongest pattern. The weakest pattern is interesting as a measure, because as long as retrieval probability is positive, it can in principle be retrieved and the system has not yet lost any memory representations. The retrieval probability of the strongest pattern is

interesting, because the higher its retrieval probability, the more likely it will turn into a runaway memory representation.

The associative memory model in which we will carry out investigations has two layers. One layer is the input layer for the other layer. They are connected in a feedforward fashion, where each group of nodes of the input layer has excitatory connections with one group of nodes of the output layer and inhibitory connections with all other nodes of the output layer. A layer may represent a neural area such as for instance the primary somato-sensory cortex, where a group of nodes or a neural assembly corresponds to a memory representation of a finger. The first layer of the model represents an input layer in which random activations occur to model autonomous self-repair, that is, self-repair with random stimuli. Given the Bernoulli process of the first layer, we derive equations that allow us to derive the probabilities of different activation configurations for the second layer. The two-layered model will be described in more detail in Section Two.

In Section Three, we will investigate the effect of weight differences on the retrieval probabilities of the weakest and strongest pattern due to learning alone and due to learning and lesioning together. It will be shown that learning together with lesioning is more detrimental to system stability than learning alone. We will, furthermore, investigate the stimulus intensity of the first layer by varying the activation probability p of the input layer. This activation probability determines for each input node its probability to be activated or fired. This probability is assumed equal for each node. It can be regarded as a random stimulus, since no neurons of a particular neural representation have a specific (higher) probability to be activated. We will show for this activation parameter that it has to remain low to maintain a stable system. Finally, we will investigate the effect of different pattern sizes on system stability. We will show that pattern size can compensate for small weights.

In addition to the above results, another result will be that changes in weight or activation probability have a non-linear effect on retrieval probability. In Section 4, we will discuss implications of the results for autonomous self-repair in connectionist systems and in the brain. Amongst others, we will argue that though we investigate a system comprising two memory representations, the results can be applied to systems having more than two memory representations.

4.2 The associative memory model

Runaway processes can occur in several associative memory models. In this paper, we will analyze an associative memory model with spike coding. The model neuron is active for one time step, after which it is deactivated. The model consists of an input-layer and an output layer of equal size. We start with a Bernoulli lattice process in the input layer, from which we will derive the expressions of one-step retrieval for the output layer (Palm & Sommer, 1996). In one-step retrieval, the output pattern is evaluated from the input pattern after one synchronous parallel calculation of all neurons as opposed to fixed-point retrieval in which the system is iterated until a stationary state. The retrieval criterion is that there is some activation in the to -be- retrieved memory representation and no activation in other patterns. We assume that both input and memory representations are non-overlapping. Consequently, in case of Hebbian learning only weight strengthening of a stored memory representation may take place and not weight strengthening between different memory representations.

From a neural point of view, a Bernoulli process can be thought of as the result of independent activations and potentials from some external (sensory) source, which activate neurons in a layer when a potential exceeds some threshold. The input-layer consists of a finite number of neurons, and we denote by $A_{1,i}$ a Bernoulli random variable that describes the activation or de-activation of neuron i in the input layer, which we mark by the index '1'. The probability distribution of layer activation is completely described by the probabilities of activation $p = \mathbf{P}\{A_{1,i} = 1\}$, which we assume to be identical for all neurons i. From the independence assumptions of the external process formulated above it follows that activations at different neurons are independent.

We now construct a second process in the output layer that is induced by the Bernoulli process of the input layer. In principle, the activation level $A_{2,j}$ of neuron j in the output layer (denoted by the index '2') is determined by the activation of all the neurons in the input layer and the "weights" w_{ji} that connect the neurons i of the input layer with neuron j of the output layer. We assume an associative memory model with orthogonal memory representations in which there are only positive connections from a memory representation of the input layer to its corresponding memory representation of the output layer. The connections of that memory representation to the other memory representations are negative. For matter of convenience and notation, memory representations will henceforth be addressed as patterns.

The random variable that describes the membrane potential $U_{2,j}$ of neuron j in the output layer, belonging to some pattern P^+ can be written as

$$U_{2,j} = \sum_{i \in P^+} A_{1,i} w_{ji} - \sum_{k \in P^-} A_{1,k} v_{jk}, \tag{5}$$

where the first sum term represents the contribution of activated neurons of the corresponding pattern P^+ in the input layer and the second term denotes the inhibition from activated neurons of the set of all other patterns P^- of the input layer. We use the same expression for the potential of neurons from P^- in output layer or layer 2. Neuron j of the output layer will be activated only when its membrane potential $U_{2,j}$ exceeds a threshold θ. We therefore introduce the activation $A_{2,j}$ of neuron j, which is a Bernoulli random variable that is defined in terms of $U_{2,j}$ as the indicator function

$$A_{2,j} = \begin{cases} 1, \text{ if } U_{2,j} \geq \theta \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

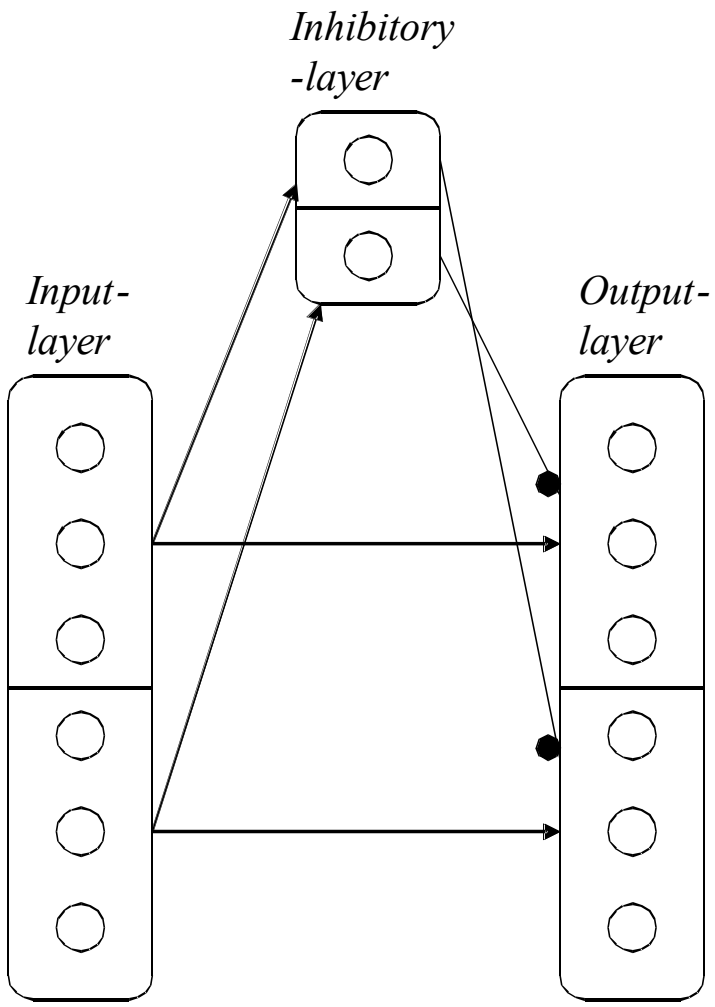


Figure 4.1. Schematic view of the associative memory model (see text).

In the sequel, we will analyze a simplified model version. We assume the excitatory weights between the neurons to be the same within each pattern. We also assume all the inhibitory weights to be equal to v. A possible neural interpretation of v is that weights w_{ji} are connected to interneurons, which in turn are connected to the output layer (Figure 4.1). Connections between layers are such that an auto-associative network is created, that is, there is global inhibition in which all neurons have an inhibitory effect on all neurons of the output layer, except to output-layer neurons of their own memory representation. Under the assumption that the feedforward inhibition (Wierenga & Wadman, 2003) is (much) faster than excitation, both excitation and inhibition of the input-layer will simultaneously affect neurons of the output-layer.

In the next section, we will derive expressions for the retrieval probability of the strongest and weakest patterns under the assumption of global inhibition present in associative memory (Amari, 1990). We will investigate the effect of an increasing weight difference on the retrieval probability of a pattern. We will mainly treat orthogonal patterns to avoid the problem of interference as was discussed by Hasselmo (Hasselmo, 1994).

4.3 Retrieval under learning and lesioning

In order to study system stability we will investigate the behavior of retrieval probabilities of stored patterns when some are reinforced or learned, while others undergo lesions. In Section 4.3.1 the concept of retrieval probability will be introduced and explained in detail. In Section 4.3.2 we will study the effect of weight differences on system stability. In this section, we will show that learning together with lesioning has a much larger (negative) effect on system stability than learning alone. In section 4.3.3 the influence of the activation probability p of a node in the input layer is investigated. It will be shown that low values for this activation probability are best for system stability. In section 4.3.4 we will consider effects of different pattern sizes on system stability.

4.3.1 Introduction: Retrieval probability

The retrieval probability of a pattern is defined here as the probability of activation of at least one neuron in that pattern, while all neurons in all other patterns are deactivated. Such a configuration of activated and deactivated neurons over the output-layer is called an event. This is a subset of all possible events, which are all possible combinations of activated and deactivated neurons over the layer. For instance, the event that three neurons are simultaneously activated in two patterns does not belong to the set of events of correct

retrieval probability. We will start our investigations with a model comprising two orthogonal patterns P_1 and P_2 with excitatory weights w_1 and w_2, respectively. We denote P_1 as the 'weak' pattern, that is, the pattern that undergoes lesions, and P_2 as the 'strong' pattern, which is the pattern that will be subjected to learning. Since we have two patterns, the weak pattern is the weakest pattern and the strong pattern the strongest of the system.

To further explain retrieval probability, we will illustrate it with the retrieval probability of the weak pattern, which we write as $\mathbf{P}(W_a \cap S_d)$. The event W_a denotes activation or retrieval of the weak pattern, while S_d is the event that the strong pattern is deactivated or not retrieved in the output layer. Activation and deactivation will be used in the sense as described above, that is, activation of at least one neuron in the weak pattern and deactivation of all neurons in the strong pattern. The retrieval probability of the weak pattern is equal to the following expression, which is derived in Appendix A:

$$\mathbf{P}(W_a \cap S_d) = \sum_{k_1=0}^{|P_1|} \sum_{k_2=0}^{|P_2|} 1_{(w_2k_2 - vk_1,\, w_1k_1 - vk_2]}(\theta) \mathbf{P}\Big(\sum_{i \in P_1} A_{1,i} = k_1\Big) \mathbf{P}\Big(\sum_{j \in P_2} A_{1,j} = k_2\Big). \tag{7}$$

The notation |.| denotes the total number of neurons in a pattern. The indicator function in (7) is equal to one if $w_2k_2 - vk_1 < \theta \leq w_1k_1 - vk_2$, and is equal to zero otherwise. Expression (7) says that the retrieval probability of the weak pattern results from all the contributions of activated neurons k_1 and k_2 in layer 1, such that the potential $w_1k_1 - vk_2$ of the weak pattern P_1 is at least equal to the threshold θ and the potential $w_2k_2 - vk_1$ of the strong pattern P_2 is smaller than θ.

The retrieval probability of the strong pattern**,** $\mathbf{P}(W_d \cap S_a)$, has the same form as (7). The only difference is that the indices of the patterns are interchanged. In the next sections, we will derive expressions of retrieval probability for the weak and strong pattern under learning and lesioning.

4.3.2 Weight differences: the effect of learning and lesioning on system stability

In this section, we will investigate the difference between the effect on retrieval probability of learning alone and learning and lesioning together. In order to do this, first in Section 4.3.2.1 the effect of learning alone on retrieval of the weaker and stronger pattern is investigated. This will show that retrieval probability of the weaker pattern can remain non-zero in an

evolving system. It, furthermore, shows that the retrieval of the stronger pattern can only grow to some maximum that can be far from one. Then in Section 4.3.2.2, the effect of learning and lesioning on the retrieval probability is investigated. We will depart from the point where we left in Section 4.3.2.2, that is, a state in which the weak pattern still has intact weights and the strong pattern has much bigger weights due to learning. Departing from this state, we will lesion the weights of the weaker pattern and calculate the retrieval probabilities of the weaker and stronger pattern. We will show that in this case the retrieval probability of the weaker pattern can become zero and the retrieval probability of the stronger pattern can grow very close to one. In Section 4.3.2.3, we will combine the retrieval probabilities of the weakest and strongest pattern in a measure of system stability. This measure elucidates the difference between learning alone and learning and lesioning together.

4.3.2.1 Weight differences: the effect of learning on retrieval

In this section we will study the effect of learning of the strong pattern on retrieval. We simulate learning by increasing the weights of the stronger pattern and keeping the weights of the weak pattern constant, in other words, w_2 increases while w_1 is fixed. We will start by analyzing the retrieval probability of the weak pattern. The point of departure is the situation where the excitatory weights w_1 and w_2 of the two patterns are equal, say, to w.

If the inhibitory weight v and the threshold θ are equal to w_1, then expression (7) simplifies to

$$\mathbf{P}(W_a \cap S_d) = \sum_{k_1=1}^{|P_1|} \sum_{k_2=0}^{k_1-1} \mathbf{P}(\sum_{i \in P_1} A_{1,i} = k_1)\mathbf{P}(\sum_{j \in P_2} A_{1,j} = k_2). \qquad (8)$$

The behavior of (8) is illustrated in Figure 4.2a as the upper curve, which is plotted as a function of the activation probability p. Notice that, in order for retrieval probability (8) to be positive, the number of activated neurons k_1 of weak pattern P_1 in the input layer must, of course, be greater than zero, while the number of activated neurons k_2 of the strong pattern must remain within an upper bound (as its potential must be smaller than the threshold).

As w_2 increases, we eventually obtain the behavior shown by the lowest curve in Figure 4.2a. This retrieval probability can be derived from (7) as follows. Remember that the indicator function in (7) implies the inequality $w_2k_2 - vk_1 < \theta \leq w_1k_1 - vk_2$. If w_2 increases,

such that $w_2 \geq \theta + v|P_1|$, then the inequality $w_2k_2 - vk_1 < \theta$ only holds for $k_2 = 0$, under which it is satisfied for all k_1 (assuming that $\theta > 0$).

The inequality $w_2k_2 - vk_1 < \theta \leq w_1k_1 - vk_2$ now simplifies to $\theta \leq w_1k_1$, so that the retrieval probability of the weak pattern P_1 decreases to

$$\mathbf{P}(\sum_{j\in P_2} A_{1,j} = 0) \sum_{k_1=\lceil \frac{\theta}{w_1} \rceil}^{|P_1|} \mathbf{P}(\sum_{i\in P_1} A_{1,i} = k_1)$$

$$= (1-p)^{|P_2|} \sum_{k_1=\lceil \frac{\theta}{w_1} \rceil}^{|P_1|} \mathbf{P}(\sum_{i\in P_1} A_{1,i} = k_1), \qquad (9)$$

which follows from the binomial distributions of the summed activations in layer 1. Expression (9) is positive if, and only if, the activation probability p in the first layer is positive and smaller than 1, and $\lceil \theta/w_1 \rceil \leq |P_1|$. Under these conditions, the retrieval probability of the weak pattern will not become equal to zero, even when the excitatory weights of the strong pattern become infinitely large. This is the case, since there is a positive probability for the strong pattern P_2 of not receiving any activation in the first layer and for potentials of the weak pattern to exceed the threshold in layer 2. The retrieval probability of the weak pattern can thus be significantly greater than zero independent of size of the weights of the stronger pattern P_2. We have calculated some values numerically, the results of which are shown as the two curves between the upper and lower ones in Figure 4.2a. The behavior of retrieval probability shows interesting properties as learning of the strong pattern continues. Figure 4.2a indicates the existence of an optimal activation probability p under which the retrieval probability of the weak pattern is maximized.

We will continue the investigation by analyzing the effect of learning on retrieval probability $\mathbf{P}(W_d \cap S_a)$ of the strong pattern. As mentioned before, this probability has the same form as (7), in which only the indices of the patterns are interchanged. In order to illustrate the effect of learning on the retrieval probability we use the same parameter values as in Figure 4.2a.

Retrieval probabilities of the strong pattern are shown in Figure 4.2b for the same set of weights as in Figure 4.2a. The lowest curve in the figure is the same as the upper curve in

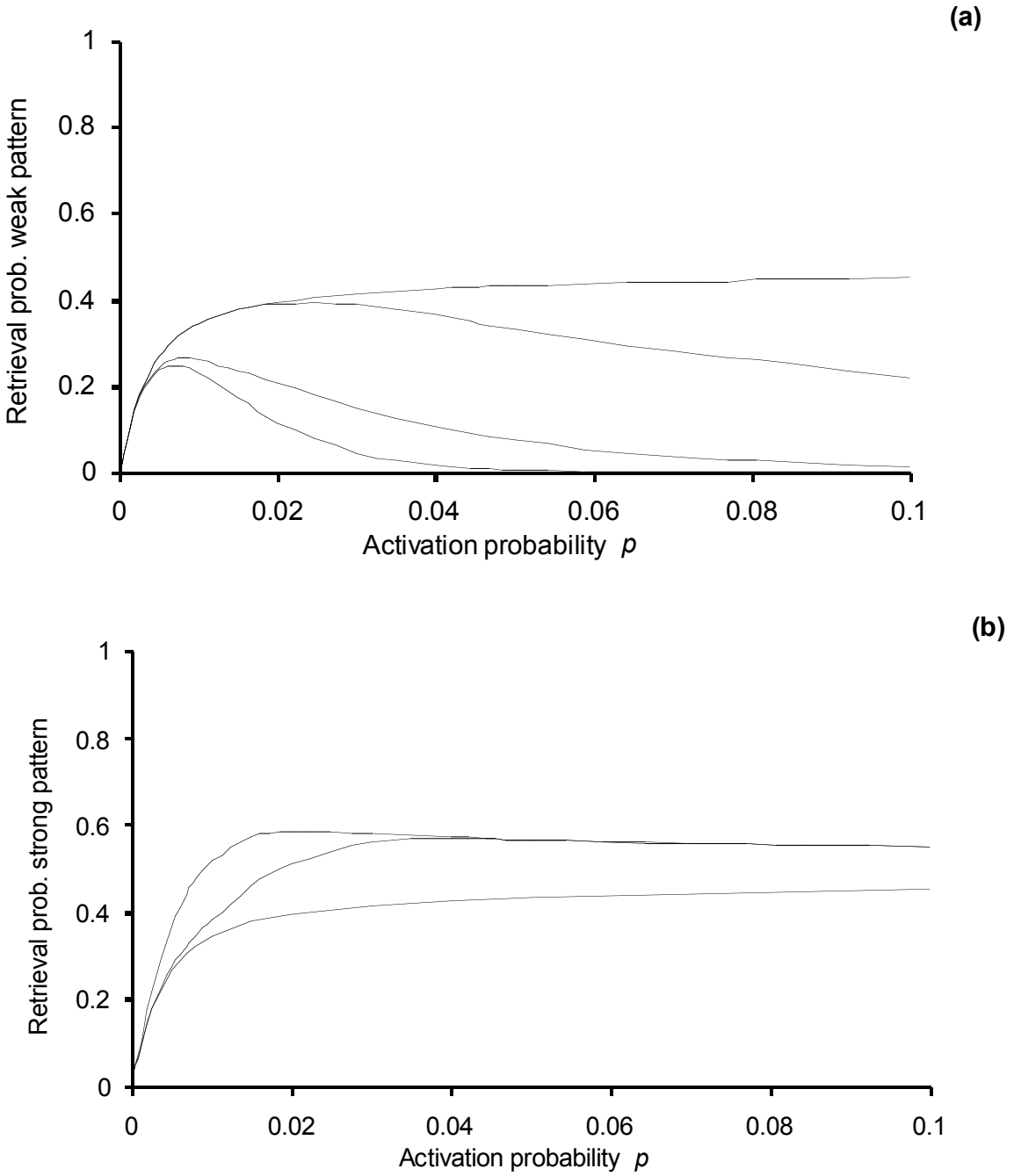


Figure 4.2. Effect of an increasing weight w_2 of the stronger pattern P_2 and a varying activation probability p (x-axis) on the retrieval probability of: (a) the weaker pattern P_1, for $|P_1| = |P_2| = 100$, $\theta = v = 0.1$. The four lines correspond with weights $w_2 = w_1 = 0.1$, $w_2 = 0.15$, $w_2 = 0.3$, $w_2 \geq 0.5$, from top to bottom line, respectively; (b) the stronger pattern P_2. All the parameter values are the same as in (a), except that the lines correspond with increasing weights w_2 from lowest to upper line.

Figure 4.2a, which represents the case $w_1 = w_2$. The retrieval probability increases as w_2 increases, reaching its maximum at

$$\mathbf{P}(W_d \cap S_a) = \sum_{k_2=1}^{|P_2|} \sum_{k_1=0}^{\left\lceil \frac{v}{w_1} k_2 + \frac{\theta}{w_1} \right\rceil - 1} \mathbf{P}(\sum_{i \in P_1} A_{1,i} = k_1) \mathbf{P}(\sum_{j \in P_2} A_{1,j} = k_2). \tag{10}$$

This expression simplifies for Figure 4.2b, since we set $v = \theta = w_1$:

$$\mathbf{P}(W_d \cap S_a) = \sum_{k_2=1}^{|P_2|} \sum_{k_1=0}^{k_2} \mathbf{P}(\sum_{i \in P_1} A_{1,i} = k_1) \mathbf{P}(\sum_{j \in P_2} A_{1,j} = k_2). \tag{11}$$

It can be shown that this expression stays the same for all $w_2 \geq 2\theta$.

4.3.2.2 Weight differences: lesioning after learning

In this section we will study the extent to which the retrieval of the weak pattern will be affected further under lesions of varying sizes, by decreasing its weight w_1. We will later do the same with the retrieval of the strong pattern.

In order to study the effect of a decreasing weight w_1 on the retrieval probability of the weak pattern, we study the behavior of (9). This probability is shown in Figure 4.3a (upper curve) and in Figure 4.2a (lowest curve) for equal pattern sizes and θ=w_1. There exists an optimum activation probability p (x-axis), where the retrieval probability is maximal (y-axis). The optimum activation probability can be calculated exactly, along with its maximal retrieval probability as we will show in Section 4.3.3, where we will investigate the influence of the activation probability p on retrieval probability. For now, the maximal retrieval probability can be made out from Figure 4.3a, where it can be seen that each curve has its own maximum.

We can furthermore see in Figure 4.3a that the retrieval probability of the weak pattern decreases rapidly as w_1 decreases. The middle curve in Figure 4.3a shows the retrieval probability for $\lceil \theta/w_1 \rceil = 2$. It easily follows from the binomial distribution for the number of activated neurons in the input layer that this probability decreases with respect to the highest curve by the amount $|P_1| p(1-p)^{2|P_1|-1}$. In the optimum of the upper curve (where p is about 0.0069), the retrieval probability decreases from ¼ to about 0.174.

The smallest value of w_1 under which retrieval probability (9) is still greater than zero is equal to $w_1 = \theta/|P_1|$. In this case we have that

$$\mathbf{P}(W_a \cap S_d) = p^{|P_1|}(1-p)^{|P_2|}. \tag{12}$$

Probability (12) is not plotted in Figure 4.3a since it is very close to zero. The figure shows the retrieval probability for $\lceil \theta/w_1 \rceil = 3$ instead. For $w_1 < \theta/|P_1|$ the retrieval probability of the weak pattern becomes zero: even with all neurons activated, w_1 is not strong enough for the potential of P_1 to exceed the threshold θ. This implies that the retrieval probability of the weak pattern can become zero in case of lesions after learning.

We will now analyze retrieval probability (10) of the strong pattern, which arose under increasing w_2 (Section 4.3.2.1). We will do this along the same line as for the weak pattern, that is, starting with the case $w_1 = \theta$ and letting w_1 decrease to $\theta/|P_1|$. The case $w_1 = \theta$ was already treated in the preceeding section, which resulted in expression (11). This retrieval probability was already shown in Figure 4.2b and is also shown in Figure 4.3b (lowest curve). The retrieval probability $\mathbf{P}(W_d \cap S_a)$ increases as w_1 decreases. As w_1 decreases to $\theta/|P_1|$, it follows that (10) increases to

$$\mathbf{P}(W_d \cap S_a) = 1-(1-p)^{|P_2|}, \tag{13}$$

when v and $\theta > 0$, which stays the same as w_1 decreases further to zero. Notice that this represents the most favorable situation for the strong pattern: as the weights w_1 and w_2 grow further apart, the strong pattern will eventually always be retrieved, except for the situation where this pattern is not activated at all in the input layer.

To summarize, this section shows that when the weights of the weaker pattern are decreased, after learning of the stronger pattern, the retrieval probability of the weak pattern can become zero. This does not only hold in the trivial case when the weights of the weaker pattern are zero. Expression (9) shows that the retrieval probability of the stronger pattern can grow to values close to one. In the following section, we will give a mathematically precise formalization and quantification of the notion of system stability under learning and lesioning.

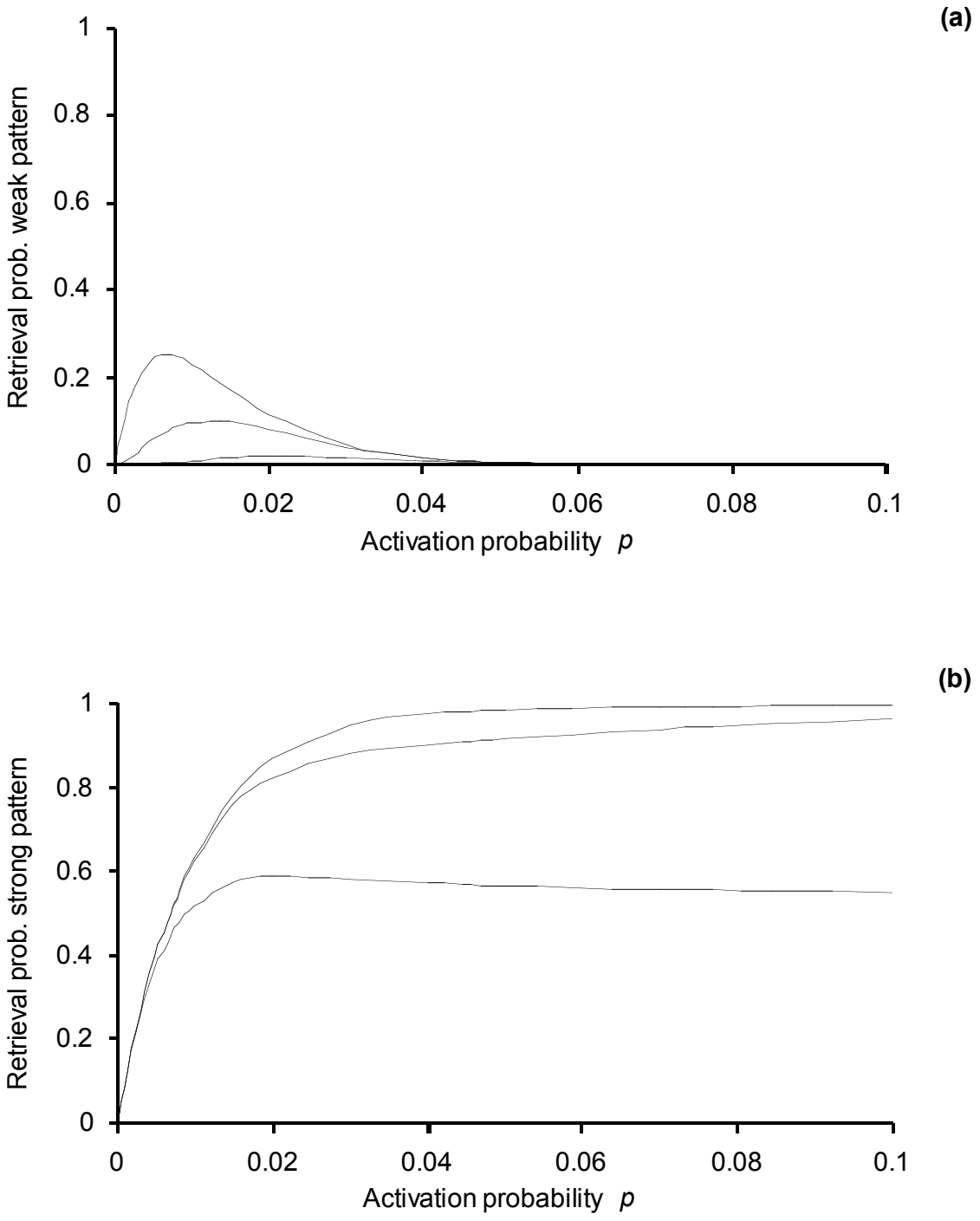


Figure 4.3. Effect of a decreasing weight w_1 of the weaker pattern P_1 and a varying activation probability p (x-axis) on the retrieval probability of: (a) the weaker pattern P_1, for $|P_1| = |P_2| = 100$, $\theta = v = 0.1$, given that $w_2 \geq 0.5$ for pattern P_2. The three lines correspond with weights $w_1 = 0.1$, 0.05, and 0.035, from top to bottom line, respectively; (b) the stronger pattern P_2. All the parameter values are the same as in (a), except that the lines correspond with decreasing weights w_1 from lowest to upper line.

4.3.2.3 A synthesis of learning and lesioning: system stability

We will combine the results of the two previous sections in order to derive implications for system stability. We will do this for learning alone and for lesioning after learning. For both cases, we derive and analyze retrieval probabilities of two events, in which either the weakest pattern or the strongest pattern is activated, and not both simultaneously. The two probabilities can be used to define system stability as the conditional probability that the weakest pattern will be activated, given that exactly one of the two patterns is activated. We thus have the following measure of system stability:

$$\mathbf{P}(\{W_a \cap S_d\} \mid \{W_a \cap S_d\} \cup \{W_d \cap S_a\}) = \frac{\mathbf{P}(W_a \cap S_d)}{\mathbf{P}(W_a \cap S_d) + \mathbf{P}(W_d \cap S_a)}. \tag{14}$$

Since we consider a system with two patterns, we could also choose the activation of the strongest pattern as event in the numerator of (14). However, this results in the complementary probability of (14), which yields the same conclusions. This measure takes into account the distance between the retrieval probabilities of the weakest pattern and strongest pattern. It filters out all probabilities of all other events. In this case with two patterns, it filters out the events that none of the two patterns are activated or that both are activated. With multiple patterns it will filter out all events involving all other patterns.

The right-hand side of (14) shows that we can derive system stability for both learning and lesioning directly by substituting the two probabilities $\mathbf{P}(W_a \cap S_d)$ and $\mathbf{P}(W_d \cap S_a)$ in (14) . In the case of learning, this means that the two retrieval probabilities shown in Figure 4.2a-b can be used to derive conditional probability (14). Of course we have to combine the two probabilities such that the weight w_2 has the same values. The results are shown in Figure 4.4a. In the case where the weights w_1 and w_2 are equal, it follows that probability (14) is equal to ½ for every $p > 0$. As w_2 increases, it will be clear that (14) decreases for every $1 > p > 0$. The results show that system stability deteriorates rapidly as w_2 increases, unless the activation probability p is close to zero. For instance, conditional probability (14) is greater than 0.3 when $0 < p < 0.01$ in Figure 4.4a.

We can also apply expression (14) in the case of lesioning of the weakest pattern. The results shown in Figure 4.3a-b can be combined in the same way as above, which leads to the behavior shown in Figure 4.4b. This figure also shows that system stability deteriorates rapidly when w_1 decreases. For example, conditional probability (14) reaches a maximum of about 0.02 when $\theta/w_1 > 2$.

To conclude, this measure shows clearly similar results as the other measures of the previous sections: with lesioning the weakest pattern, after learning of the stronger pattern, the measure of system stability decreases rapidly. Another remarkable result is that in case of lesioning after learning, stability expression (10) goes to zero as activation probabilities p tends to zero. We will discuss this in the next section.

4.3.3 The influence of the activation probability *p* on system stability

Figure 4.2 and Figure 4.3 show that retrieval probability of the strong pattern is less sensitive to change in weights than the weak pattern. To investigate the effect of p on system stability we study its effect on the retrieval probability of the weaker pattern. We discuss the influence of activation probability p on retrieval probability in the same order as we discussed the effect of weight differences on retrieval probability. We start with a system in which only the stronger pattern is reinforced by learning. After this we analyze a system in which the weak pattern undergoes lesions and the strong pattern already has large weights.

To investigate the influence of the activation probability p on the retrieval probability of the weak pattern in case only the weights of the stronger pattern are reinforced due to learning we analyze the behavior of equation (9). The behavior of this equation is different from the initial condition with equal excitatory weights, both in the values of the retrieval probability and in the sensitivity of this probability with respect to the activation probability p. The latter refers to the rapid decline of the retrieval probability after reaching its maximum value. Small deviations around the optimal value for p will lead to a rapid decrease of the retrieval probability. This implies that the weak pattern should be activated or cued in a very precise way, with small p, in order to have a significant probability of being retrieved, thus limiting the risk of being lost rapidly. As can be noted in Figure 4.2a, the retrieval probability in the case $w_1 = w_2$ is less sensitive to variations in p. This suggests that sensitivity to p plays an important role in system stability.

The retrieval probability (5) of the weakest pattern reaches its largest values when it is equal to

$$\mathbf{P}(W_a \cap S_d) = (1-p)^{|P_2|} \sum_{k_1=1}^{|P_1|} \mathbf{P}\Big(\sum_{i \in P_1} A_{1,i} = k_1\Big) = (1-p)^{|P_2|} \left\{1 - (1-p)^{|P_1|}\right\}, \qquad (15)$$

since $\theta > 0$, so that $\lceil \theta / w_1 \rceil \geq 1$ in (5) for all nonnegative w_1. One can furthermore derive an equation to calculate the optimal activation probability p, such that (15) is maximized. Retrieval probability (15) has a maximum at

$$p = 1 - \left(\frac{|P_2|}{|P_1| + |P_2|} \right)^{\frac{1}{|P_1|}}. \tag{16}$$

The corresponding retrieval probability can also be expressed in terms of pattern sizes and is equal to

$$\frac{|P_1|}{|P_1| + |P_2|} \left(\frac{|P_2|}{|P_1| + |P_2|} \right)^{\frac{|P_2|}{|P_1|}}. \tag{17}$$

When the pattern sizes are equal, the maximal retrieval probability is ¼.

To investigate the influence of the activation probability p on the retrieval probability of the strong pattern after reinforcement due to learning, we analyze the behavior of equation (10) and (11). A thorough mathematical analysis of (10) and (11) with regard to their behavior in p is hard to accomplish, but it is possible to gain insight into certain properties. First, it is easily verified that both expressions go to zero as p goes to zero. Second, from Figure 4.2b it can be seen that retrieval probability (11) increases rapidly as p increases. It can be proven that the retrieval probability settles at values around ½ when p approaches ½, which can be explained from the distribution of the bivariate probability mass for the number of activated neurons k_1 and k_2 for the two patterns on the subset $1 \leq k_2 \leq |P_2|$, $0 \leq k_1 \leq k_2$ in the input layer.[1]

Expression (14) for system stability gives a value greater than 0.3 in the lowest curve in Figure 4.4a. This means that the weak pattern will be retrieved about three times and the strong pattern seven times, on average, in ten retrievals. The weak pattern, therefore, still has a significant probability of being retrieved. Figure 4.4b shows that system stability deteriorates rapidly when the weakest pattern is lesioned. The results also show that system stability is very sensitive to variations in p under learning and lesioning.

[1] The behavior of the retrieval probability when p goes to 1 depends on the pattern sizes $|P_1|$ and $|P_2|$. It can be proven that $\mathbf{P}(W_d \cap S_a)$ either goes to 0 or to 1. Since we only focus on small values of the activation probability p, we will not analyze this case in this paper.

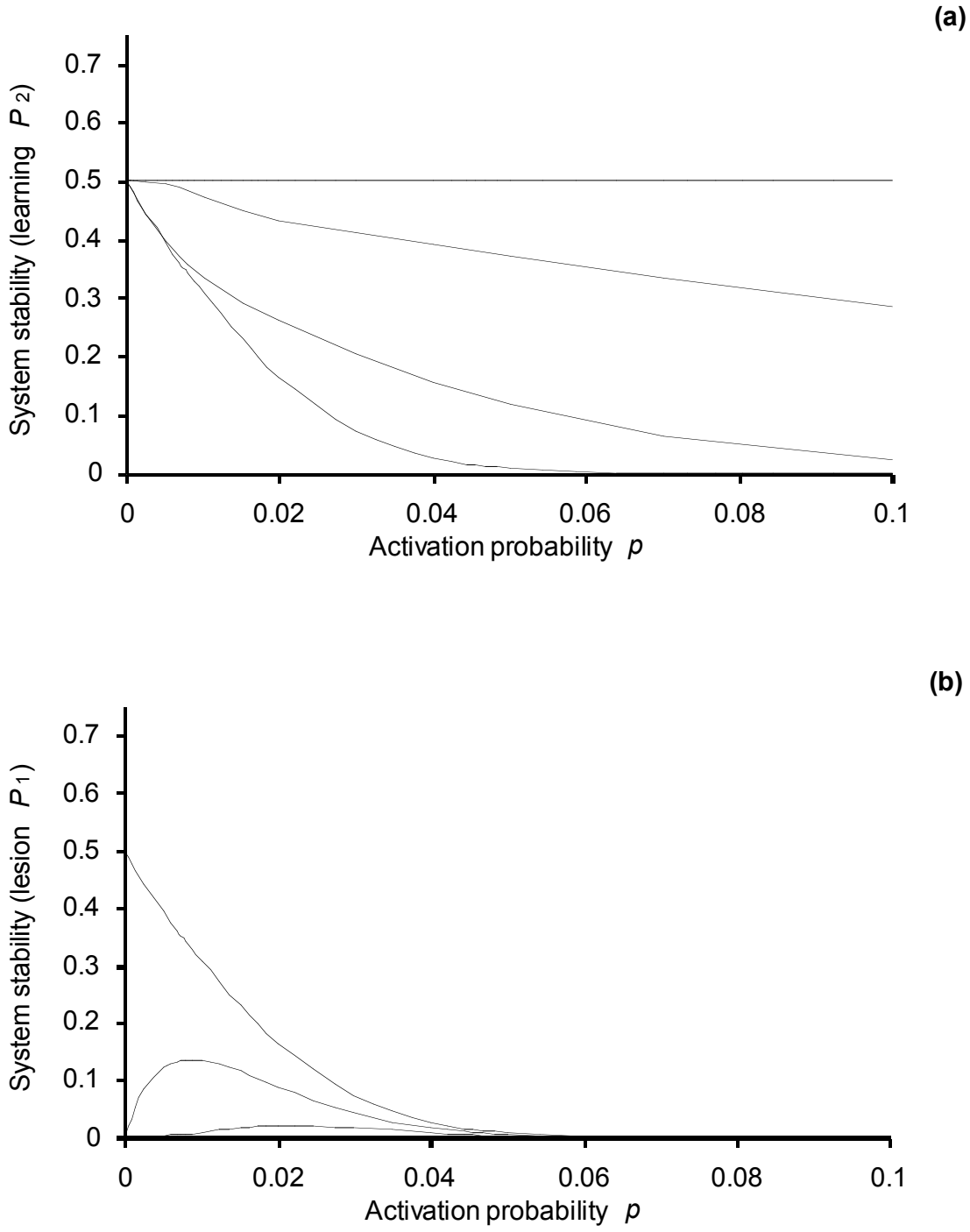


Figure 4.4. Behavior of system stability expression (14) for: (a) learning of the strongest pattern. The four curves correspond with weights $w_2 = w_1 = 0.1$, $w_2 = 0.15$, $w_2 = 0.3$, $w_2 \geq 0.5$, from top to bottom, respectively; (b) lesioning of the weakest pattern. The three curves correspond with weights $w_1 = 0.1$, 0.05, and 0.035, from top to bottom, respectively. The parameter values are the same as in Figures 2 and 3.

Another characteristic of system stability, which was also mentioned in Section 3.4.2.3, is the behavior of (14) as p goes to zero, when θ/w_1 increases, in case of lesioning after learning. In the example of Figure 4.4b, this conditional probability goes to ½ when $\theta/w_1 \leq 1$, but goes to zero when $\theta/w_1 > 1$. A decrease of p will improve system stability in the case of learning of the strong pattern, but this does not necessarily hold when the weak pattern is lesioned afterwards. Figure 4.4b shows that system stability expression (14) reaches a maximum for some value of $p > 0$ when $\theta/w_1 > 1$.

The reason for this behavior lies in the conditions for the number of activated neurons in the input layer under which the weak and strong pattern will be activated in the output layer. Activation of one neuron of the strong pattern in the input layer may be sufficient for its activation in the output layer. If $\theta/w_1 \leq 1$, then this also holds for the weak pattern, so that the retrieval probabilities of both patterns are of the same order in p. This, however, does not hold when $\theta/w_1 > 1$, in which case the weak pattern will need more than one activated neuron in the input-layer in order to have a chance of becoming activated in the output layer. When p goes to zero, this implies that conditional probability (14) tends to zero. This behavior holds for all pattern sizes, threshold values, and inhibition weights.

Another conclusion that can be drawn from the measure of system stability of Section 4.3.2.3, the main conclusion of this section is that in case of lesioning after learning the system is more sensitive to p compared to a system with only learning. This can be observed from Figure 4.4, where for instance a p-value of 0.04 in the only learning case (Figure 4.4a) still has a relatively high retrieval probabilities compared to the same p value in the lesioning after learning case (Figure 4.4b). This means that in such a system (lesioning after learning) there are fewer p-values for which the retrieval probabilities are fairly high and the weak pattern can still be retrieved.

4.3.4 The effect of pattern size $|P|$ on system stability

In this section, we will investigate the effect of pattern size $|P|$ on system stability. The importance of the pattern size was already demonstrated in the previous section, where it was shown that the optimal activation probability p and the corresponding optimal retrieval probabilities can be calculated with pattern sizes only.

It is also interesting to investigate the effect of different pattern sizes on system stability when p goes to zero. In Section 4.3.2.3 we noted that system stability expression (14) goes to zero when the weak pattern is lesioned, such that $\theta/w_1 > 1$, irrespective of pattern size. We thus consider the situation where $w_1 \geq \theta$ and where w_2 increases, and let p go to zero. The

two retrieval probabilities in (14) are given by expressions (10) and (15). The limit of the resulting conditional probability can be found by L'Hopital's rule, which yields:

$$\lim_{p \downarrow 0} \frac{\mathbf{P}(W_a \cap S_d)}{\mathbf{P}(W_a \cap S_d) + \mathbf{P}(W_d \cap S_a)} = \frac{|P_1|}{|P_1| + |P_2|}. \tag{18}$$

The results shown in Figure 4.4 illustrate the importance of values of the activation probability p close to zero for system stability, and hence the possibility of retrieving the weak pattern under processes of learning and lesioning. Expression (18) shows that the weak pattern can be boosted significantly by its size alone when it has undergone lesions. An interpretation of (18) is that activation tends to a binary process when p decreases to zero, so that conditional probability (14) approaches the fraction of the number of neurons in the weak pattern relative to the total number of neurons in both patterns. It can be verified that expression (18) holds for all positive model parameters, such that $w_1 \geq \theta$.

To illustrate the effect of differences in pattern size, we calculated for the weak pattern P_1 its retrieval probability for different pattern sizes of $|P_1| = 50$, 100, 150, and 200 with a constant pattern size $|P_1| = 100$ of the strong pattern P_1 (see Figure 4.5).

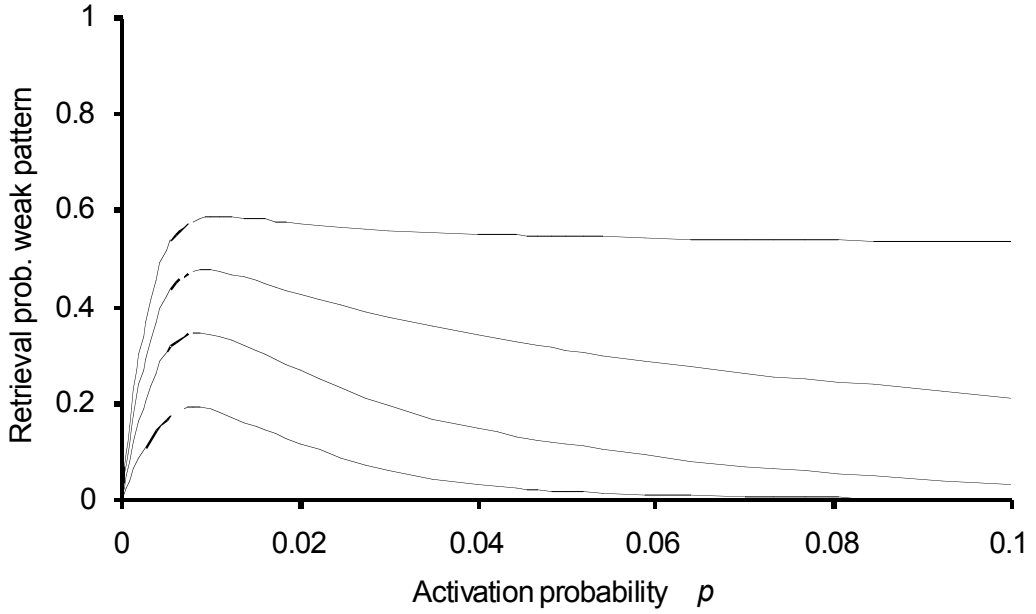


Figure 4.5. Effect of pattern size or increasing number of patterns of the weaker pattern P_1 and a varying activation probability p (x-axis) on the retrieval probability. The four lines correspond with pattern size $|P_1| = 50$, 100, 150, 200 from lowest to top line, respectively. The other parameter values were $w_1 = \theta = \nu = 0.1$ and $w_2 = 10$ (a very strong pattern P_2).

The results depicted in Figure 4.5 demonstrate clearly that an increase in pattern size increases the retrieval probability of the weaker pattern. This result is even valid in case the weights of the strong pattern are very large. We have thus shown that the effect of pattern size $|P|$ can be considerable. A pattern can compensate for its weakness in weight strength by its pattern size.

4.4 Discussion

We investigated the effect of different neural network parameters on system stability in an associative network. The research questions we addressed concerned the effect on system stability (i) of weight differences due to learning alone, (ii) of weight differences due to learning and lesioning together, (iii) the activation probability p and, (4) the pattern size $|P|$ of the memory representations. We will list the main results. Then, we will discuss the extent of these results concerning more complex systems. We will, furthermore, discuss each result in more detail with respect to more complex systems in particular the brain.

The main results of this investigation for autonomous self-repair are that:

1. Learning after lesioning increases destabilization compared to randomly cued learning alone.
2. There is a significant influence of the input activation probability p on retrieval probability. In particular, very small activation probabilities p have a limited effect on system stability for learning, but a huge impact under lesioning after learning.
3. There is an effect of pattern size on retrieval probability and system stability, which may be substantial.
4. The effects of weights and activation probability on retrieval probability are non-linear.

These results also apply to memory systems having more than two memory representations. Both the weak pattern as well as the strong pattern can be interpreted as a set of patterns. If a system has more than two patterns, they can be either categorized in the weak or strong pattern, because stability in this chapter is about losing one weak pattern or one very strong runaway pattern. If a pattern consists of multiple patterns the retrieval probability of each pattern changes a little, because it misses connections compared with a single pattern. An approximate calculation of the missing number of connections is *number of patterns* * *pattern size*. This approximation has the following assumptions: the pattern size is such that it can be divided by a natural number and the result is a natural number, the patterns are orthogonal, patterns are equally strong, and patterns are fully connected. In case of the one

pattern case when the original retrieval probabilities were large, the overall behaviour of the many memory representations interpretation will be the similar. There are other differences between the system of this chapter and the brain that can affect the results. It is expected that these differences are more favorable for memory stabilization. For instance, the brain possesses other mechanisms for memory stabilization. In other words, the results of this chapter are a lower bound for real performance in the brain.

1. The first result is derived from the outcome that with learning alone the retrieval probability of the strongest pattern may reach a maximum that can be far from one, depending on the weights of the weakest pattern. This means for the weakest pattern that it always has a positive probability. The explanation for this behavior is that there is always a probability that there is no activation in the input map of the strongest pattern. As long as the weakest pattern is strong enough to exceed its threshold there is a probability that it can be activated. Thus when the weaker pattern is lesioned, after the strongest pattern has been learned, the maximum of the strongest pattern can grow. If some lesion size is reached, the retrieval probability of the weakest pattern can become zero and the retrieval probability of the strongest pattern can become close to one. The statement that (randomly cued) learning with lesioning is worse than randomly cued learning alone depends on the type of learning. If learning can weaken connections, as is the case with some types of Hebbian learning (T.J. Sejnowski, 1977), this suffers the same consequences as random cued learning with lesions. Our analysis elucidate why this is the case.

As was mentioned in the Introduction, self-repair is a process taking place over time, where in each time step self-repair and damage can take place. Intuitively, one would think that a larger lesion or higher learning rate at each time step will decrease system stability over time. This is indeed the case. It is illustrated with the numerical calculation of Appendix B. Figure 4.6 of the appendix shows three times an increase and decrease of a strong and weak pattern, respectively, for two different parameter values (0.1 and 0.3). The figure shows the effect of lesion size and learning rate. A larger size or higher learning rate increases faster the retrieval probability of the stronger pattern and decreases faster the retrieval probability of a weaker pattern compared with a smaller size or lower learning rate. Lesion size and learning rate thus influence the rate of instability.

These results can be applied directly to the brain: randomly cued learning with lesions is more difficult than randomly cued learning alone, and large lesion sizes and high repair learning rates increase a brain’s instability.

2. The second result is that there is an effect of the activation probability of the first map on system stability. Small values (< 0.03) are best according to the graphs of Section 4.3.2. In this case the retrieval probability of the weak pattern is high relative to its other probabilities. A reason for this is that with a low activation probability there is a higher chance that the strongest pattern is not activated at all and the retrieval probability of the weakest is therefore relatively high. However, as can be seen in Figure 4.4b, in case the activation probability approaches zero the retrieval probability can collapse to zero. This is the point where the weights of the weak pattern change from the starting weight to a lower value than the threshold. Thus, although most often low activation probabilities are favorable for system stability, this is not a general rule. Another observation is that the activation probability together with retrieval probability can be a measure of system stability. If a system is stable, and the patterns have equal weights, the range of values of the activation probability is larger with an unstable system with patterns having different strengths. The result may have consequences for rehabilitation therapies. For example, in case the rehabilitation stimuli are diffuse, the results imply that the intensity of stimulation during therapy should be low. Otherwise, only the strong patterns will be activated and strengthened, while the weak, damaged memory representations will have no chance of being retrieved and consequently strengthened. For example, in case of aphasia, were nearly all representations are weakened, rehabilitation should not concentrate on intensive learning of a small subset, but from the beginning target on a large vocabulary.

3. The third result concerns the effect of the pattern size. It shows that retrieval probability of the weaker pattern can be positive and significantly greater than zero, if the weaker pattern is larger than the stronger pattern. It can even be larger than the retrieval probability of the stronger pattern. These results imply for self-repair that a weaker pattern can be retrieved and thus be repaired with a Hebbian learning rule. Thus under the condition of differences in pattern size, a system can be stable even if there are (large) weight differences.

The weak pattern P_1 can be interpreted as a set of patterns. In this case, it represents all other patterns of the memory system except for the strong pattern P_2. The retrieval probability of P_1 as a set of patterns is less than the retrieval probability of P_1 as a single pattern, because in the multiple patterns case it misses connections compared with the single pattern case as was mentioned above. Even taking into account that the probability is smaller, it is clear that an increase in number of patterns decreases the probability of a runaway memory representation. For the brain this result implies that one very strong memory

representation cannot easily destabilize the brain, since the brain contains a very large number of memory representations.

4. The fourth result concerns the non-linear effect of the weights and activation probability on retrieval probability. This is due to the fact that the probability distribution of the number of activated neurons is discrete. Nodes fire in all or none fashion because of the non-linear threshold function. Weight changes, therefore, can cause sudden changes, or no change at all, in the number of neurons participating in the competitive strength or, in probabilistic terms, contributing to the probability mass of retrieval of a pattern. Figure 4.3a, for instance, shows that a change from the starting value $w_1 = 0.1$ to 0.095 decreases the maximum retrieval probability of the weakest pattern from ¼ to about 0.1. Then, when weight strength decreases further from 0.095 to 0.05, the retrieval probability does not change at all. The non-linear effect of weights implies that there are regions in the weight space of connections possessing very little redundancy and where small changes have a large impact. Furthermore, there are regions that possess much redundancy in which weight changes have (almost) no effect. This result can be directly applied to the brain, because the non-linear threshold function of our artificial neurons is a property from real neurons (Koch, 1999). This non-linear effect of changes in weights on the retrieval of memory representations therefore may also be found in the brain. For instance, there might be a brain disease that shows similar behavior as in Figure 4.3a, where first a relatively small lesion can have a large effect on the performance (of memory retrieval), while subsequent lesions have no effect until a certain lesion exceeds a threshold that will cause a huge drop in performance again. Our analysis, thus, predicts stable ‘plateaus’ in the progression of degenerative diseases, such as dementias.

This chapter shows critical conditions for long-term stability in a neural system with autonomous self-repair and lesions: small weight differences between patterns, small activation probabilities, and neural systems with many patterns. This last condition improves the stability much, allowing the other conditions to be weakened. Since the brain is such a huge memory system possessing many patterns it seems that autonomous self-repair in the brain is feasible. It is important to note that these conditions are valid under the most difficult case of autonomous self-repair with random stimuli. If autonomous self-repair is carried out with more informed stimuli, the conditions can be weakened.

Appendices

Appendix A

We give the derivation of expression (7). The probability $\mathbf{P}(W_a \cap S_d)$ can be written as

$$\mathbf{P}\{A_{2,j} = 0 \ \forall j \in P_2\} - \mathbf{P}\{A_{2,i} = 0, A_{2,j} = 0 \ \forall i \in P_1 \ \forall j \in P_2\}. \tag{A1}$$

We write the first probability in terms of the potential $U_{2,j}$ according to (5). The assumptions of global inhibition and equal weights for both patterns imply that

$$\mathbf{P}\{A_{2,j} = 0 \ \forall j \in P_2\} = \mathbf{P}\{w_2 \sum_{j \in P_2} A_{1,j} - v \sum_{i \in P_1} A_{1,i} < \theta\}.$$

In order to evaluate this probability, we apply the theorem of total probabilities by conditioning on the number of activated neurons k_1 in P_1 and k_2 in P_2, which gives the following expression:

$$\sum_{k_1=0}^{|P_1|} \sum_{k_2=0}^{|P_2|} \mathbf{P}\{w_2 k_2 - v k_1 < \theta\} \mathbf{P}\{\sum_{i \in P_1} A_{1,i} = k_1\} \mathbf{P}\{\sum_{j \in P_2} A_{1,j} = k_2\},$$

where $|P_1|$ and $|P_2|$ denote the total number of neurons in P_1 and P_2, respectively. The first probability in this expression can be written as an indicator function of θ, so that

$$\mathbf{P}\{A_{2,j} = 0 \ \forall j \in P_2\} = \sum_{k_1=0}^{|P_1|} \sum_{k_2=0}^{|P_2|} 1_{(w_2 k_2 - v k_1, \infty)}(\theta) \mathbf{P}\{\sum_{i \in P_1} A_{1,i} = k_1\} \mathbf{P}\{\sum_{j \in P_2} A_{1,j} = k_2\}. \tag{A2}$$

We proceed with the derivation of the second probability in (A1). Switching from activations to potentials according to expression (5), we obtain the probability

$$\mathbf{P}\{w_1 \sum_{i \in P_1} A_{1,i} - v \sum_{j \in P_2} A_{1,j} < \theta, w_2 \sum_{j \in P_2} A_{1,j} - v \sum_{i \in P_1} A_{1,i} < \theta\}.$$

By applying the theorem of total probabilities as above we get the expression

$$\sum_{k_1=0}^{|P_1|}\sum_{k_2=0}^{|P_2|}\mathbf{P}\{w_1k_1 - vk_2 < \theta, w_2k_2 - vk_1 < \theta\}\mathbf{P}\{\sum_{i\in P_1} A_{1,i} = k_1\}\mathbf{P}\{\sum_{j\in P_2} A_{1,j} = k_2\},$$

which can be written as

$$\sum_{k_1=0}^{|P_1|}\sum_{k_2=0}^{|P_2|}1_{(w_1k_1-vk_2,\infty)}(\theta)1_{(w_2k_2-vk_1,\infty)}(\theta)\mathbf{P}\{\sum_{i\in P_1} A_{1,i} = k_1\}\mathbf{P}\{\sum_{j\in P_2} A_{1,j} = k_2\}. \quad \text{(A3)}$$

Expressions (A2) and (A3) can be combined by calculating the difference between the indicator functions in (A2) and (A3). By doing so we obtain expression (7) in the text.

Appendix B. The effect of different step sizes of learning rate and lesion size

To investigate the effect of different step sizes of the learning rate and the lesion size we did the following numerical simulation. We model two cases of a simultaneously lesioning of the weak pattern P_1 and strengthening of the strong pattern P_2 over three time steps. The lesion size and the learning rate were similar. They were in the first case 0.1 and in the second case 0.3. Starting from a situation where w_1=w_2 we increase the weights of pattern P_2 with 0.1, simulating a learning rate of 0.1, and decrease the weights of pattern P_1 with 0.1, which simulates a lesion of 0.1. We repeat this weight increase and weight decrease with the same step size two times. We compare the step size of 0.1 with a step size of 0.3. For a weight increase and weight decrease of 0.1 we repeat the procedure of the previous step size. We depict the two cases in Figure 4.6.

The middle line represents the situation where w_1=w_2. The two closest upper lines (indicated with 2x0.1 and 3*0.1) and two closest lines below (2*-0.1 and 3*-0.1) represent the situation where the increase and decrease with a *step size of 0.1* is repeated 2 and 3 times, respectively. The most two outer upper lines and two lower lines represent the situations where the increase and decrease with a *step size of 0.3* is repeated 2 and 3 times, respectively.

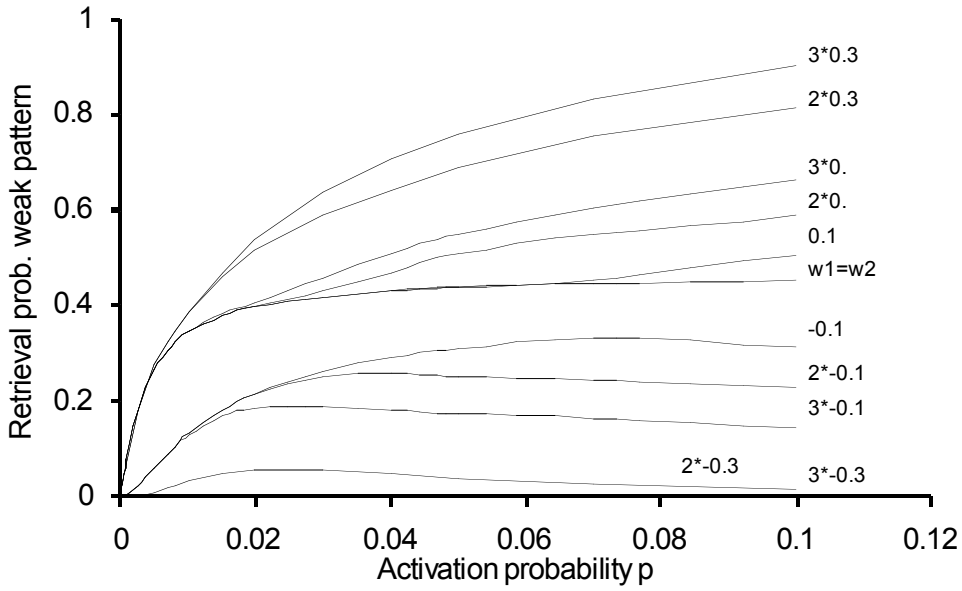


Figure 4.6. Effect of a decreasing weight w_1 of the weaker pattern P_1, a simultaneous increasing weight w_2 of the stronger pattern P_2, and varying activation probability p (x-axis) on the retrieval probability of the weaker pattern P_1 and stronger pattern, for $P_2 = 100$, $\theta = v = 0.1$. The line in the middle corresponds to the situation where $w_1 = w_2$ (the dashed line labeled with w1=w2). All the lines below this middle line represent the retrieval probability of the weaker pattern P_1 of a simultaneously increase and decrease of 0.1, 0.2, 0.3, 0.6, and 0.9. They are labeled in the figure by -0.1, 2*-0.1, 3*-0.1, 2*-0.3, and 3*-0.3, respectively. All the lines above this middle line represent the retrieval probability of the stronger pattern P_2 of a simultaneously increase and decrease of 0.1, 0.2, 0.3, 0.6, and 0.9. They are labeled in the figure by 0.1, 2*0.1, 3*0.1, 2*0.3, and 3*0.3, respectively.

From the result that a simultaneous lesioning and learning of 0.1 after three time steps is equivalent to a simultaneous lesioning and learning of 0.3 after one time step, it can be easily inferred that a larger lesion size and learning rate is destabilizing the system faster.

5

Investigating self-repair in a cortical neural network

Abstract

In the previous chapters we have demonstrated the approach of self-repair as maintenance of redundancy in various connectionist models. In this chapter, we will demonstrate autonomous self-repair in a more plausible neurobiological neural network that may represent a small part of somato-sensory cortex. This demonstration will bring us a step further in the proof of the hypothesis that self-repair is process that takes place in the brain. We will, furthermore, show that for successful self-repair the amount of self-repair and amount of damage have to be in balance and that this balance is dynamic. We will also discuss how this model and its parameters should be viewed with respect to the brain.

5.1 Introduction

In this chapter we will study self-repair as maintenance of redundancy in a more detailed cortical neural network as compared to the models of our previous work (Murre et al., submitted). The main research question addressed in this chapter is whether this idea of self-repair can work in such a model. We will present a neural network, in which memory representations are neural assemblies. These assemblies have integrate and fire neurons with lateral inhibitory connections with other assemblies, as can be found in different parts of the cortex. The model may represent for instance a small part of the primary somato-sensory cortex (SI), in which each neural assembly represents a finger of one hand.

The process of self-repair consists of 1) a cue activating neurons in the network, 2) spreading of activation to connected neurons according to a neuron activation rule, and 3) updating of connections according to a learning rule. As in previous chapters, we will use an activation cue or stimulus that is generated from a random probability process to retrieve a stored memory representation. It is possible to quantify the distance of a retrieval stimulus from a training stimulus. For instance, if retrieval and training stimuli are binary one can use the Hamming-distance to determine distance or similarity. Repair strategies can be classified on the basis of the distance of the stimuli used. A recovery or repair strategy using randomly generated retrieval cues, which are most probably very dissimilar from training stimuli, is defined as autonomous repair, because repair is unsupervised and there is no control over which patterns are repaired and the number of times they are repaired.

Autonomous repair, if left uncontrolled, has one large drawback, namely runaway repair. In this case, competitive strength of the runaway memory representation is constantly enhanced by increasing its weights, which increases the probability for its subsequent selection during retrieval (see Chapter Four). This degenerates into the predominant repair of one pattern that gradually overtakes all resources. In this chapter, we will refer to these ‘runaway’ patterns as dominant assemblies, because these memories will dominate the memory system at a certain moment. Runaway repair resembles the problems of runaway consolidation and runaway synaptic modification (Hasselmo, 1994; Meeter, 2003). A permanent derivation from its starting number of assemblies is defined as network destabilization (see Chapter Four).

Runaway repair suggests that it may not be easy to achieve stable self-repair with an autonomous repair strategy. We have, nonetheless, taken this type of strategy, because if it is successful, any other repair with more informed stimuli can also succeed in artificial neural

networks. The brain utilizes probably more informed stimuli than random stimuli for self-repair, because it is adapted evolutionary and is fine-tuned throughout lifetime to the stimuli of its environment. In other words, self-repair in the brain is easier than autonomous self-repair in (artificial) neural networks. Demonstrating autonomous self-repair in a neural network means therefore that self-repair in the brain can also work.

This chapter is organized as follows; First in Section 5.2.1 we will discuss the neural network model: its architecture in 5.2.1.1, the neuron model in 5.2.1.2, and the plasticity rule in 5.2.1.3. In Section 5.2.2, we will present the general format of a self-repair simulation. In Section 5.2.3 methods for analyzing a dynamic neural network will be presented. In Section 5.3, we will demonstrate self-repair in the cortical network and investigate the effect of the amount of damage and the amount of self-repair on network stability. In the final section, we will discuss how the model, some of its parameters, and their values should be viewed in the context of the brain.

5.2 Methods

5.2.1.1 The neural network

The neural network model comprises three parts of the brain: a sensory-, a sub-cortical-, and a cortical map, each having a two-dimensional torus topology structure comprising 100 excitatory neurons. The cortical map also contains 100 interneurons. The model and its projections are depicted in Figure 5.1.

The network has 5 projections: (1) a sensory–subcortical projection, (2) a subcortico–cortical projection, (3) a cortico-cortical projection between pyramidal neurons, (4) a cortico-cortical projection from the pyramidal neurons to the interneurons, and (5) a cortico–cortical projection from the interneurons to the pyramidal neurons. The last two projections form a feedback inhibition loop of the cortical map.

Connections are initialized with a connection density function and a synaptic weight function. The connection density function determines how many neurons are connected. It controls both the connectivity within an assembly (intra-connectivity) and between different assemblies (the inter-connectivity). To keep distance computations simple for the density and synaptic weight function, the cortical pyramidal and interneuron map are of equal size. A functionally equivalent type of inhibition could be achieved with the more neurobiological plausible solution of fewer inhibitory neurons relative to the pyramidal neurons.

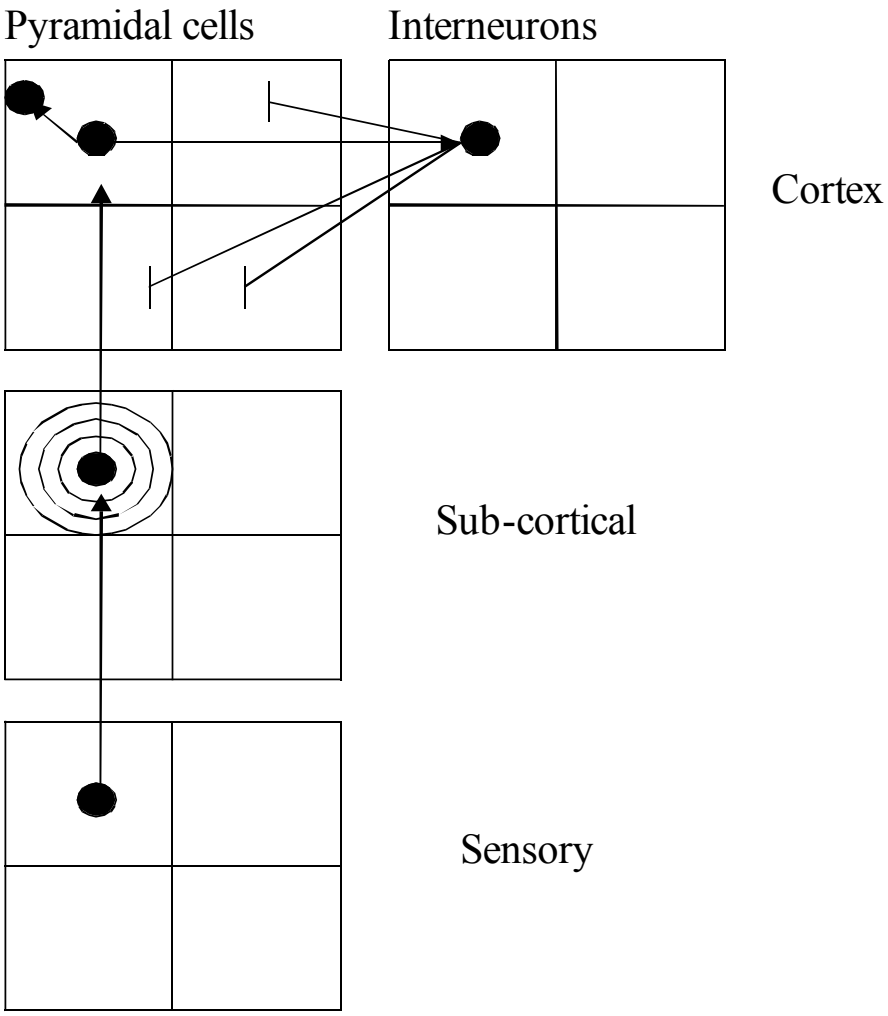


Figure 5.1. The figure shows that each map is divided into four assemblies. For one neuron in each map an example connection is depicted. For a sub-cortical neuron its rings determining its distance are shown. The picture, furthermore, shows a cortico-cortico connection for a pyramidal neuron and feedback inhibition to neurons of other assemblies.

The synaptic weight function determines the connection strength or weight from one neuron to its surrounding neurons. We use the function for the normal distribution $n(0,\pi/50)$ as a synaptic weight function, where 0 is the mean and $\pi/50$ is the variance. Both the connection density and the synaptic weight function are dependent on the distance in a hexagonal topology between the different neurons. The synaptic weight and density function are such that in the cortical network a computational map of competitive neural assemblies emerged, as in the model of von der Malsburg (Malsburg, 1973) and its abstraction by Kohonen (Kohonen, 1982). More details about each function and each projection are given in Appendix A. In the simulations described below, all connections were kept constant, with the

exception of the cortico-cortical connections of the pyramidal neurons. The latter connections are referred to as the self-repair map.

5.2.1.2 The neuron model

The model neuron of the neural network is a simplification of the MacGregor neuron (MacGregor & Oliver, 1974), which in turn is derived from the Hodgkin-Huxley neuron (Hodgkin & Huxley, 1952). The model neuron is a tradeoff between neural plausibility and computational cost of simulating neuronal spiking and adaptation. In numerical simulations, the state of the neurons is updated in discrete time steps, which in our simulations took two msec.

The neuron's membrane potential, U, depends on the sodium, potassium and chloride currents flowing over the membrane. It can be described by the following differential equation:

$$\frac{dU}{dt} = -\delta U - g_k(U - U^k) - g_{ex}(U - U^{ex}) - g_i(U - U^{in}) - U^0. \quad (1)$$

Here $-\delta U$ is a leakage current, g_{ex} the excitatory conductance, E_{ex} the natrium reversal potential, g_i the conductance due to inhibitory synapses, and E_i the chloride reversal potential. The parameter governing the leakage current is different for each map and given in Appendix B.

Adaptation is modeled by allowing the potassium conductance g_k to vary over time according to the following equation:

$$\frac{dg_k}{dt} = \frac{-g_k}{\tau} + bS, \quad (2)$$

where S is a dichotomous spiking variable. The time constant τ differs for each map and is specified in Appendix B.

Excitatory and inhibitory input to the i^{-th} node is a linear summation of weighted inputs to that node

$$g_{ex/in} = \sum_j w_{ij} S_j, \quad (3)$$

where w_{ij} is the weight from node j to node i, and S_j is the binary spiking variable of node j. The weight w_{ij} is negative in case it is inhibitory.

The model neuron emits a spike every time the membrane potential U crosses the threshold Θ.

$$S = \begin{cases} 0 & if\ U < \Theta \\ 1 & if\ U \geq \Theta \end{cases} \tag{4}$$

S is a dichotomous variable with a value 1 if a spike is emitted and 0 otherwise.

For the computer simulations, the discrete time approximation formulas of MacGregor & Oliver (MacGregor & Oliver, 1974) were used. The model was implemented in Nutshell, the neural network simulator developed in our group (www.neuromod.org/nutshell).

5.2.1.3 Hebbian learning

Since we argue that self-repair taking place in the network model should correspond to self-repair in biological brains, the learning rule adopted in the network model should also have a reference to a learning rule in the brain. The most likely candidate for Hebbian learning is long-term potentiation (LTP) (Bliss & Lomo, 1973; Martin et al., 2000; McEachern & Shaw, 2001). In the current network model the Singer-Hebb learning rule (Singer, 1990) is applied. This learning changes a weight when the postsynaptic neuron is active at time $t+1$: a weight increases with amount μ ($0 \leq \mu \leq 1$) if the pre-synaptic neuron was active and decreases if it was inactive at time t. When the post-synaptic neuron was inactive at time t weights remain constant:

$$\Delta w_{ij} = \begin{cases} \mu & if \quad S(j,t) = 1 \quad and \quad S(i,t+1) = 1 \\ -\mu & if \quad S(j,t) = 0 \quad and \quad S(i,t+1) = 1 \\ 0 & else \end{cases}, \tag{5}$$

where Δw_{ij} is the change in weight of the connection from node j to node i. The neuronal output S - a spike or no spike - is determined by equation (4). To keep the weights between a minimal and a maximal bound of 0 and 1 we apply the following update rule for the weights:

$$w_{ij}(t+1) = \min(\max(w_{ij}(t) + \Delta w_{ij}, 0), 1). \tag{6}$$

We use an inward normalization rule, where all incoming weights of a neuron are scaled:

$$w_{ij}(t+1) = \frac{w_{ij}(t)}{\sum_j w_{ij}(t)}, \tag{7}$$

where the weights of a neuron are updated synchronously. The use of normalization seems justified since simulation studies have shown that normalization is an intrinsic property of spike-based temporal learning (Kempter *et al.*, 2001), a type of learning that can be found in the brain (Bi, 2002). There is, furthermore, direct evidence for normalization known as synaptic scaling (Turrigiano et al., 1998; Turrigiano & Nelson, 2004).

5.2.2 Simulation procedure

The general format of a self-repair simulation is as follows: a number of memory representations is stored in the network with the connection density and synaptic weight function (Section 5.2.1.1), then for a given number of time steps the memory representations are damaged, self-repair takes place, and the network is tested. We store memory patterns in the network with the in 5.2.1.1 described synaptic density and weight function such that in each map four clusters are formed according to the cluster algorithm. Damage to the synapses is modeled by

$$w_{ij}(t+c) = w_{ij}(t) - \varepsilon, \tag{8}$$

where time t is a given time and ε is a pertubation. The constant c determines the period with which the damage is administered and represents an accumulated lesion over that period. In this case the period is 40 time steps. The stochastic perturbation ε, added to each synapse, is modeled by drawing a random number from a uniform distribution between 0 and a maximum l, which we call the lesion size.

Self-repair is modeled by a three step process in which (1) neurons of the sensory map are randomly activated, (2) activation is allowed to spread over the rest of the network according to the activation equations 1 to 3 described in Section 5.2.2, and (3) connections are added or updated in the cortico-cortical projection of pyramidal neurons with the Hebbian learning rule and normalization rule described in 5 and 7, respectively. For the normalization of the cortico-cortical projection, we leave the subcortical-cortical projection out of the calculation, because it is constant during a simulation. The random activation of the sensory

neurons is modeled by a spatially and temporally homogeneous Bernoulli process, which is a discretized analogue of the Poisson point process (Stoyan *et al.*, 1997). This leaves us with one parameter, referred to as the intensity parameter that determines the activation probability of each neuron at each time step. To find the optimal probability for the self-repair process we conducted simulations as described in Appendix C and selected an intensity of 0.03.

The network performance is tested with the methods described in Section 5.2.3. In all simulations we use the same values for the neural parameters. For a description of the parameters and their values see Appendix B.

The time scheme of the above-described processes, such as lesion and repair, is as follows. First, the network is initialized, then:

1. At each time step self-repair takes place.
2. At each 10^{th} time step the cortico-cortical projection of the pyramidal neurons is tested for the number of assemblies with the cluster algorithm and the total projection weight.
3. At each 40^{th} time step, damage is administered to all weights according to equation 8.

5.2.3 Analysis of memory representations

The measurement of memory representations in a spatio-temporal domain is not a trivial problem. The neural network consists of a recurrent neural network that receives (stochastically) independent input from the sensory map. The spatio-temporal dynamics in the neural network presented here is equivalent to a (stochastic) non-autonomous system in which the neuronal activity is much more complex than the usual point attractors or limit cycles. Indeed, of this type of neural networks it is not even clear whether it possesses attractors (Arbib *et al.*, 1998). To nonetheless be able to analyze network behavior, we identify memory representations or memory traces by their weights instead of by their neural activity. The assumption here, is that neurons of a same memory representation will have a higher spike correlation, if they have more connections to neurons of a same memory representation than to neurons of another memory representation.

Memory representations in this study are represented by neural assemblies of which the intra-connectivity is higher than their inter-connectivity to any other assembly. When the plasticity of the cortico-cortical projection of the pyramidal neurons is enabled, the connectivity within the cortical map is altered, which may affect the number of assemblies. A measure of network performance, therefore, is the number of assemblies. To identify them by their weights, we used a variant of a cluster algorithm by Xing & Gerstein (Xing & Gerstein, 1996a). The algorithm finds the number of clusters or assemblies such that each cluster is

most strongly connected to itself, which is the 'self-connectivity' of a cluster. The average self-connectivity of all clusters in a map is the map-connectivity. In other words, the algorithm tunes the number of assemblies such that the map-connectivity is maximal. The map-connectivity is a real number between zero and one. There is no or random coherence between the different clusters if the map-connectivity is 0.5. For more information on the cluster algorithm we refer to Appendix D.

A network stabilizing strategy that may be imposed is one that guards the initial number of assemblies to be approximately preserved. Merely checking at each time step whether the network retains its initial number of assemblies would be too strict, because self-repair and damage can impose transient changes in the network that should not be immediately regarded as faulty network behavior.

We do not analyze whether a memory representation has moved. The cluster algorithm checks only for the number of assemblies, so the assembly may have shifted. We checked for this for some simulation runs (not reported in this chapter), but it was not the case. We should add that the moving of memory representations does not necessarily have to be regarded as a problem. It is known that memory representations can move within the cortex (for a review see for example Buonomano and Merzenich (Buonomano & Merzenich, 1998)).

5.3 Results

5.3.1 Demonstration of self-repair in the cortical neural network

In this section we will demonstrate self-repair as maintenance of redundancy in the above described model. Figure 5.2 shows a simulation with self-repair and one without self-repair, both having a lesion size of 0.002. In the simulation with self-repair the number of assemblies is changing frequently in the beginning, but after about 1500 time steps, it fluctuates less, remaining close to the initial four assemblies. It thus seems that a balance between damage and self-repair emerges. The balance is dynamic, since the number of assemblies can fluctuate over time. In the simulation that ran without self-repair all assemblies eventually disappeared. The degradation of that network could be divided into two phases: in the first phase the number of assemblies was slowly increasing until some large number of assemblies had been reached. In the second phase, the number of assemblies rapidly decreased until none were left.

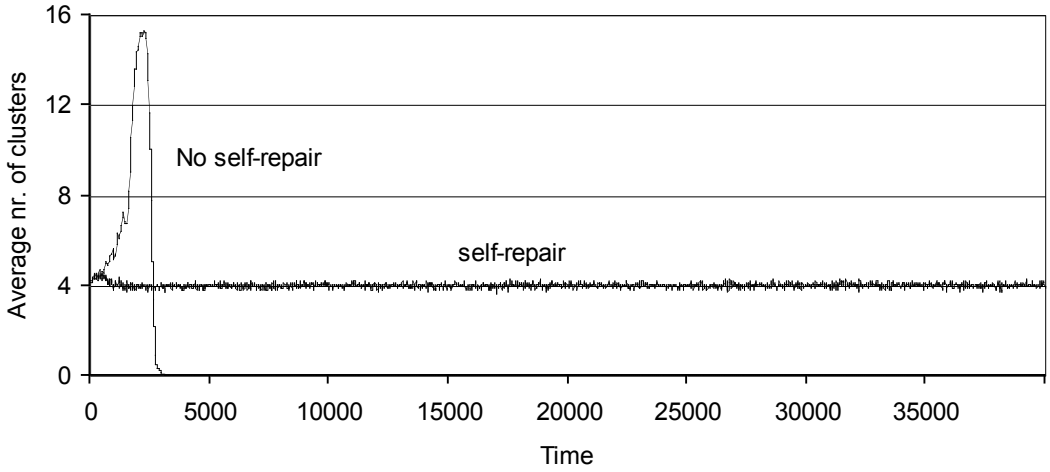


Figure 5.2. This figure illustrates the effect of self-repair. It shows a simulation with self-repair and one without self-repair. In both simulations the lesion size is 0.002. The results give the number of assemblies (y-axis) over time (x-axis). The figure clearly shows the effect of self-repair. For a more detailed discussion of the results see Section 5.3.1. All results of the figure are the average results of 20 simulations.

These results demonstrate that self-repair is able to protect the network from small cumulative lesions that otherwise would have destroyed the network connectivity structure.

5.3.2 Investigating the effect of the amount of self-repair and amount of damage

In this section, we will investigate the effect of different amounts of self-repair and damage on network stability. The amount of self-repair is determined by the learning rate and self-repair frequency and the amount of damage is determined by the lesion size and lesion frequency. Since both frequencies in this study are constant according to the simulation scheme described in Section 5.2.2, we will only vary the learning rate and the lesion size. The simulations were computationally expensive. We, therefore, shortened the simulation period and first carried out an exploratory search with single runs (reported in Appendix E), in which all variations of parameter values were used. The learning rate and the amount of damage studied were 0, 0.002, 0.005, and 0.008. We have chosen these parameter values, because they correspond to about the average weight of the cortico-cortical connection of the pyramidal neurons, one order of magnitude smaller than the average weight of cortico-cortical connection of the pyramidal neurons, and a value in between the two previously mentioned values. Each simulation took 4000 time steps and has the format as described in Section 5.2.2.

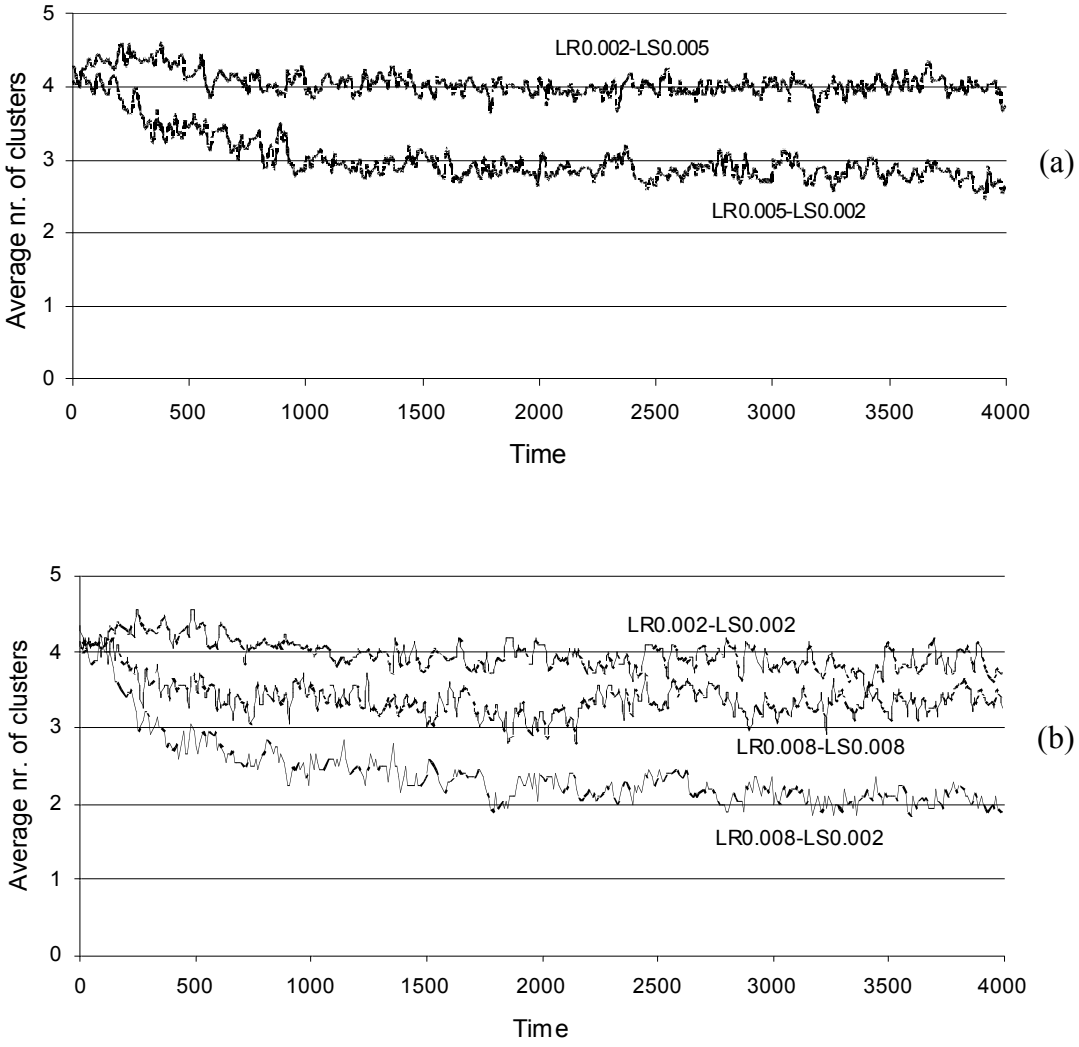


Figure 5.3. The simulations in these figures were carried out to investigate the effect of the amount of damage and the amount of self-repair. (a) Figure 5.3a shows the results of a simulation that had a learning rate of 0.002 and a lesion size of 0.005 and a simulation that had a learning rate of 0.005 and a lesion size of 0.002. The upper line represents the result of the first simulation and is labeled with "LR0002LS0005". The result of the second simulation is represented by the bottom line and is labeled with "LR0005LS0002". The results give the number of assemblies (y-axis) over time (x-axis). Both lines represent the results of an average of 20 simulations. (b) Figure 5.3b shows from top to bottom the results representing the self-repair simulations with a learning rate of 0.002 and a lesion size of 0.002, a learning rate of 0.008 and lesion size of 0.008, and a learning rate of 0.008 and a lesion size of 0.002. The results give the number of assemblies (y-axis) over time (x-axis). Each line represents the results of an average of 20 simulations.

A first conclusion is that self-repair is possible with a range of parameter values. This is shown by the simulations with learning rate 0.002 and 0.005 both with a lesion size of 0.002. A second conclusion is that the processes of self-repair and damage must be balanced. From Figure 5.3a and 5.3b (also see Tables Three and Four of Appendix E), it seems that if the learning rate is larger than the lesion size self-repair cannot retain the initial number of four assemblies. The more the learning rate exceeds the lesion size, the larger the error in the number of assemblies. This is clearly illustrated by Figure 5.2b, where the average error in number of assemblies is larger for a simulation with lesion size 0.002 than lesion size 0.008 both having a learning rate of 0.008.

5.4 Discussion

The main result of this chapter is that we were able to build a more neurobiological detailed model in which we demonstrated self-repair. This connectionist model possesses more realistic neurobiological details than the models of the previous chapters. This brings us a step closer of proving the hypothesis that self-repair is taking place in the brain. This is also supported by the fact that we demonstrated autonomous self-repair in this model. This type of self-repair is more difficult in artificial neural networks than in the brain, because the brain probably uses better cues for memory retrieval. The results, furthermore, suggest that for network stability there has to be a dynamic balance between self-repair and damage. We speak of a dynamic balance, since both processes take place continuously and the balance is in constant flux. The results showed how the balance is influenced by the amount of damage and the amount of self-repair. The network is stable if the amount of damage and self-repair are correctly tuned. If either one is dominant, the network will become unstable, resulting in the loss of assemblies or its disintegration indicated in a first phase by an increasing amount of assemblies. In the remainder of this section, we will discuss model parameters such as the amount of damage, the learning rule, and the initial connectivity between assemblies with respect to the brain.

Amount of damage. In our study damage took place on a timescale of milliseconds. In the real world such damage in the brain takes place on a timescale of days to months. Thus, if we demonstrate that autonomous self-repair on a millisecond timescale can counteract a given lesion size, we are certain that it can also counteract those lesion size on timescale of days to months. We regarded a single lesion size as an accumulated lesion over 40 time steps. Using this lesion size, the content of the cortex vanished in a few seconds, unless autonomous self-repair counteracted the lesion. In that case stability could be achieved. Having demonstrated

the fact that stability in the brain can be achieved with "large" lesion sizes means that autonomous self-repair can also counteract "smaller" lesion sizes on a timescale of days to months, as it is easier to attain stability with a smaller lesion size than with a larger lesion size (see Chapter Four).

Learning rule. In this model we use a Hebbian learning mechanism with linear normalization as described in Section 5.2.2. One of the ideas of the self-repair theory is that normal plasticity mechanisms can carry out self-repair. Since there is proof for normalization in the brain (Turrigiano et al., 1998) and it has been suggested that homeostasis mechanisms active during sleep may lead to a more optimal functioning of the brain system (Tononi & Cirelli, 2003), it is interesting to investigate whether normalization alone is able to carry out self-repair. Neural regulation is an interesting type of normalization mechanism for future investigations as it has been shown that it can undo small lesions to connections (Horn et al., 1998a). Spike timing dependent synaptic plasticity (STDP) is another type of normalization mechanism that seems to be promising for future research. In STDP the update value of a connection depends on a period or window of activity of pre- and post-neuron instead of two time steps. Computational studies have shown for this mechanism that it has intrinsic normalization properties (Kempter et al., 2001). Moreover, there is empirical proof of the existence of STDP (for a review see (Bi, 2002)) in the brain. If the STDP mechanism would be successful, it will validate our idea that plasticity mechanisms found in a normal healthy brain are able to carry out self-repair.

Initial connectivity of the cortical map. This map in the simulations of Section 5.3.1 was initialized with assemblies that are very similar with strong local connectivity and few long-distance inter-connections. This resembles a small-world network of regular coupled networks with some disorder that gives the complete network the characteristic short path length between nodes (Watts & Strogatz, 1998). Since our model represents the cortex, it implies that the connectivity of the cortex possesses also the small-world property. This has indeed been suggested by a number of researchers (Salvador *et al.*, 2005; Sporns & Zwi, 2004; Watts & Strogatz, 1998). It has, furthermore, been suggested that small-world connectivity has advantages like efficient information exchange on a local as well as global scale (Latora & Marchiori, 2001), fast learning (Simard *et al.*, 2005), and provides a good tradeoff between computational efficiency and wiring costs (McGraw & Menzinger, 2003). We suggest that the few extra long-distance connections between the assemblies can have the advantage that self-repair is not restricted to one assembly, but can spread to other assemblies. Moreover, in a small world any other region can be reached in few steps meaning that any

region in the cortex can contribute to repair of any other region. Self-repair can, however, only have a positive effect if the activity of the different assemblies is spaced sufficiently in time, so that Hebbian learning cannot grow many new connections between different assemblies to keep different memory representations separated This separation is, for instance, needed to obtain linking fields in long-term memory that can be used by working memory (Luck & Vogel, 1998; Raffone & Wolters, 2001; Raffone, Wolters, & Murre, 2003). To keep the neural assemblies separated, the time-window of plasticity has to be small such that learning can take place only between nearly activated assemblies. This or any other mechanism should keep the inter-connectivity of assemblies below a critical limit such that the probability of simultaneous activation of many assemblies remains low. If this is the case, the small-world network structure can have a positive effect on self-repair.

Appendices

Appendix A. Network connectivity

The typical connection density of the network is show in Table 5.1.

Table 5.1. This table shows the connection density between the different maps.

Connection density function

Table 5.1

Projection	Probability of a connection within an assembly								Probability of a connection between assemblies	
From map -To map	r 0	1	2	3	4	5	6	7	r 0	1
sensory - subcortical	1.0	1.0	1.0	1.0	0	0	0	0	0	0.1
subcortical - cortical	1.0	1.0	1.0	1.0	0	0	0	0	0	0.1
cortical - cortical	0	1.0	1.0	1.0	1.0	1.0	0	0	0	0.1
cortical - cortical-inhibitory	0	1.0	1.0	1.0	1.0	1.0	0	0	0	0.1
cortical-inhibitory - cortical	0	0	0	0	0	0	1.0	1.0	0	0.1

The first column indicates the projection. The second column indicates the probability of a connection between a neuron from the from-map and a neuron from the to-map within an assembly depending on the radial distance *r*. The third column indicates the probability of a connection between a neuron from the from-map and a neuron from the to-map between assemblies. Every map has the same size so that every neuron of the from-map has a corresponding neuron in the to-map. For reasons of simplicity the distance between these neurons is zero (no time-delays). The distance between the neuron in the from-map and other neurons in the to-map is calculated relative to the corresponding neuron of the to-map. The table shows the typical connectivity between assemblies. Neurons in the cortical map are not connected to themselves. There is full connectivity within an assembly in a certain radius *r*, for instance sensory neurons are connected to subcortical neurons of their assembly in a radial distance of three. The simulations of Section 5.3.1 had a ten percent probability of connections between neural assemblies (inter-connectivity) of neurons that are one step away from each other in a hexagonal topology. There is no inter-connectivity of the simulations of Section 5.3.2. Cortical neurons have (inhibitory) connections with inter-neurons that have a radial distance of 6 and 7.

The synaptic weight function determines the connection strength or weight from one neuron to its surrounding neurons. We use the function for the normal distribution $n(0,(\pi/50))$

as a synaptic weight function, where *0* is the mean and *π/50* is the variance. The synaptic weight *w* is

$$w = c\frac{1}{\sqrt{2\pi}}e^{-\frac{(r-\mu)^2}{2\sigma}}, \quad (9)$$

where *r* is the distance from the neuron to one of its neighbors and *c* is a constant controlling the neural activity in a map. It is 5 for the cortical-cortical – inhibitory projection, 8 for the cortical-inhibitory – cortical projection, and 1 for all other projections.

Appendix B. Neural parameters

The neural parameters are such that a stimulus of nine neurons administered to the network for 5 ms would activate the cortical map for 5-10 ms. The values of the different neural parameters are given in Table 5.2.

Table 5.2. This table shows the values of the neural parameters of each map.

Table Neural parameters

Table 5.2

Neural parameter	Map Sensory	Sub-cortical	Cortex pyramidal	Cortex interneurons
τ_U	0.5	7.5	0.1	10
τ_{Gk}	0.5	7.5	0.1	10
b	0.5	7.5	0.1	10

Appendix C. Simulations to find the appropriate stimulus intensity

In this study we had set ourselves the task to implement self-repair with a random cue. We modeled it by activating a number of neurons in the sensory map according to a spatially and temporal homogeneous Bernoulli process, which is a discretized analogue of the Poisson point process (Stoyan et al., 1997). This left us with one parameter, referred to as the intensity parameter that determines the activation probability of each neuron at each time step. With the manipulation of this parameter we tried to find a stimulus that gave the best response for self-repair in the cortical map. The best cortical activation pattern for each time step is an

activation pattern that restricts itself to one assembly. To measure the performance of the network, while taking the restriction of activation into account, we had the following signal to noise ratio measure (SNR). It is the number of neurons of the most activated neural assembly in the cortical map (the signal), divided by the total number of activated neurons in the cortical map.

$$R = \max_a \frac{A_a}{\sum_i^n A_i}, \tag{10}$$

where A_a indicates the assembly that has the largest activation and n is the number of assemblies. To investigate the effect of the stimulus intensity parameter on the activation pattern in the cortical map of a certain weight configuration, learning was disabled and no lesions were administered to the weights. We did four simulations, two simulations with unnormalized weights and two simulations with normalized weights each one differing in inter-connectivity of the assemblies. There was one simulation with 10% connected assemblies and another with isolated assemblies.

In each simulation we used the following procedure to find the stimulus intensity with the highest signal to noise ratio. A Bernoulli process in which every node had an equal spike probability generated the stimuli at each time step. The stimulus intensity parameter was varied from one to nine. The homogeneous Bernoulli stimuli were administered to the network for 300 time steps. The membrane potential of every spiking neuron of the input map was kept constant during the simulation. At each time step, activation in the self-repair map was recorded and the signal to noise ratio was calculated. In Figure 5.4, we show the results. Only the results of simulations without normalized weights are depicted. The results with normalized weights show a same pattern. The results show a decreasing signal to noise ratio when intensity is increasing, with the exception of intensity one, which was unable to activate the self-repair map.

The main difference between a model with a self-repair map with interconnected assemblies and one without interconnected assemblies is that the overall signal to noise ratio is lower for a model with a self-repair map with interconnected assemblies. This can be explained by the fact that nodes of different assemblies were able to activate each other, which resulted in a lower signal to noise ratio.

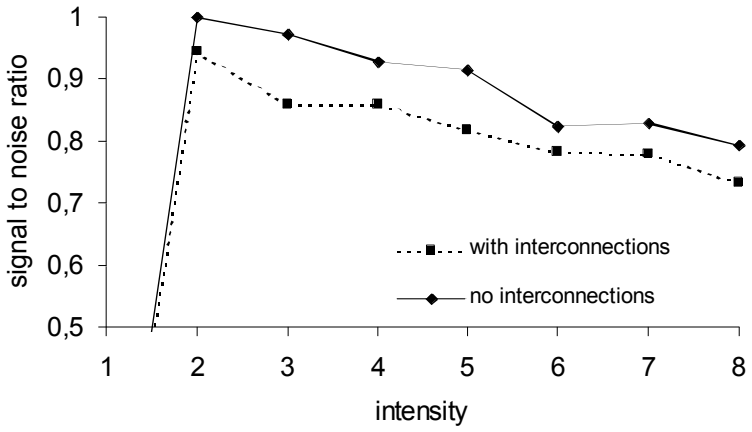


Figure 5.4. The figure shows the results of simulations to find the optimal stimulus for retrieval. In addition to the stimulus parameter two other parameters were varied: with or without normalized weights and inter-connectivity (10 percent) or no inter-connectivity. Each data point represents a simulation that took 300 time steps and gives the mean signal to noise ratio (y-axis) over a simulation. The probability of spiking of each neuron in the sensory map at each time was varied, taking integer values from 1 to 8, over the different simulations (x-axis).

The parameter of stimulus intensity determines both the number and the place of connections that are added. The amount is simply regulated by the degree of intensity: the higher the intensity the more activity in a specific area and the more repair in that area. A drawback of a high intensity is that activity is diffuse and random. In case of a high intensity, therefore, there will not be activity in a specific assembly but in many assemblies.

Even though a low intensity has a higher signal to noise ratio, it also has a drawback. If the intensity is too low there is a drawback of a high probability of a dominant assembly. For example, a stimulus intensity of two has the highest signal to noise ratio, but it does not suit our purposes as the measure of spikes per assembly shows that at each time step the same assembly is active. If this is the case, only one neural assembly is repaired, leaving the others damaged.

In general, it is desirable that the activity does not reside too long in one assembly, as this leads to an unstable network. We thus can conclude that the stimulus intensity has to be high enough for the network to switch its activity from one assembly to another, but not so high that it causes too much noise. In other words, there is a trade off between a high signal to noise ratio and the switching behavior of a network from one assembly to another.

Appendix D. The cluster algorithm

The number of clusters or assemblies in this algorithm is determined by optimizing the maximal map-connectivity, which is the average self-connectivity of all assemblies in a map. This is the sum of weights of neurons in the assembly to other neurons of its assembly, divided by the total outgoing weights of all neurons of the assembly. The algorithm begins with the calculation of the inward excitation, because observations show that those locations where neurons have larger inward excitations correspond to neuronal groups (Xing & Gerstein, 1996b). The complete algorithm is as follows:

1. The inward excitation of every neuron is calculated, which is the total sum of incoming weights of the direct neighbors in a rectangular topology divided by the total excitatory strength. The direct neighbors form the first ring of the calculated neuron. The following rings around the center neuron are formed by NxN patch around the center neuron, excluding the center neuron and neurons belonging to previous rings.
2. The M neurons with the highest inward connectivity are chosen and defined as center neurons around which neuronal assemblies are formed. To each of those M neurons, neurons of a given ring are assigned if all neurons in the ring are connected to the center neuron.
3. Neurons not identified with any assembly so far, will be assigned to an assembly with which it has the highest outward connectivity, which is the sum of the weights of a neuron to the neurons of the assembly, divided by the total outward weights of the neuron.
4. For all neurons, the outward connectivity with all assemblies in the map is calculated and they are assigned to the assembly with the highest outward connectivity
5. Finally, for all assemblies the assembly connectivity with all other assemblies, including itself (the self-connectivity of an assembly), is calculated. The connectivity of an assembly is the sum of weights of neurons in an assembly to other neurons of the assembly, divided by the total weight of all neurons of the assembly. If the self-connectivity is lower than the assembly connectivity to other assemblies, it is merged with the assembly with which it has the highest connectivity.

A map with random connectivity has a map-connectivity of 0.5. A map with fully isolated assemblies has a map-connectivity of 1.0.

Appendix E. Exploratory search of the amount of self-repair and amount of damage

In this appendix, we will investigate the effect of different amounts of self-repair and damage by varying the learning rate and the lesion size. The results reported are from single simulation runs. They are noisier than the results of the averages of 20 simulations. It is, therefore, harder to interpret the results only on the basis of the number assemblies. To anticipate this problem, we have two additional measures based on the number of assemblies: (1) the assembly loss measure and (2) the retention percentage of the initial number of assemblies for a number of tests in a given period. We speak of assembly loss, when in a time-frame of 200 time steps the occurrence of the initial number of assemblies is less than 10%. The second measure is the total number of correct retention of the initial number of assemblies in a period divided by the number of tests in the given period. With the last measure we can compare network performance even if they have an equal number of assemblies at a given moment.

In this section we will investigate how to achieve successful self-repair by varying the learning rate and lesion size. The learning rate and the amount of damage studied were 0, 0.002, 0.005, and 0.008. As mentioned in Section 5.3.2, we chose these specific values because they are about the average weight of the cortico-cortical connection of the pyramidal neurons, one order of magnitude smaller than the average weight of cortico-cortical connection of the pyramidal neurons, and a value in between the two previously mentioned values. Each simulation takes 4000 time steps and has the format as described in Section 5.2.2.

The results of the simulations, with different learning rates and amount of damage, are given in Tables 3 and 4. Table 5.3 indicates the onset time of the loss of an assembly. Table 5.4 shows for each simulation the percentage of the network in which it retained its original number of assemblies for a total simulation (first part of the table) and for the last 100 tests (second part of the table). A minus ‘-‘ means that no simulation has taken place. The meaning of the star ‘*’ symbol is explained in Table caption 3.

The results that concern the spike activity (not depicted) show for all simulations that there are no dominant assemblies for any simulation. The results of Tables 5.3 and 5.4 show that network stability is attained in case both processes of self-repair and damage are active. The results with either damage or self-repair, but not both, show that a higher lesion size or a higher learning rate accelerates the network degeneration.

Table 5.3. Table 5.3 shows the onset time of the loss of one assembly according to the assembly loss measure. A star '*' in the table means that there were no assemblies lost during the simulation period. The results are positive, no assemblies are lost, when both processes of self-repair (learning rate > 0) and damage (lesion size > 0) are taking place, and the learning rate is equal to or lower than the lesion size.

Onset time of first Assembly loss

Table 5.3

Damage amount	Learning rate time			
	0	0.002	0.005	0.008
0	-	320	350	60
0.002	670	*	270	50
0.005	280	*	*	80
0.008	200	*	*	*

Table 5.4. Table 5.4 shows the occurrence percentage of the initial number of assemblies for a total simulation (right part of the table) and for the last 100 tests (left part of the table) for a varying lesion size and learning rate, taking the values of 0, 0.002, 0.005, and 0.008. This table indicates the same trend as in Table 5.1, namely: The percentage correct for a total simulation period and for the last 100 tests are high (1) in case of self-repair and damage and (2) the learning rate is equal to or lower than the lesion size. For the last 100 tests, the better results are expressed as a percentage correct larger than 0. An exception is the simulation with a learning rate of 0.005 and a lesion size of 0.002 that has a result larger than 0. The results, however, are still worse, because the occurrence percentage is only 2%.

Maintenance of initial number of groups

Table 5.4

Damage amount	Learning rate Percentage correct total simulation				Percentage correct last 100 tests			
	0	0.002	0.005	0.008	0	0.002	0.005	0.008
0	-	8.25	9.25	2	-	0	0	0
0.002	21	74.75	11.5	1.75	0	70	2	0
0.005	8	82.25	72.5	7	0	89	69	0
0.008	5	72.5	62.25	31.75	0	89	68	64

If the lesion size is zero and the learning rate is higher than zero self-repair does not seem to work.An explanation for this erroneous/deleterious repair is that the network is learning the random stimuli. Both tables, furthermore, show that results are better when the learning rate is equal or smaller than the lesion size indicating the importance of a balance between self-repair and damage.

6

Self-repair of neural circuits during sleep

Abstract

In this chapter, the hypothesis is proposed that self-repair by maintenance of redundancy takes place during sleep, in particular autonomous self-repair with randomly cued activation. We will investigate this hypothesis with a neurobiological plausible model of sleep. We will demonstrate in this model that self-repair with randomly cued activation works and extends the lifetime of memories. The hypothesis is, furthermore, investigated by a review of models and data of processes of memory maintenance and memory consolidation taking place during sleep. This review suggests that these processes during sleep may be able to carry out self-repair, because (1) it implies that the brain possesses redundancy as is proposed by self-repair and (2) it shows that the processes of memory maintenance and memory consolidation are similar to self-repair. The second point implies that they share the same algorithmic procedure with self-repair. Autonomous self-repair in the brain driven by random cues is supported, since the sleep processes share the algorithmic part of randomly cued activation.

6.1 Introduction

We investigated self-repair by maintenance of redundancy in connectionist networks, where redundancy resides in the network's connections and maintenance is carried out by updating connections. We proposed the hypothesis that this type of self-repair is a property of the brain. In this Chapter, we investigate the hypothesis that self-repair is taking place during sleep, in particular autonomous self-repair driven by random cues. This is different from the previous chapters in which we argued mostly that autonomous self-repair in the brain consists of cues that resemble the cues that stored the memory traces and autonomous self-repair in artificial neural network consists of random cues. Thus, in this chapter autonomous self-repair in the brain is equivalent to autonomous self-repair in artificial neural networks. Below we will further discuss why autonomous self-repair with random cues could be taking place during sleep in the brain. We will first discuss connectionist redundancy and maintenance.

Redundancy is present at different levels of the brain: from the synaptic to the neural systems level. The typical connectionist redundancy at the neural systems level resides in the connections with memory representations distributed over these connections. Even when some connections are lost, the activity pattern is nearly the same as before. This is possible, because the information of the activation pattern is still available in the remaining parts of the damaged memory representation and in other parts of the network system. Parts of a memory representation may be specialized in processing specific aspects of information (e.g. perceptual or motor processing), but most of parts will be involved in many brain functions. The latter property of memory was identified by Friston (Friston, 2002) as being functional integration. Bach-y-Rita (Bach-y-Rita, 1990) named it multiplexing. In this chapter, we will encounter other researchers that attribute the brain this specific connectionist quality.

Maintenance of redundancy is carried out by a *three-step self-repair algorithm* that consists of:

(1) an activation cue, after which

(2) the activity is allowed to spread over its nodes according to a connectionist spreading activation rule, while

(3) a plasticity rule updates the connections between the nodes.

A memory representation is selected by the activation cue and the spreading activation rule. As a plasticity rule, we used a Hebbian mechanism complemented by normalization. We argued that these mechanisms can repair damage (Chapter Two) and have shown that they are able to carry out self-repair (see Chapter Three and Chapter Five). We presented behavioral

evidence for maintenance of redundancy coming from the scientific fields of the use-it-or-lose-it principle and the serial lesion effect (Chapter Two).

The data reviewed in Chapter Two suggest that self-repair may be mediated by plasticity mechanisms of the normal, intact brain. During the day plasticity is triggered by daytime activities like learning, which activate and reinforce the participating memory traces. In this chapter, we investigate whether self-repair is taking place during sleep, in particular autonomous self-repair driven by random cues. An indication of random brain activation is the unstructured order of dreams. Furthermore, internal random activation would probably also not be desirable during daytime, because it would interfere with external stimuli. Nowadays, there is a considerable amount of research available indicating a relationship between memory and sleep (Maquet, 1995, 2001; Terrence J. Sejnowski & Destexhe, 2000; R. Stickgold *et al.*, 2001). We focus on memory processes of memory maintenance and memory consolidation during sleep. Briefly, *memory maintenance* is the regularization of memory for proper functioning: strong memories are downscaled in order not to dominate memory, while weak but necessary memory traces are strengthened in order not to disappear. *Memory consolidation* is the post-processing of memory traces, during which traces may be reactivated, analyzed and gradually incorporated into the brain's long-term memory (Maquet, 2001).

The hypothesis of self-repair during sleep will be investigated as follows. We will demonstrate autonomous self-repair driven by random cues in a neurobiological plausible model of sleep: it extends the lifetime of memories. Further proof will be provided by a review of different sleep theories of memory maintenance and memory consolidation. We will investigate whether the theoretical and computational models of the sleep processes and their data support the self-repair ideas of redundancy and its maintenance. Maintenance will be investigated by identifying similarities with one or more steps of the three-step self-repair algorithm mentioned above. Autonomous self-repair with random cues will be supported, if they share the first step of randomly cued activation with the self-repair algorithm.

The exact model specifications and methods of analysis of model performance of the sleep model will be described in the next section. In Section 6.3, the exact description of the simulation and its results will be described. In Section 6.4, we will first introduce theories of memory maintenance and memory consolidation. The relationship between the two sleep processes and redundancy and maintenance of redundancy will be discussed in Section 6.5 and Section 6.6, respectively. In Section 6.7, we conclude with a summary of the similarities on the algorithmic level between memory consolidation and memory maintenance on the one

hand and self-repair on the other hand. The summary shows that self-repair can be carried out by the other two memory processes. It implies that self-repair can carry out maintenance and consolidation: they are interchangeable on the algorithmic level. We, therefore, hypothesize that it is possible that the other two processes are side-effects of self-repair.

6.2 The self-repair sleep model

The sleep model is an extension of the neural network model of Chapter Five. The main difference is an additional cortical-subcortical feedback projection to emphasize the interaction between the cortex and subcortical parts of the brain during sleep, for instance it is known that the thalamo-cortical loop generates the typical sleep waves (Lumer *et al.*, 1997; Steriade, 2001; Steriade *et al.*, 1993). The model is depicted in Figure 6.1.

The model represents an input, subcortical, and cortical part of the brain, where self-repair takes place in the cortical part. Each part consists of a map of 100 neurons. In the sleeping brain, the input component represents brain regions in which sleep oscillations are initiated like the brain stem core or forebrain structures (Steriade, 2001). The subcortical component of the model represents the dorsal thalamus, a region that receives its afferent signals during sleep from the reticular nucleus.

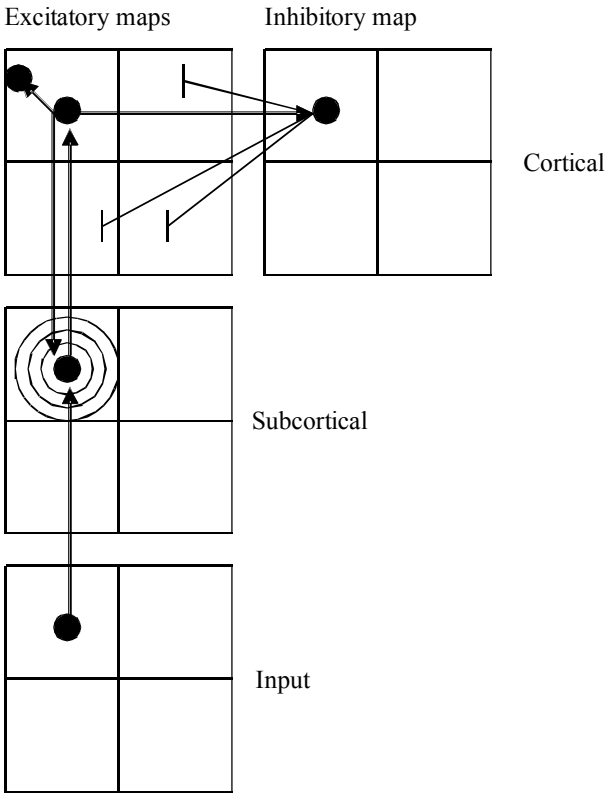


Figure 6.1. The neural network model (for a further explaination see text).

Regarding the connectivity tracts, in addition to the afferent signals from the reticular nucleus, the thalamus has both thalamocortical and corticothalamic projections. Afferent signals from sensory systems are not considered in the model since they are inhibited during sleep. The candidate brain regions for representation by the cortical component of the model should at least be involved in some form of learning or memory retrieval. In other words, there should be some form of plasticity, for instance long-term potentiation. Moreover, the modelled cortical area should have projections to and from the thalamus. In principle this could be any part of the cortex, such as the hand part of the primary somato-sensory cortex (SI).

The model, then, has six projections: (1) an input–subcortical projection, (2) a subcortico–cortical projection, (3) a feedback cortical-subcortical projection, (4) a cortico-cortical projection between pyramidal neurons, (5) a cortico-cortical projection from the pyramidal neurons to the interneurons, and (6) a cortico–cortical projection from the interneurons to the pyramidal neurons. The last two projections form a feedback inhibition loop of the cortical map. All projections connect neurons that tend to be spatially close. This type of connectivity results in clustered groups of neurons that we call neural assemblies. Each assembly or group of connected assemblies placed over the different maps form a memory representation or memory trace. The formulas of the connectivity and other details concerning it are given in Appendix A. Thus, the model is constructed in such a way that neuronal assemblies represent memory representations of neurons lying close together on a grid. This topology is similar to a cortical topological feature map such as the somato-sensory (auditory and visual) cortex or the motor cortex (Kandel et al., 1991).

The neuron model is a simplification of the MacGregor neuron (MacGregor & Oliver, 1974), which is based on the Hodgekin & Huxley model (Hodgkin & Huxley, 1952). It is a tradeoff between neural plausibility and computational cost; it models neural spiking behaviour and adaptation. For the computer simulations, the discrete time approximation formulas of MacGregor & Oliver (MacGregor & Oliver, 1974) were used. The model was implemented in Nutshell, the neural network simulator developed in our group (www.neuromod.org/nutshell). In all simulations we use the same values for the neural parameters. For a description of the parameters we refer to Appendices A and B, for other details of the model see the previous chapter.

6.3 Self-repair sleep simulation

The self-repair sleep simulation was as follows: a number of memory representations is stored in the network, then, after a given number of time steps the memory representations are damaged, then self-repair takes place, and finally the network is tested. We store memory patterns in the network with the synaptic density and synaptic weight function that is described in Appendix A. Storage is such that in each map four clusters are formed according to the cluster algorithm. Damage to the synapses is modelled by

$$w_{ij}(t+1) = \begin{cases} w_{ij}(t) - \varepsilon & \text{if } t \bmod c = 0 \\ w_{ij}(t) & \text{else} \end{cases}, \qquad (1)$$

where ε is noise and t is the integration time of a neuron. The constant c determines the time step at which damage was administered to the network. (This was at every 40th time step in the simulations.) Damage at that time step is represented by a stochastic perturbation ε that is added to each synapse. It is modeled by a random number drawn from a uniform distribution between zero and a maximum l that is indicated in the simulations by the parameter lesion size.

The self-repair algorithmic scheme is modeled by a three step process in which (1) neurons of the input map are randomly activated, (2) activation is allowed to spread over the rest of the network, and (3) connections are added or updated in the cortico-cortical projection of pyramidal neurons with a Hebbian learning rule as specified in Appendix B.

The initiation of oscillations is modeled by a spatially and temporally homogeneous Bernoulli process, which is a discretized analogue of the Poisson point process (Stoyan et al., 1997). This type of activation results in stimuli that are very likely temporally and spatially very dissimilar from the stimuli that stored the cortical memory representations. If we show that self-repair is possible with this type of stimuli, we show that it is also possible with stimuli that stored the memory representation, since the latter are most likely stronger associated with the stored memory representations. In the real brain self-repair will be carried out by stimuli that stored the memory representations, because the brain has been adapted to these stimuli during evolution and lifetime. Showing that self-repair can be carried out with the artificial stimuli used in the neural network of this chapter, is thus showing that self-repair can work with the stimuli of the real brain. The spread of activation is determined by the neuron model described in detail in Appendix B (equations B1-B4).

We use the Singer-Hebb learning rule (Singer, 1990) for (self-)repair. In this learning rule, a weight changes when the postsynaptic neuron is active at time *t+1*. A weight increases with amount μ ($0 \le \mu \le 1$) if the pre-synaptic neuron was active and decreases if it was inactive at time *t* given that the post-synaptic was active. In case the post-synaptic neuron was inactive at time *t*, weights remain constant. After Hebbian learning we apply an inward (in-star) normalization rule (see Appendix B equation B5-B7) to every neuron. The use of normalization is justified since simulation studies have shown that normalization is an intrinsic property of spike-based temporal learning (Kempter et al., 2001), a type of learning that can be found in the brain (Bi, 2002). Furthermore, direct evidence for normalization known as synaptic scaling has been found (Turrigiano et al., 1998; Turrigiano & Nelson, 2004). Normalization of the subcortical-cortical projection is not considered, because it is constant during a simulation.

After the network is initialized, the time scheme of the above described processes of lesion and repair is as follows:

1. At each time step self-repair takes place.
2. At each 10^{th} time step the cortico-cortico projection of the pyramidal neurons is tested for the number of assemblies with the cluster algorithm and for the total projection weight.
3. At each 40^{th} time step, damage is administered to all weights according to equation 1.

As we discussed in the previous section memory traces are represented by neural assemblies. To analyze network behavior we use a cluster algorithm that analyzes the neurons with respect to their connections. It is variant of a cluster algorithm by Xing & Gerstein (Xing & Gerstein, 1996a). The algorithm optimizes the number of clusters or assemblies such that each cluster is most strongly connected to itself, which is the 'self-connectivity' of a cluster. The average self-connectivity of all clusters in a map is the map-connectivity. In other words, the algorithm tunes the number of assemblies such that the map-connectivity is maximal. For more details of the cluster algorithm we refer to Chapter Five.

To show the effect of self-repair we performed two types of simulation, one with self-repair and another without self-repair. We ran both types of simulation for 4000 time steps. Figure 6.2 depicts the number of clusters retained in case of self-repair and without self-repair for each time step. The results are an average of 25 simulations.

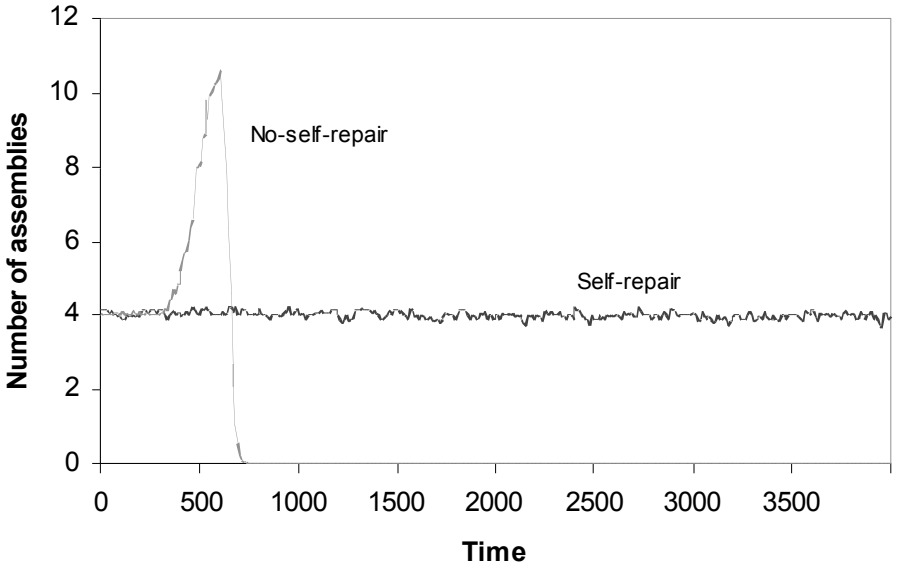


Figure 6.2. This figure illustrates the effect of self-repair. It shows a simulation with self-repair and one without self-repair. The results give the number of assemblies (y-axis) over time (x-axis). The figure clearly shows the effect of self-repair. For a more detailed discussion of the results see Section 6.3. All results of the figure are the average results of 20 simulations.

In case of simulations with self-repair and damage the network retains its original number of assemblies. Without self-repair all assemblies of the network eventually disappear. The degradation of the network can be divided in 2 phases, a first phase of a slowly increasing number of assemblies until a large number of assemblies is reached. This is followed by a second phase of a rapidly decreasing number of assemblies until none are left. In case of simulations of only damage and no self-repair, the network is stable for every four subsequent measurements. The explanation for this period of stability is that lesions are administered at every 40^{th} time step, while the network is tested at every 10^{th} time step. The results of both types of simulations clearly show the effect of self-repair.

6.4 Memory maintenance and memory consolidation

In Section 6.4.1 we introduce several theories of memory maintenance. In Section 6.4.2 we discuss theories of memory consolidation. An overview of theories of memory maintenance and memory consolidation is given in Table 6.1.

Table 6.1. Overview of the different sleep theories

The different sleep theories

Table 6.1

Sleep theory	Short description	References
Dynamic stabilization	Old memories have to be reinforcement to counteract forgetting over the lifetime	Kavanau (1996,1997)
Unlearning	Weakening of memories to counteract runaway and spurious attractors	Crick and Mitchison (1983)
Synaptic weakening	Downscaling of all synapses to counteract synaptic noise	Tononi and Cirelli (2003)
Neural regulation	Strengthening of weak memories and weakening of strong memories to counteract runaway attractors	Horn, Levy, and Ruppin (1998a, 1998b)
Skill improvement	Improvement of a skill	Karni et al. (1995); Shadmehr and Holcombe (1997)
Skill stabilization	Consolidation of a skill to interference of subsequent other type of skill acquisition	Robertson, Pascual-Leone, and Miall (2004)
Standard theory of systems consolidation	Reinforcement of newly acquired memories for consolidation	Squire, Cohen and Nadel (1984); Murre (1996)
Alternative theory of systems consolidation	Reinforcement of newly acquired memories for consolidation	Nadel and Moskovitch (1997)

6.4.1 Theories of memory maintenance

Several sleep theories of memory maintenance are in existence. One can distinguish the different theories by the way maintenance is carried out either by strengthening of connections, weakening of connections, or both.

A theory that proposes maintenance by connectivity strengthening is dynamic stabilization (Kavanau, 1996, 1997), where neural circuits are stabilized or maintained by extrinsic and intrinsic induced neural activity. This activity re-activates mechanisms that have an effect on synaptic efficacy. Reactivation is necessary to continue the synaptic change, because the mechanisms only have an effect for a limited amount of time. Kavanau proposes long-term potentiation (LTP) and gene expression together with facilitated entry as mechanisms. The effect of LTP lasts for days up to weeks, while the effect of genetic expression mediating the synthesis of new messenger ribonucleic acids and proteins lasts from weeks to months. LTP is a likely neural correlate of Hebbian learning (discussed in Chapter Two). The effect of gene-expression is the production of a higher amount of (among others transmitter) molecules in the cell-core. An increased amount of neurotransmitter does not automatically imply increased information transfer, because the entry of molecules into the terminals from where they are transmitted is limited. This is referred to as facilitated entry.

Thus, in order to affect information transfer, the limitation of synaptic entry in the terminals has to be changed. One such a mechanism proposed by dynamic stabilization is a change in the shape of synaptic terminals. Reactivation of phylo-genetic memories, memories created by genes, and onto-genetic memories, memories created by experience, can come about by external stimuli through daytime use. In that case, the reactivation is sufficiently frequent for stabilization. Memories not being reactivated by external stimuli are supposed to be reactivated by intrinsic spontaneous neural activation during sleep. This could be the case for phylo-genetic and ontogenetic neural circuitry both. Although Kavanau admits that he is not very certain about this matter, he sketches two scenarios of selection of memory types. The first possibility is the reactivation of all types of memory or in other words (a random sample of) the complete memory set. The second possibility is the reactivation of only the type of memory set that is not activated by frequent use. In the next section, we will say more about these two possible scenarios.

Sleep theories of memory maintenance proposing memory maintenance by the weakening of connections can be distinguished by their different objectives: to keep memory from overloading by removing old or parasitic memories (Crick & Mitchison, 1983), to suppress runaway processes in memory (Hopfield *et al.*, 1983), or to increase the signal to noise ratio of memories (Tononi & Cirelli, 2003). These authors argue that weight decrease is necessary for the regulation of memory in order for it to function properly or more optimized.

The theories of Crick and Mitchison (Crick & Mitchison, 1983) and Hopfield et al. (Hopfield et al., 1983) emphasize different goals of weight decrease. They do, however, share the same mechanism of unlearning. Crick and Mitchison (Crick & Mitchison, 1983) provide most of the theory while Hopfield et al. (Hopfield et al., 1983) were the first to demonstrate the mechanism in a computational model. The theory of unlearning by Crick and Mitchison (Crick & Mitchison, 1983) sketches a model of a cortical brain processing information that 1) is distributed over many synapses, 2) is robust, where in case of synapse loss the information is not completely lost, and 3) is superimposed: one synapse is involved in several pieces of information. If such a system is overloaded the overlap of stored memory patterns becomes too large, leading to a system that produces self-excitatory nodes that are mixtures of the stored memory representations (referred to in the literature as spurious memory representations). To prevent overloading, spurious patterns have to be removed. This is supposedly carried out by unlearning. The mechanism works as follows: random activation of the cortex (specifically the forebrain) will select the spurious patterns, which will then be unlearned by anti-Hebbian learning.

The theory of Tononi and Cirelli (Tononi & Cirelli, 2003) postulates that during slow wave sleep (SWS) there is synaptic homeostasis in the form of downscaling of synaptic weights (in our terminology weakening of connections). This optimizes memory performance, because it increases the signal to noise ratio at the neuronal level. That is, the noise in synapses accumulated during daytime due to potentiation will be suppressed and will fall under a baseline that silences them, while synapses coding memory will stay above the baseline.

A sleep theory of memory maintenance proposesing memory maintenance by way of the weakening and strengthening of connections is neuronal regulation (Horn et al., 1998a, 1998b). Crick and Mitchison presented the idea of a smarter mechanism than unlearning that does not select memories randomly, but 'knows' what to store and what to erase. Neuronal regulation is such a mechanism. It scales the weights of connections according to the measure of the basin of attraction strength of a memory representation.

To conclude, the purpose of memory maintenance is to regulate memory in order for it to function properly or for the maintenance of crucial memory circuits. The different theories of memory maintenance ascribe different functions to sleep with respect to memory. A clear example is memory strengthening during slow wave sleep by the theory of dynamic stabilization (Kavanau, 1996, 1997) and memory weakening by the synaptic homeostasis mechanism (Tononi & Cirelli, 2003) in the same period. We will later address this difference in Section 6.6.3.

6.4.2 Theories of memory consolidation

Memory consolidation is the post-processing of memory traces, during which the traces may be reactivated, analyzed and gradually incorporated into the brain's long-term memory (Maquet, 2001). Consolidation refers to processes occurring at different timescales (Squire & Alvarez, 1995) ranging from hours to years (decades) (Meeter & Murre, 2004). In this paper we will focus on consolidation taking place in humans from hours to days and also on consolidation taking place from months to years that is referred to here as systems consolidation.

Consolidation from hours to days, has been observed for instance in visual skill learning (Karni *et al.*, 1995) and motor skill learning (Shadmehr & Holcomb, 1997), in which after a night of sleep a particular skill is improved or is less vulnerable to interference by other skill learning (E. M. Robertson *et al.*, 2004). The latter is also referred to as memory stabilization. Skill improvement and stabilization of all types of skills can be regarded as

consolidation of procedural memory. A general theory as to how this type of consolidation is taking place in the brain does not exist, but there is an assumption in the field that during consolidation memory is transferred to other parts of the brain. Some empirical evidence supports this. For motor skill learning it has been established that during learning the skill seems to be moved from the pre-frontal regions of the cortex to the pre-motor cortex, posterior parietal, and cerebellar cortex structures (Shadmehr & Holcomb, 1997). Other experimental proof of movement of memory through the brain has been provided by Izquirdo et al. (Izquierdo *et al.*, 1997). They showed in rats for a step-down inhibitory avoidance task that the task dependence moved from the amygdala and hippocampus through the entorhinal cortex to the parietal cortex.

Systems consolidation is the consolidation taking place from months to years in humans and foremost explains the Ribot gradient. This gradient of retrograde amnesia, where recent memory traces are not available, was first proposed by Theodule Ribot (Ribot, 1881), who suggested that recent memories might be more vulnerable to brain damage than remote memories. This has indeed been found in experimental animals and patients with damage to the hippocampal memory system (Kim & Fanselow, 1992; Kopelman, 1989; Squire, 1992). It can be explained by assuming that memories are first dependent on a hippocampal memory system for their retrieval. Through consolidation they gradually become stored in the neocortex, making them independent of the hippocampal system (Squire & Alvarez, 1995; Squire *et al.*, 1984). This interpretation of memory consolidation is the conventional or standard theory of systems consolidation (Meeter & Murre, 2004). We will further discuss systems consolidation and its relation with self-repair in Section 6.6.2.

6.5 Redundancy

Redundancy, as we have argued in Chapter Two is provided by the connectionist properties of memory distribution and the possibility of neural groups to participate in multiple memory traces. To show that models of memory maintenance and memory consolidation have redundancy we will show that their theory and models are either embedded or implemented in connectionist models.

The theory of dynamic stabilization has not been implemented in a computational model until now (Kavanau, 1996, 1997). As we will argue in Section 6.6, however, the simulation of this chapter can be regarded as a possible implementation of the theory, because of its resemblance to the self-repair algorithmic scheme. In addition to its similarity on

algorithmic level, it also shares its ideas about how memory is embedded in a system of neurons and synapses. These ideas about memory lead to a similar conclusion as was made by the theory of self-repair. In the words of the theory of dynamic stabilization: "*From these studies and the foregoing, it is evident that the neural substrates for many functions are redundant, that they are widely distributed in cortical and subcortical areas, and that the areas may be multi-functional, involved in both cognitive and non-cognitive functions (pp.30)*" (Kavanau, 1997). With "*these studies*" is meant lesion studies and with "*foregoing*" is meant distributed interactive neural systems that can participate in several functions. The latter is similar to what the theory of self-repair is arguing about the typical connectionist property of shared information in different neural traces, discussed in the introduction of this chapter, which is providing the brain redundancy. The main point of the theory of dynamic stabilization for the theory of self-repair is that Kavanau (Kavanau, 1997) provides its claim of brain redundancy with extensive argumentation and illustrates it with many examples.

Weight decrease or unlearning has been implemented (and investigated by implementation) by Hopfield (Hopfield et al., 1983), Christos (Christos, 1996), and van Hemmen (Hemmen, 1997) in connectionist models. In addition to this, as was discussed in the Section 6.4 about maintenance, Crick and Mitchison (Crick & Mitchison, 1983) ascribe the brain connectionist properties. Amongst them was the property of redundancy that Crick and Mitchison (Crick & Mitchison, 1983) named robustness. Other memory properties they mention are the properties of distribution and superimposedness. The latter refers to the fact that a brain structure can have multiple functions (see Chapter Two). The multi-functionality of memory is similar to the typical connectionist redundancy discussed above by the theory of dynamic stabilization that was mentioned before in the introduction of this chapter. A difference being that the theory of self-repair does not take redundancy as an axiomatic property of the brain, but explains it in terms of the other two properties (see Chapter Two).

Tononi and Cirelli did not (yet) implement the theory of synaptic homeostasis (Tononi & Cirelli, 2003) in a model. Tononi, however, did built with Hill (S. Hill & Tononi, 2004) a very large scale visual thalamocortical model to investigate thalamocortical functioning in waking and sleep. Their model comprises detailed physiological properties as well as detailed anatomical organization at different levels: from the level of intrinsic cellular currents and synaptic conductances to that of the connectivity within and between cortical and thalamic areas. Although the synaptic homeostasis mechanism was not implemented in this model it is suggested that the model can be a starting point for investigating different theories of sleep. This illustrates how the sleep hypothesis is embedded in connectionist theory.

The theory of neural regulation has been implemented in a connectionist model see for example Horn et al. (Horn et al., 1998a) and Horn et al. (Horn et al., 1998b).

To our knowledge, there are no computational models simulating procedural consolidation, which is not surprising since there is no general theory proposing how it takes place in the brain. There are numerous models of systems consolidation of which some will be discussed in Section 6.6.2. To our knowledge they are mostly connectionist models implying that they all possess the property of redundancy.

All discussed theories and models of memory maintenance and memory consolidation theories can be modeled in a connectionist system. As Fenn et al. (Fenn *et al.*, 2003) put it: "*If performance is reduced by decay, sleep might actively recover what has been lost, presumably by an interaction between partially retained memories (words) and partially retained mappings that resulted from learning the word set (pp 616)*". Their remark was meant for memory consolidation, but it is evident that the same principle also applies to memory maintenance and repair of memories.

6.6 Self-repair: maintenance of redundancy

In the previous section we discussed one aspect of the self-repair theory, namely redundancy. In this section we discuss the other aspect of the self-repair theory, that is, the process of continuous repair or self-repair. In Section 6.6.1, we first go into the process the similarities between self-repair and memory maintenance. We will then in Section 6.6.2 list the similarities between self-repair and memory consolidation. Finally, in Section 6.6.3 review all empirical evidence of memory maintenance and memory consolidation that is also supporting self-repair.

6.6.1 Self-repair and memory maintenance

The dynamic stabilization theory of Kavanau (Kavanau, 1996, 1997) proposes that during sleep internally generated rhythms activate phylogenetic critical neural circuitry, after which it is reinforced by, for instance, long-term potentiation (LTP). One possible scheme, suggested by dynamic stabilization, purports that all memories have an activation probability, irrespective whether they are ontogenetic or phyologenetic memories. This is very similar to the self-repair scheme in which there is random selection of memories that is able to select any memory. Connections are, furthermore, updated with a Hebbian learning rule. The similarities between dynamic stabilization with the random selection scheme of memories and self-repair imply that the self-repair simulation of this chapter can be regarded as a simulation

of dynamic stabilization. The other scheme of dynamic stabilization selects only memories of the phylogenetic type. This scheme is easier than the random selection scheme, because to select such a specific type the cues have to be more informed than random cues avoiding runaway problems. With an informed cue is meant that it is strongly associated to a stored memory.

In Section 6.4 we suggested that unlearning is incorporated in a similar algorithm as consolidation, where memory representations are selected by random cueing. The only difference between the self-repair algorithm and unlearning algorithm is that instead of Hebbian learning anti-Hebbian learning is used, where neurons that fire together are decreased instead of increased. Unlearning does not seem to work well in suppressing runaway attractors in a system with ongoing learning and unlearning. One of the main reasons is that the newly acquired memories are those most preferred for unlearning (Christos, 1996; Meeter, 2003). In other words, they have the highest probability of being selected for unlearning (see Chapter 4 for an explanation).

As was mentioned in the previous section, neural regulation has been implemented in a connectionist model. It consists of two phases. A first phase in which the basin of attraction of the different memories is tested by randomly activating the network. A second phase in which the different memories are scaled according to their attractor strength by a normalization plasticity rule that is adapting the weights (Horn et al., 1998a, 1998b). There seem to be two differences with the self-repair algorithm, namely the time scheme of the different parts of the algorithm and the learning rule. The first is not a real difference, as we have argued in the previous section, self-repair models a process over time. We have modeled this by self-repair at each time step of the simulation. The time scheme of self-repair, however, can be modeled differently with self-repair at every 50^{th} or 100^{th} etc. time step. Critical for successful self-repair is that the amount of self-repair is in balance with the amount of damage, where the frequency of self-repair is only one parameter determining the balance (see Chapter Five for a more elaborate discussion). The second difference between self-repair and neural regulation concerns the learning rule. This is also not a problematic difference, as will be explained in section 6.6.3. Interesting to note, is that neural regulation may be able to substitute the learning rule used in the model of this chapter, because computational simulations have shown that it can to some extent rid noise from connections similar to self-repair.

In summary, the theory of dynamic stabilization, hitherto not implemented, can be regarded as an instance of the simulation presented in this Chapter. Two of the maintenance

theories share the randomly generated cues and therefore the random selection of memory representations with autonomous self-repair. In the other theories the rhythms driving the maintenance process play a pivotal role, but they do not have to be necessarily randomly generated. This is also the case with the self-repair theory. As was argued in Section 6.5, all theories can be regarded as connectionist theories and they, therefore, possess a similar connectionist spreading activation rule as self-repair. One theory, the theory of dynamic stabilization, shares with the theory of self-repair the assumption of a memory strengthening mechanism that might be Hebbian.

6.6.2 Self-repair and memory consolidation

Although neurobiological models for procedural consolidation do not exist, one assumes that Hebbian learning is part of the mechanism responsible for procedural consolidation (Muellbacher *et al.*, 2002). Several models exist for the consolidation of declarative memory. One model implementing the standard theory of consolidation entails a fast learning system, the hippocampus, that integrates newly acquired knowledge in the cortex, which is a slow learning permanent memory store (Alvarez & Squire, 1994; McClelland *et al.*, 1995; Murre, 1996). In the standard theory, the consolidation of (long-term) memory is by the strengthening of cortico-cortical connections. There are also computational models implementing another type of system consolidation, among them the multiple trace theory (Nadel *et al.*, 2000) that assumes consolidation is the strengthening of hippocampal-cortical connections.

Models implementing the standard theory of consolidation slowly incorporate memories in the cortex which eventually become independent from the cortex by a rehearsal or pseudo-rehearsal procedure. In the rehearsal procedure the memory representations themselves are rehearsed. In the pseudo rehearsal procedure a memory representation is retrieved by randomly cued activation (Alvarez & Squire, 1994; Meeter, 2003; Murre, 1996; Robins, 1996). After the memory representation has been activated, it is strengthened through Hebbian learning or a backpropagation learning rule. The algorithmic scheme of consolidation is thus exactly similar to the self-repair scheme. Consolidation models implement the idea that the cortical trace relies on the hippocampus until it is strong enough, thereby underlining the idea that cortical memory reinforcement is taking place. This is a difference with the self-repair theory. Although the emphasis of this thesis is on cortical self-repair (Chapter Two and Chapter Five), self-repair is not restricted to the cortex and may take place in other places of the brain.

The multiple trace model (Nadel & Moskovitch, 1997) implements another type of system consolidation. Cortical memory representations or traces are selected by random cues from the hippocampus. Traces are then created between hippocampus and cortex by Hebbian learning. Thus, the algorithmic scheme is the same as the first type of consolidation and also of self-repair. However, instead of weight strengthening in the cortex the emphasis is on the formation of hippocampal-cortical connections, although their model does not exclude the formation or renewal of cortico-cortical connections. Moreover, support for cortical memory reinforcement comes from another computational model of the second type of systems consolidation. This takes the plasticity of cortex as an assumption (Kali & Dayan, 2004). This being so they want to show with their simulations that the hippocampus is always necessary as an index system to associate an input pattern with a given (changing) cortical pattern. Their theory implies that the cortex is plastic, and therefore can carry out cortical self-repair.

The discussed models of systems consolidation, thus, show a great similarity to the model of self-repair. The consensus in the consolidation models regarding the activation cue is that patterns are selected by random activation. As to the learning mechanism, most of the models comprise weight strengthening through Hebbian learning. A significant difference, however, with the model of self-repair is that consolidation is supposed to strengthen new memories, while for self-repair the type of memory is not important. Moreover, the process of rehearsal that strengthens the cortical trace can also automatically strengthen other neural assemblies involved in the new cortical memory trace by Hebbian learning. This is made possible by the above mentioned connectionist redundancy of memory representations of participation of neural groups in multiple memory traces.

6.6.3 Experimental data of memory maintenance and memory consolidation supporting self-repair

The self-repair algorithm consists of three steps. The first and second step of the self-repair algorithm concerns the activation cue and the activation rule of the nodes. Together they determine the selection of memories. The second step concerns the spreading activation rule that is depending on the particular connectionist model. Relevant to the type of self-repair is whether the activation rule is similar for every node in the model, in which case random selection according to a uniform distribution is possible. For most connectionist models discussed in this Chapter this holds. We will, therefore, not consider the second step further and will only discuss the first and third step of the self-repair algorithm.

The first step of the self-repair scheme consists of the activation cue. Since we try to approximate a random selection of memory representations, we used a random cue. We simulated this in the model of this chapter by giving each input node an equal activation probability. Three computational models of maintenance, dynamic stabilization, unlearning, and neural regulation, and all computational models of systems consolidation use random cues for unlearning or consolidation. In the theory of dynamic stabilization, activation does not necessarily have to be random, but it is essential for dynamic stabilization that cues during the night are able to activate important phylo- and ontogenetic memory representations. The theory of unlearning assumes that random selection selects spurious memories. In case of synaptic homeostasis, random cues are used to determine the basis of attraction of memory representations. With consolidation, random cues are used to activate at least newly acquired memories.

Several intrinsic brain rhythms during sleep have been mentioned to drive maintenance and consolidation, viz. the theta rhythm of REM-sleep (Pavlides *et al.*, 1988), the field irregular sharp waves of non-REM sleep (Buzsaki, 1989), and ponto geniculo occipital (PGO) waves (Horn et al., 1998b). Of the first two waves it is not known with any certainty whether they can be regarded as random cue(s). Supposedly, they are involved in memory reinforcement (Kavanau, 1997). Only the ponto geniculo occipital wave is supposed to be random (Crick & Mitchison, 1983; Horn et al., 1998b). According to Sejnowski and Destehxe (Terrence J. Sejnowski & Destexhe, 2000) the rhythms during slow wave sleep seem perfect for consolidation. They suggest that slow wave sleep spindles that are low amplitude low coherent fast oscillations select or prime certain memory representations. The spindles are alternated with the typical slow wave pattern that has high amplitude and highly coherent slow oscillation, which stores the selected memories. Spindle periods are supposed to be brief compared to the typical slow wave pattern.

In the literature, different candidate rhythms have thus been suggested. It is, however, hard to establish whether they can be identified as the activation cues of self-repair and whether they are random.

The third part of the self-repair algorithm is the plasticity mechanism. With the dynamic stabilization theory and also the different systems consolidation theories, the emphasis is on reinforcing weights in particular by Hebbian learning. Evidence for memory reinforcement, therefore, will have to come from these theories.

The theory of dynamic stabilization provides the following data for memory reinforcement. It first mentions data of Roffwarg, Musio, and Dement (Roffwarg *et al.*, 1966)

of the requirement of spontaneous, repetitive excitations of neural circuits during REM sleep to facilitate circuit development and maintenance. Then, it mentions several other data confirming the requirement of spontaneous, stereotyped activations of neural circuits for maintenance in development (Changeux & Danchin, 1976; Hobson, 1989; Jacobson, 1991). Other supporting experimental data for memory reinforcement found in the adult are the data of circuit consolidation and reinforcement during REM sleep (Karni *et al.*, 1994) and NREM sleep (Pavlides & Winson, 1989; M. A. Wilson & McNaughton, 1994).

Memory consolidation is the strengthening or fixating of memory representations into memory. Any data supporting memory consolidation therefore support memory reinforcement as proposed by the self-repair theory. Supportive data have been found for different forms of consolidation of procedural memory, i.e. motor consolidation (Brashers-Krug *et al.*, 1996; Huber *et al.*, 2004; Karni et al., 1995; Muellbacher et al., 2002; Shadmehr & Holcomb, 1997), perceptual consolidation (Karni et al., 1994; Mednick *et al.*, 2002), consolidation of categorization (Gais *et al.*, 2000; Robert Stickgold *et al.*, 2000), and the consolidation of generalization of words (Fenn et al., 2003). For instance, Mednick et al. (Mednick et al., 2002) found that short naps with only SWS are sufficient for perceptual memory consolidation. There is no direct evidence that Hebbian learning is involved in procedural learning, only indirect evidence. For instance, for motor skill learning evidence is available that the primary motor cortex is involved in procedural consolidation (Muellbacher et al., 2002). Also LTP and LTD are observed after motor skill learning (see for example Hess et al. (Hess *et al.*, 1996) and Hess and Donoghue (Hess & Donoghue, 1996)).

Systems consolidation also supposes that memory is strengthened, but strengthening takes place in a larger time scale than procedural consolidation. The hypothesis of systems consolidation models as described in the previous section is that strengthening of memory takes place in the cortex. Meeter and Murre (Meeter & Murre, 2004) provide evidence for this coming from neuropsychology, fMRI, and neurobiology. Important evidence from neuropsychology is the Ribot gradient mentioned in Section 6.4.2 that has been found many times in humans (Albert *et al.*, 1981; Beatty *et al.*, 1988; Kopelman, 1989; Kritchevsky & Squire, 1989; Squire *et al.*, 1989). The Ribot gradient has, furthermore, been found in different kinds of animals (Squire, 1992). One has to keep in mind that in animals it is of a smaller timescale, ranging from weeks to months instead of months to years. Another interesting finding from neuropsychology, where the emphasis is on cortical strengthening, is that in case of reversible damage of retrograde amnesia, recovery can take place after a night's sleep (Whitty & Zangwill, 1977). This supports the notion of self-repair during sleep. Meeter

and Murre (Meeter & Murre, 2004) mention several convincing neurobiological studies of a temporary role of the hippocampus and the strengthening of cortical traces. For instance, the experiment of Frankland et al. (Frankland *et al.*, 2001) shows that mice with impaired cortical long term potentiation are able to acquire new memories, but unable to retain them. This suggests that the hippocampus is able to acquire new memories, but that the cortex is unable to retain them, because cortical traces cannot be strengthened by long term potentiation. Other relevant studies are from Bontempi et al. (Bontempi *et al.*, 1999) and Izquierdo et al. (Izquierdo et al., 1997). For more details and a discussion see Meeter and Murre (Meeter & Murre, 2004).

As we mentioned at the end of the section about theories of maintenance, there seems to be a discrepancy between reinforcement and normalization, in particular the downscaling of synaptic homeostasis. Although the two processes seem to be contradicting each other, this is not necessary so and they may even be complementary as was also suggested by Tononi and Cirelli (Tononi & Cirelli, 2003). As was investigated in the previous chapters, neither a self-repair process with only Hebbian learning (Chapter 3) nor simple normalization as is used in the model of this chapter can work (not reported data). Otherwise, memory problems like for instance spurious and runaway memories will arise. They both have to be present for successful self-repair. These processes can take place (almost) simultaneously in a single time step as is the case in the computational simulations of this chapter. The simultaneously taking place of the two processes is also suggested by simulation studies showing that normalization is an intrinsic process of spike time dependent plasticity. They may, however, take place in separate periods too. An example of the latter comes from Tononi and Cirelli (Tononi & Cirelli, 2003) who suggest that potentiation takes place during daytime and normalization takes place during sleep. Until now other possibilities like potentiation and normalization during the same sleep stage are not excluded by the empirical data.

There thus is a wealth of data provided by different theories of memory maintenance and memory consolidation for memory reinforcement. In addition other plasticity mechanisms, not necessarily taking place during sleep, can play a role in self-repair and might be complementary to memory reinforcement.

6.7 Discussion

In this chapter, we investigated the hypothesis whether self-repair through maintenance of redundancy takes place during sleep. The extended research question was whether

autonomous self-repair with randomly cued activation takes place during sleep. We will first address the first research question and then the extended research question.

To investigate the first research question we constructed a neurobiological sleep model in which we demonstrated self-repair with a three step algorithm. We opted for an overall fairly abstract connectivity with some biological plausible features to cover a thalamo-cortical model, but which also leaves open the possibility, for instance, for a hippocampal-cortical model. We reviewed, furthermore, two sleep processes of memory maintenance and memory consolidation. The similarities between self-repair on the one hand and the two sleep processes on the other hand can be summarized as follows. All theories of memory maintenance and memory consolidation are embedded in a connectionist framework. Therefore, all these models possess redundancy. They share parts, one or more steps, with the three step self-repair algorithm. (1) With respect to the first step, the cue, if neurons are activated by a random cue the self-repair algorithm is similar to many memory consolidation algorithms (Alvarez & Squire, 1994; Murre, 1996; A. Robins & McCallum, 1998). In this case the activation cues are supposed to be associated with newly acquired memories. If the activating cues are associated with phylogenetic old memory circuits critical for the functioning of an organism that are not activated during daytime (Kavanau, 1996, 1997), it is similar to memory maintenance. (2) With respect to the second step, the spreading activation rule, since all theories of maintenance and consolidation are implemented or embedded in connectionist models, the general way of how memories are selected is also similar to the model of self-repair. (3) As far as the plasticity mechanism is concerned, the third step of the self-repair algorithm, all theories of memory consolidation assume a Hebbian like plasticity rule. Some theories of memory maintenance also assume such a Hebbian plasticity rule. Other maintenance theories assume different plasticity rules, but are not contradicting self-repair and can be complementary to the Hebb rule. A summary of similarities between the self-repair theory and theories of memory maintenance and memory consolidation is given in Table 6.2. The conclusion is that both the model and the data support the hypothesis of self-repair during sleep.

Table 6.2. Summary of similarities between the self-repair theory and theories of memory maintenance and memory consolidation (a + is a similarity, a – is a difference and a ? is that it is unknown).

Similarities between self-repair theory and theories of memory maintenance and memory consolidation
Table 6.2

Sleep theory	**connectionist theory: redudancy + spread of neural activation**	**Activation by random stimuli**	**Hebbian Plasticity rule**
Dynamic stabilization	+	+	+
Unlearning	+	+	-
Synaptic weakening	+	?	-
Neural regulation	+	+	-
Skill consolidation interference	+	?	+
Skill consolidation consolidation	+	?	+
Standard theory of systems consolidation	+	+	+
Alternative theory of systems consolidation	+	+	+

The second research question whether autonomous self-repair with randomly cued activation takes place in the brain, was simultaneously addressed with the first research question. We constructed a neurobiological sleep model in which we demonstrated *autonomous self-repair with randomly cued* activation. The review, furthermore, shows that many theories and models of the discussed sleep processes use random activation. Also in case of dynamic stabilization the activation cues can be random, because they can as well activate old phylo-genetic memories. Thus, the model and the data also support the extended hypothesis of randomly cued self-repair during sleep.

Despite the similarities between the three processes of memory maintenance, memory consolidation, and self-repair, differences can be noted also. The main difference lies in their objectives: maintenance aims to stabilize memory, consolidation aims to consolidate newly acquired memory, and the goal of self-repair is to repair (small) damage. Though the objectives may be different, the many similarities in the algorithmic scheme suggest that one process can carry out other processes. For instance, consolidation may be able to repair memory. As a consequence, self-repair may be a by-product of consolidation and maintenance, or they can all be aspects of a single mechanism.

It is hard to determine what the main process of the brain is and what the by-product, as it is already hard to validate any of these processes. Even for the most advanced research topic of memory consolidation, there is not yet definite proof. This has lead Hairston and

Knight (Hairston & Knight, 2004) to suggest that the 'reverberating circuits' may be a by-product of design for another purpose (than consolidation). Moreover, according to Meeter and Murre (Meeter & Murre, 2004) consolidation may continue for years and even decades. If this is the case, consolidation will be very similar to the theory of dynamic stabilization, where memories have to undergo stabilization or fixation for the rest of the lifetime. Thus, if the brain would have a main process, all possibilities are still open. A speculative argument that this may be self-repair runs as follows: this chapter and other chapters show that self-repair can provide the brain with the advantage of extending memory lifetime. It can be the immune system of the brain. A memory system possessing this property provides it with a great evolutionary advantage. It may be that evolution first selected organisms possessing this property and that later, as a side effect, the same mechanism or slight variations of it could be used for the maintenance of old phylogenetic memory and/or for integrating newly acquired memory. The theory of self-repair is simple and may be favored by Occam's razor, as it explains strengthening of any type of memory and not just of one particular type of memory. Whatever the answer may be, this chapter shows that there is evidence for a self-repair process during sleep that deserves further investigation.

Appendices

Appendix A. Connectivity of the model

The typical connection density of the network is shown in Figure 6.3.

Table 6.3. This table shows the connection density between the different maps.

Connection density function

Table 6.3

Projection	Probability of a connection within an assembly								Probability of a connection between assemblies
From map -To map	r 0	1	2	3	4	5	6	7	r 0
sensory - subcortical	1.0	1.0	1.0	1.0	0	0	0	0	no inter-connectivity
subcortical - cortical	1.0	1.0	1.0	1.0	0	0	0	0	no inter-connectivity
cortical - subcortical	1.0	1.0	1.0	1.0	0	0	0	0	no inter-connectivity
cortical - cortical	0	1.0	1.0	1.0	1.0	1.0	0	0	no inter-connectivity
cortical - cortical-inhibitory	0	1.0	1.0	1.0	1.0	1.0	0	0	no inter-connectivity
cortical-inhibitory - cortical	0	0	0	0	0	0	1.0	1.0	no inter-connectivity

The first column indicates the projection. The second column indicates the probability of a connection between a neuron from the from-map and a neuron from the to-map within an assembly depending on the radial distance *r*. The third column indicates the probability of a connection between a neuron from the from-map and a neuron from the to-map between assemblies. Every map has the same size so that every neuron of the from-map has a corresponding neuron in the to-map. The distance between these neurons is zero. The distance between the neuron in the from-map and other neurons in the to-map is calculated relative to the corresponding neuron of the to-map. The table shows the typical connectivity between assemblies. Neurons in the cortical map are not connected to themselves. There is full connectivity within an assembly in a certain radius *r*, for instance sensory neurons are connected to subcortical neurons of their assembly in a radial distance of three. There is no connectivity between assemblies, except when indicated (see for example the simulations of Figure 6.3). Cortical neurons have (inhibitory) connections with inter-neurons that have a radial distance of 6 and 7.

The weight of a connection is determined as follows. An average weight of each connection is calculated by dividing 1 by the number of connections in an assembly. We call this number the average (weight). The weight of a connection is further calculated by adding a

random number drawn from a uniform distribution *U(*min, max*),* where min is -0.5*average and max is 0.5*average. Each weight is multiplied by a constant *c* that controls the neural activity in a map. This constant is different for the different projections. It is five for the cortico - cortical-inhibitory projection, eight for the cortico-inhibitory – cortical projection, and one for all other projections.

Appendix B. The neural model and plasticity rule

The model neuron of the neural network is a simplification of the MacGregor neuron (MacGregor & Oliver, 1974), which in turn is derived from the Hodgkin-Huxley neuron (Hodgkin & Huxley, 1952). The model is a tradeoff between neural plausibility and computational cost of simulating neuronal spiking and adaptation. In numerical simulations, the state of the neurons updated in discrete time steps, which in our simulations lasted two milliseconds.

The membrane potential U is dependent on the sodium, potassium and chloride currents over the membrane. It can be described in the following differential equation:

$$\frac{dU}{dt} = -\delta U - g_k(U - U^k) - g_{ex}(U - U^{ex}) - g_i(U - U^{in}) - U^0. \quad \text{(B1)}$$

Here $-\delta U$ is the leak current, g_{ex} the excitatory conductance, E_{ex} the natrium reversal potential, g_i the inhibitory conductance and E_i the chloride reversal potential. The parameter governing the leak current is different for each map and given in Appendix D.

Adaptation is modeled by the potassium conductance g_k with the following equation:

$$\frac{dG_k}{dt} = \frac{-g_k}{\tau_{Gk}} + bS, \quad \text{(B2)}$$

where S is the dichotomous spiking variable. The time constant τ_{Gk} differs for each map and specified in Table 6.4, along with other neural parameters.

Excitatory and inhibitory input to the *i*'th node is a linear summation of weighted inputs to that node

$$g_{ex/in} = \sum_j w_{ij} S_j, \quad \text{(B3)}$$

where w_{ij} is the weight from node j to node i, and S_j is the dichotomous spiking variable of node j. The weight w_{ij} is negative in the case it is inhibitory.

Table 6.4. This table shows the values of the neural parameters of each map.

Table Neural parameters

Table 6.4

Neural parameter	Map Sensory	Sub-cortical	Cortex pyramidal	Cortex interneurons
τ_U	0.5	7.5	0.1	10
τ_{Gk}	0.5	7.5	0.1	10
b	0.5	7.5	0.1	10

The model neuron emits a spike every time the membrane potential U crosses the threshold Θ.

$$S = \begin{cases} 0 & if\ U < \Theta \\ 1 & if\ U \geq \Theta \end{cases} \tag{B4}$$

S is a dichotomous variable with a value 1 if a spike is emitted and 0 otherwise.

In the model the Singer-Hebb learning rule (Singer, 1990) is applied. In this learning rule, a weight changes when the postsynaptic neuron is active at time $t+1$: a weight increases with amount μ $(0 \leq \mu \leq 1)$ if the pre-synaptic neuron was active and decreases if it was inactive at time t. When the post-synaptic neuron was inactive at time t weights remain constant. This results in the following learning rule:

$$\Delta w_{ij} = \begin{cases} \mu & if \quad S(j,t) = 1 \quad and \quad S(i,t+1) = 1 \\ -\mu & if \quad S(j,t) = 0 \quad and \quad S(i,t+1) = 1 \\ 0 & else \end{cases}, \tag{B5}$$

where Δw_{ij} is the change in weight of the connection from node j to node i. The neuronal output S is determined by equation (4). To keep the weights between bounds of 0 and 1 we apply the following update rule for the weights:

$$w_{ij}(t+1) = \min(\max(w_{ij}(t) + \Delta w_{ij}, 0), 1). \tag{B6}$$

We use the following inward normalization rule:

$$w_{ij}(t+1) = \frac{w_{ij}(t)}{\sum_j w_{ij}(t)}, \tag{B7}$$

where the weights of a neuron are synchronously updated.

For the computer simulations, the discrete time approximation formulas of MacGregor & Oliver (MacGregor & Oliver, 1974) were used. The model was implemented in Nutshell, the neural network simulator developed in our research group (www.neuromod.org/nutshell).

7 Discussion

7.1 Introduction

This thesis had two main goals. One goal was to show that self-repair through maintenance of redundancy is a possible process taking place in the brain. Another goal was to lay the foundation for models of brain recovery after damage. To achieve the first goal we constructed a connectionist self-repair model based on neurobiological and behavioral empirical data, where redundancy resides in the connectivity and self-repair is carried out by plasticity mechanisms. In a simple connectionist model we demonstrated that guided self-repair, an easy type of self-repair, can work extending the lifetime of memory representation. In that same model we also demonstrated autonomous self-repair. Autonomous self-repair in artificial neural networks is difficult, because it uses random cues to activate memory representations. Since this type of cues is not associated with stored memory representations, there is no control over which memory is repaired and the amount of times it is repaired. In the brain an easier type of self-repair using more informed cues is used, which is closer to guided self-repair, may take place, because the cues of the environment have shaped our brain during evolution and lifetime. In other words, we argued that a more difficult type of autonomous self-repair is possible in a connectionist system than might actually be taking place in the brain. This suggests that provided the brain is a connectionist system self-repair is possible in it. In the rest of thesis we, therefore, focused on autonomous self-repair with random cues. We derived general properties of autonomous self-repair (Chapter Four) of which the most relevant one for self-repair in the brain is that self-repair is much more stable with many memory representations. We demonstrated it in a neurobiological detailed model (Chapter Five) and we argued that sleep may be the time in which it is taking place (Chapter 6).

Did we proof with these simulation models that self-repair is taking place in the brain? We proved that we can build a model implementing the idea of self-repair through maintenance of redundancy based on empirical data. As we mentioned, however, in Chapter One these model have limitations. A first limitation is that it is impossible to incorporate all known properties of the brain due to computational cost. A second limitation is that we do not know the essential computational properties of the brain that have to be in the model. Thus the answer to the question whether we can prove self-repair in the brain with these models is no.

We demonstrated that given the available data it is possible. Showing it in a more neurobiological detailed model makes it more likely. It could be that the simple model already possesses the essential properties of the brain and we did prove it. It could also be that we are far from understanding the brain and completely different models are needed in which to demonstrate self-repair.

For proving self-repair in the brain, it is easier to investigate self-repair directly in the brain than acquiring a full understanding of the computational properties of the brain. We proposed in Chapter Two that the model of autonomous self-repair with small periodic diffuse lesions represents aging of the normal adult brain (Chapter Two). A first investigation is to identify in the brain these intermittent diffuse lesions. We will do that in the next section (Section 7.2). In Section 7.3, we will discuss a possible experiment testing the self-repair hypothesis.

The second goal of this thesis was to lay a foundation for models of brain recovery after injury. The models of chapters Five and Six are a starting point for such models. These models are the state of art in the tradeoff between neurobiological detail and computational cost. Moreover, we showed that autonomous self-repair is possible in these models, which is a difficult type of self-repair to model. It is, therefore, likely that we are also able to model the other types of self-repair. We will discuss models of brain recovery after injury further in Section 7.4. Adapting the models of autonomous self-repair to model brain recovery is one possibility for future research. Other possible future research topics will be discussed in Section 7.5.

7.2 Small damage: The necessity of self-repair

An assumption in this thesis is that if autonomous self-repair with small diffuse intermittent lesions models normal aging is the existence of small damage. As is shown in this thesis self-repair itself can also be damaging if it is not in balance with damage (Chapter Five). Consequently, without any damage autonomous self-repair is unnecessary and might be harmful to a memory system. The existence of damage is thus a vital assumption of the self-repair theory. We argued that aging may be accompanied by a process of damage, which self-repair counteracts (Chapter One). There is evidence on different levels suggesting that during aging memory is damaged. For instance, at the behavioral level different kinds of tests show a decrease of memory performance (Christensen, 2001; Grady & Craik, 2000; Phillips & Sala,

1996). At the anatomical level there is evidence of gray and white matter damage (Resnick et al., 2003; Salat et al., 1999).

Damage during aging is probably not one process, but represents many processes with different causes. We distinguish between external causes and internal causes of damage. External causes of damage come from outside the brain, while internal causes are intrinsic to the brain's memory system.

There are several external causes of damage. A first one is oxidation (Barja, 2004; Harman, 1956). Experiments suggest that oxidation has an impact on learning and memory (Forster *et al.*, 1996; Fukui *et al.*, 2001; Urano *et al.*, 1998). Other external damage may be caused by other toxics such as alcohol (Collins *et al.*, 1998). Another external source of damage may be brain hemorrhage due to sports, like heading in soccer (Downs & Abwender, 2002; Matser *et al.*, 1998; Matser *et al.*, 2001; Stephens *et al.*, 2005) and boxing (Chappell *et al.*, 2006; Clausen *et al.*, 2005; Erlanger *et al.*, 1999; Mendez, 1995; Ryan, 1998; L. Zhang *et al.*, 2006). There is a debate on whether these sports lead to brain damage (see for example Butler (R. J. Butler, 1994; Porter & O'Brien, 1996; Rutherford *et al.*, 2003)). One possible explanation is that the brain has a very strong self-repair capacity, because it is clear that in some cases there is damage to the brain after incidence, but that it is transient (Webbe & Ochs, 2003). Another possible explanation, is that damage is not noticeable behaviorally or too small to be spotted by behavioral tests, but can be detected by a physiological measure (L. Zhang et al., 2006). This type of damage is not only caused by participating in sports, but can happen in daily life accidentally just by bumping your head.

The other type of damage is internal damage. Internal brain damage is due to the fact that the brain, in particular the cortex, is plastic. The existence of internal brain damage is more speculative. Nonetheless, several researchers have taken internal damage due to plasticity as an assumption. Tononi and Cirelli (Tononi & Cirelli, 2003) speculated that LTP enhanced synapses that were not involved in the newly learned memory trace lead to noise or damage. Others researchers stating that a plastic brain leads to memory errors were Kali and Dayan (Kali & Dayan, 2004). In their theory, they alleged that because the cortex is plastic the hippocampus has the function of an index system to keep track of the moving memories in the cortex. Of course there is ample evidence of a plastic brain as was indicated in many places of this thesis. There is, however, no empirical evidence that plasticity in the normal intact brain causes a decrease in memory performance. On the other hand, from a theoretical point-of-view of neural network modeling it is hard to imagine, given that memory representation are overlapping and are sharing parts, that there would be no noise in the system.

The existence of internal damage is arguable. It might be hard to measure in the brain, since the brain may have mechanisms to counteract the internal noise. Familiar mechanisms are the so called homeostasis or normalization mechanisms, but it can also be homeostasis mechanisms as suggested by Tonino and Cirelli (Tononi & Cirelli, 2003) or the self-repair mechanism, as proposed in this thesis. The existence of external damage is more difficult to refute. We mentioned several causes of which it is likely that people do encounter them during life. The issue is more whether the damage has the uniform distribution as we assumed in this thesis, because it is possible that different types of damage hit particular regions of the brain or that different areas of the brain differ in vulnerability to specific types of damage. If the distribution of damage is uniform, autonomous self-repair can counteract this type of damage, as we have shown in this thesis. If, however, damage is not uniform, but targets for example particular areas of the brain, autonomous self-repair is probably not able to repair damage and needs to be more informed of where to repair in the brain.

7.3 Testing the self-repair hypothesis

In Chapter Two, we reviewed data from different scientific fields that supports the self-repair hypothesis. They supported the concepts of redundancy and maintenance at different levels: from the neuronal level to the behavioral level. Since autonomous self-repair with diffuse lesions models aging, the most straightforward research to test autonomous self-repair empirically is longitudinal research. The problem with this research is that data is very noisy and suffers from the disadvantage of having many confounding variables. Moreover, research would take many years and is very hard to control, because human subjects will not always behave accordingly to the experimental set-up during the time of the experiment.

The serial lesion effect is a better controlled experiment than longitudinal research. There are rats specially bred for research and are for instance genetically very similar. They can, furthermore, be raised under the same conditions. Unfortunately, the results of the serial lesion effect experiments were not unambiguous, since in one of the last discussed experiments (Ramirez et al., 1999) of Chapter Two only the axonal sprouting after the second lesion lead to behavioral observable differences. We, therefore, recommended to extend this experiment with more stages to obtain clearer results. Even with the extension the serial lesion effect experiment is not the best experiment to test the self-repair hypothesis, because the size of the lesions may be too large. Large lesions have as a disadvantage that they cannot be neurally restituted, but only neurally compensated. By definition neural restitution will

reinstate the memory representation with the original neural tissue, while with neural compensation other neural tissue will take over the function(s) of the original tissue. In case of neural restitution we can be certain that redundancy is reinstated. With large lesions self-repair may be by neural compensation, which can lead to decreased redundancy. Another disadvantage of large lesions is that they might have as a side-effect that entirely different plasticity mechanisms will carry out repair than in case of small lesions that occur during aging. The latter is only a disadvantage if we want to test the hypothesis that autonomous self-repair is modeling normal aging, where it is stated that normal plasticity mechanisms carry out self-repair. Smaller lesions can be obtained by hemorrhage that for example mimics the lesions due to boxing or heading with soccer. Smaller lesions could also be obtained by administering neuro-toxins to the rats. Another important parameter in this experiment would be the amount of repair. An obvious way to do this is to keep one group of rats in a stimulus rich environment and another group in a stimulus poor environment or normal environment. To enrich the environment the particular group of rats can be given training for different tasks. To measure performance of the different groups of rats, performance on trained tasks may be taken as measure, but preferable a battery of tasks should be taken measuring more the full spectrum of behavior that rats are able to perform. This increases the chance to detect possible effects of self-repair and damage.

The serial lesion effect experiment is a good experiment to test the self-repair theory, because it possesses the two components of the self-repair theory, namely redundancy and maintenance. Differences in redundancy can be tested by raising two groups of rats in different environments during their development. One group can be raised in a stimulus poor environment and the other in a stimulus rich environment. As the rat group in a stimulus poor environment will probably still receive stimuli, self-repair in those rats will still take place. We, therefore, speak of the reduced self-repair rat group. To further reduce the amount of self-repair one could use knock-out techniques to impair learning (Bartoletti *et al.*, 2002; Linnarsson *et al.*, 1997). To keep the experiment simple and reduce the number of experimental groups, we will leave out this option for the moment. Each group is further split into two groups to undergo the serial lesion experiment as described above. Thus, the experiment has a total number of experimental groups of four. Expected is that the group raised in a stimulus rich environment, which is expected to undergo self-repair, will perform best. The group raised in a stimulus poor environment, which is expected to undergo reduced self-repair, is expected to perform worst. Interesting would be which of the two other groups performs best. Is it the group raised in a stimulus rich environment with reduced self-repair or

the group raised in a stimulus poor environment with self-repair? If the first group performs best this would imply that redundancy is very important to withstand small serial damage. In case the second group performs best, self-repair during aging is the best protection against small serial lesions.

The self-repair theory is invalidated if all four groups perform similarly or in case the the group raised in a stimulus rich environment with self-repair does not perform best. The self-repair group has to perform best, because the self-repair theory does a very particular prediction for this group, namely it predicts an *interaction effect* between redundancy and self-repair. On the one hand maintenance keeps redundancy at some level or even enhances it. On the other hand, in case of more redundancy there is more information present in the brain to guide repair processes, thereby increasing the probability for successful self-repair. Such an interaction effect was for instance found in data of cognitive reserve, where data suggested that brain use leads to more cognitive reserve, but data suggested also that for a healthy brain, with more cognitive reserve, it was easier to stay healthy.

The in this section described serial lesion experiment can (in-)validate the self-repair theory. In Chapter Two we reviewed other experiments supporting the self-repair hypothesis. We will argue in the last section that self-repair is not invalidated if one of those experiments will yield a negative result.

7.4 Application of models of self-repair: models of brain recovery

In this thesis we investigated mostly small diffuse lesions that is counteracted by autonomous self-repair. The self-repair model is able to model more than this particular type of lesion and particular type of self-repair. Lesions can vary in size and place. Self-repair can be guided or supervised instead of unsupervised as used in this thesis. In case of supervised self-repair, the stimuli administered to the network are by definition strongly associated with the stored memory representations in the network and we have more control over how much each memory representation is repaired. One can imagine that there is a continuum of stimulus types between the randomly generated stimuli of autonomous self-repair and the stimuli of supervised learning. We modeled one particular type of lesion and one particular type of self-repair scheme. The framework of self-repair, however, can be easily applied to the abnormal impaired brain by using different type of lesions and different types of self-repair.

In Chapter One, we introduced a model of recovery by Robertson and Murre (I. H. Robertson & Murre, 1999). Based on an elaborate literature study they distinguished in this

model a triage of recovery: autonomous recovery, guided recovery, and compensatory recovery. This distinction was based on lesion size, where the lesion size increased from autonomous recovery to compensatory recovery. The relationship between the self-repair model and the model of triage of recovery was that autonomous self-repair was a model for autonomous recovery and supervised self-repair was a model for guided recovery. If the lesion size was small, there was sufficient information present in the brain to retrieve the memory representation with a cue that is not strongly associated with it. If the lesion size was large, there was insufficient information present in the memory representation and only a cue strongly associated to it was able to retrieve it. In both cases of lesion size, it was assumed that the damaged neural region, where the memory representation resided, can be repaired. Here, we can speak of neural restitution (see Chapter One, Introduction). In the case of compensatory recovery, no stimulus was able to retrieve the damaged memory representation and the behavioral pattern belonging to it had to be carried out by another neural area. This can be called neural compensation. If no neural area was able to replace this behavior the only way to carry out a particular behavioral task that relied on functioning of the damaged area was by behavioral compensation.

Another way to clarify the relationship between the triage of recovery and self-repair is the explanation of the relationship between autonomous self-repair in artificial neural networks and autonomous self-repair in the brain. In Chapter One we explained a difference in meaning between them. In artificial neural networks, autonomous repair was repaired with randomly generated stimuli that were not associated with any stored memory representation. In neural networks of the brain, autonomous self-repair was repair with more informed non random stimuli, i.e. they were associated to some stored memory. The similarity between both is that we can speak of unsupervised repair, because there is no *external* control over the variable which memory representation is repaired and the variable of the number of times it is repaired. With supervised self-repair one can control these variables. This similarity between autonomous self-repair in the brain and in artificial neural networks shows how to apply our model of autonomous self-repair to recovery after brain damage. If we want to model guided recovery or rehabilitation, the important parameter to adapt is the stimulus batch we provide to the model. With the stimulus batch we can control which memory representation is repaired and the number of times it is repaired.

The models of this thesis allow us to investigate easily the following variables: the stimulus with which the network is recovered, stimulus magnitude and frequency, the type of damage that can differ in amount and place of lesion in the network, the initial connectivity

like the relative amount of inhibition and excitation, and learning rule differing in type and parameter values. We can investigate the limits of the different models. For instance, with autonomous self-repair we can start with a certain learning rule and investigate the lesion size that it still can repair. When the critical limit size is exceeded, it can then be investigated whether another learning rule can succeed. Yet another possibility is to investigate temporal effects of different lesion types and different types of repair. For instance, we might investigate the effect of guided self-repair alternated with autonomous self-repair. The sessions of guided self-repair may represent a rehabilitation scheme, while autonomous self-repair may represent the time between two rehabilitation sessions.

With the extended models and their parameters we can try to investigate the huge amount of animal and human data of recovery after damage that is present nowadays. It can be investigated whether this model with the mentioned parameters, is able to model the different type of data as described by Robertson and Murre (I. H. Robertson & Murre, 1999), ranging from recovery after mild lesions to maladaptive repair. Some other interesting data to model can be animal data of reorganization like for example the data of Kilgard and Merzenich (Kilgard *et al.*, 2001), brain reorganization of Braille readers (Elbert & Rockstroh, 2004), differences in the brain because of music (Münte et al., 2002), effect of rehabilitation therapies like constrained-induced therapies (Taub et al., 2002), motor reorganization like writers cramp (Quartarone *et al.*, 2003), etc. The amount of available data is huge. It is probably useful to do a literature review that classifies the data with the variables of the model, as for instance classifying data in which the type of stimulus is most important and data in which the learning rule is most important. If necessary, we can go beyond the present model and extend it to brain structures that model emotion or allow the addition of new neurons modeling stem cell treatments.

The development of models of recovery after damage can unify the different data of recovery after damage similar to the data of (small) diffuse lesions and the recovery from this as discussed in this thesis' models. Models of recovery after larger lesions can integrate data of different fields and put them into a new perspective by giving one coherent explanation. Furthermore, they provide a common terminology and framework with which researchers from different fields can communicate. Theories expressed in natural language are also able to do this, but (computational) models can explicate underlying, sometimes hidden, assumptions present in theories expressed in natural language. Furthermore, by modeling, shared (algorithmic) components are easier identified, as is demonstrated in Chapter Six for models of sleep. Other advantages of models are that they allow us to have full control over the

experiments and give us the opportunity to investigate extreme cases that could not be so easily investigated in animals or humans.

Models also have restrictions. One major limitation is that models cannot provide definite proof for a theory or hypothesis. This has to come from experimental research. In my opinion, models should be regarded as a guide to show the possibilities of a system, similar to the hypothesis of this thesis that small diffuse lesions in the brain can be counteracted by self-repair. Ideally, there should be a continuous dialogue between models and experiments: models have to incorporate and integrate (the latest) experimental data from which new predictions can be derived, fuelling new experiments to verify the newly made predictions. This bootstrap method will possibly allow us to get insight into the reorganization of the brain after damage and the opportunity to develop worthwhile scientifically founded rehabilitation programs.

7.5 Future research

There are several avenues to extend the work presented in this thesis. There are two main directions concerning computer modeling or simulation research, namely to investigate the self-repair hypothesis of autonomous self-repair and to model recovery of brain damage as discussed in Section 7.4. Models of both directions can be extended by making more neurobiological detailed models or extend the existing models at the systems level. There are several ways to investigate the hypothesis that self-repair takes place in the brain experimentally. We provided one possible experiment in Section 7.3. We will discuss other ways of how self-repair can be (in-)validated. First, we will elaborate further how to do future simulation research.

At the systems level the simulation models can be extended by modeling brain structures involved in neuro-modulation as for example cholinergic (Hasselmo, 1995; Kilgard et al., 2001) or dopaminergic (Bao et al., 2001; Robbins & Everitt, 1996; Waelti et al., 2001) systems or other systems involved in learning, emotion and attention. Simulation models can, furthermore, be extended by changing the learning rule, by elaborating the neuronal model, or both. This has as advantage that the model becomes more ‘realistic’. To illustrate advantages of a ‘realistic’ network, in the Hopfield network of Chapter Three we needed an artificial stop criterion that stopped a self-repair cycle. Such an artificial criterion was not needed in the more neurobiological plausible model of Chapter Five. In future research, a more realistic feature is that we can get rid of the artificial normalization used in this thesis by introducing a

spike time dependent plasticity rule (STDP) learning rule. In these rules the magnitude of weight change is dependent on the activity of pre- and post-synaptic neurons of a certain time window and they posses intrinsic normalizing properties (Kempter et al., 2001). Future research can test whether this intrinsic normalizing property is sufficient for self-repair. It can, furthermore, be tested to what extent a neural network using this learning rule can withstand damage and to develop variant rules modeling plasticity after large damage. Another advantage of more realistic neural network features is that with more detailed neuronal models we would expect to have more realistic population dynamics. This would allow us to do predictions that can be validated empirically afterwards. An example of such a model that possesses realistic population dynamics is the model of sleep by Hill and Tononi (S. Hill & Tononi, 2004). This sleep model is able to produce characteristic sleep waves that can be tested with several brain measurement techniques like electroencephalography.

As was mentioned in the introduction of this chapter, the fastest way to validate whether self-repair takes place in the brain is to carry out experiments. In Section 7.3, we discussed one possible experiment that is able to validate the self-repair hypothesis. In this thesis, especially in Chapter Two, we have discussed many other experiments that tested one of the two components of the self-repair theory of redundancy and maintenance. These components can be found at different levels of the brain. For example, maintenance can be found at the synaptic level by plasticity mechanisms and at the behavorial level as use-it-or-lose-it. The self-repair hypothesis will be invalidated if redundancy and maintenance cannot be found at any level of the brain. In the future, however, different forms of redundancy and maintenance may be found for different levels of the brain. For instance, it is possible that other mechanisms than the two suggested plasticity mechanisms are able to carry out self-repair. Since research has started unraveling brain mechanisms and their effect on the neural circuit level, self-repair is not invalidated if it turns out that present known mechanisms are unable to carry out self-repair. For now, self-repair seems a fruitful and practical concept, since it provides a consistent framework for different findings and concepts for different levels of the brain.

To conclude, the most important contribution of this thesis is the idea that self-repair by maintenance of redundancy is possible in the brain. This idea can be applied to aging in the normal intact brain, as was the case for the models used in this thesis. It can also be applied to model recovery from brain damage. There are many possibilities for future research, like extending the current model of the normal intact brain or applying it to model functioning of the injured brain. Pursuing this research will hopefully contribute to preserve and prolong

mental health of mankind. At least this stimulating research will prolong the mental health of the researcher concerned.

References

Aartsen, M. J., Smits, C. H. M., Van Tilburg, T. G., Knipscheer, C. P. M., & Deeg, D. J. H. (2002). Activity in older adults: Cause or consequence of cognitive functioning? A longitudinal study on everyday activities and cognitive performance in older adults. *Journal of Gerontology, B57*, 153-162.

Albert, M. S., Butters, N., & Brandt, J. (1981). Patterns of remote memory in amnesic and demented patients. *Archives of Neurology, 38*, 495-500.

Altman, J. (1969). Autoradiographic and histological studies of postnatal neurogenesis. Iv. Cell proliferation and migration in the anterior forebrain, with special reference to persisting neurogenesis in the olfactory bulb. *J. Comp. Neurol., 137*(4), 433-457.

Altman, J., & Das, G. D. (1965). Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *J. Comp. Neurol., 124*(3), 319-335.

Alvarez, R., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of National Academy of Sciences (USA), 91*, 7041-7045.

Amari, S.-I. (1989). Characteristics of sparsely encoded associative memory. *Neural Networks, 2*, 451-457.

Amari, S.-I. (1990). Mathematical foundations of neurocomputing. *Proceedings of the IEEE, 78*(9), 1443-1463.

Anderson, J. A. (1983). Cognitive and psychological computation with neural models. *IEEE Transactions Systems Man and Cybernetics, SMC-13*, 799-815.

Arbib, M. A., Érdi, P., & Szentagothái, J. (1998). *Neural organization: Structure, function and dynamics*. Cambridge, Massachusetts: The MIT Press.

Artola, A., & Singer, W. (1993). Long-term depression of excitatory synaptic transmission and its relation to long-term potentiation. *Trends in Neurosciences, 16*, 480-487.

Ayala, F. J., & Kigger, J. A. (1982). *Modern genetics*. Menlo Park, CA: Benjamin/Cummings.

Bach-y-Rita, P. (1990). Brain plasticity as a basic for recovery of function in humans. *Neuropsychologia, 28*(6), 547-554.

Bailey, C. H., & Kandel, E. R. (1993). Structural changes accompanying memory storage. *Annual Review of Physiology, 55*, 397-426.

Bao, S., Chan, V. T., & Merzenich, M. M. (2001). Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature, 412*(79-83).

Barja, G. (2004). Free radicals and aging. *Trends in Neurosciences, 27*(10), 595-600.

Bartoletti, A., Cancedda, L., Reid, S. W., Tessarollo, L., Porciatti, V., Pizzorusso, T., et al. (2002). Heterozygous knock-out mice for brain-derived neurotrophic factor show a pathway-specific impairment of long-term potentiation but normal critical period for monocular deprivation. *Journal of Neuroscience, 22*(23), 10072-10077.

Beatty, W. M., Salmon, D. P., Butters, N., Heindel, W. C., & Granholm, E. L. (1988). Retrograde amnesia in patients with alzheimer's disease or huntington's disease. *Neuropsychology of Aging, 9*, 181-186.

Bengtsson, S. L., Nagy, Z., Skare, S., Forsman, L., Forssberg, H., & Ullén, F. (2005). Extensive piano practicing has regionally specific effects on white matter development. *Nature Neuroscience, 8*(9), 1148-1150.

Bertoni-Freddari, C., Fattoretti, P., Casoli, T., Meier-Rung, W., & Ulrich, J. (1990). Morphological adaptive response of the synaptic junctional zones in the human dentate gyrus during aging and alzheimer's disease. *Brain Research, 517*, 69-75.

Bertoni-Freddari, C., Meier-Runge, W., & Ulrich, J. (1988). Quantitative morphology of synaptic plasticity in the aging brain. *Neurobiology of aging, 9*, 181-186.

Bi, G.-Q. (2002). Spatiotemporal specificity of synaptic plasticity: Cellular rules and mechanisms. *Biological Cybernetics, 87*, 319-332.

Bienenstock, E. (1995). A model of neocortex. *Network: Computation in Neural Systems, 6*, 179-224.

Biernaski, J., & Corbett, D. (2001). Enriched rehabilitative training promotes improved forelimb motor function and enhanced dendritic growth after focal ischemic injury. *J. Neurosci., 21*, 5272-5280.

Bliss, T. V. P., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology, 232*, 331-356.

Bollobás, B. (1985). *Random graphs*. London: Academic Press.

Bolt, G. (1991). *Investigating fault tolerance in artificial neural networks* (Technical report No. YCS 154). Heslington, York: University of York.

Bontempi, B., Laurent-Demir, C., Destrade, C., & Jaffard, R. (1999). Time-dependent reorganization of brain circuitry underlying long-term memory storage. *Nature, 400*, 671-675.

Bosma, H., Boxtel, M. P. J. v., Ponds, R. W. H. M., Houx, P. J., Burdorf, A., & Jolles, J. (2003). Mental workload protects against cognitive impairment: Maas prospective cohort study. *Experimental Aging Research, 29*, 33-45.

Bosma, H., Boxtel, M. P. J. v., Ponds, R. W. H. M., Jelicic, M., Houx, P. J., Metsemakers, J., et al. (2002). Engaged lifestyle and cognitive function in middle and old-aged, non-demented persons: A reciprocal association. *Zeitschrift für Gerontologie und Geriatrie, 35*(6), 575-581.

Braitenberg, V., & Schüz, A. (1991). *Anatomy of the cortex: Statistics and geometry*. Berlin: Springer Verlag.

Brashers-Krug, T., Shadmehr, R., & Bizzi, E. (1996). Consolidation in human motor memory. *Nature, 382*, 252-255.

Buell, S., & Coleman, P. (1979). Dendritic growth in aged human brain and failure of growth in senile dementia. *Science, 206*, 854-856.

Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review Neuroscience, 21*, 149-186.

Butler, R. J. (1994). Neuropsychological investigation of amateur boxers. *British journal of sports medicine, 28*(3), 187-190.

Butler, S. R. (1988). Mechanisms for recovery from brain damage. *Archivio Italiano Riabilitazione di Scienze Neurology, 2*, 10-26.

Buzsaki, G. (1989). Two stage model of memory trace formation: A role for "noisy" brain states. *Neuroscience, 31*, 551-570.

Castro de, J. M., & Zrull, M. C. (1988). Recovery of sensorimotor function after frontal cortex damage in rats: Evidence that the serial lesion effect is due to serial recovery. *Behavioral Neuroscience, 102*(6), 843-851.

Cerella, J., & Hale, S. (1994). The rise and fall of information-processing rates over the life span. *Acta Psychologica, 86*, 109-197.

Changeux, J. P., & Danchin, A. (1976). Selective stabilization of developing synapses as mechanism for the specification of neuronal networks. *Nature, 264*, 705-712.

Chappell, M. H., Ulug, A. M., Zhang, L., Heitger, M. H., Jordan, B. D., Zimmerman, R. D., et al. (2006). Distribution of microstructural damage in the brains of professional boxers: A diffusion mri study. *Journal of magnetic resonance imaging, 24*(3), 537-542.

Christensen, H. (2001). What cognitive changes can be expected with normal ageing. *Australian and New Zealand Journal of Psychiatry 2001, 35*, 768-775.

Christos, G. A. (1996). Investigation of the crick-mitchison reverse-learning dream sleep hypothesis in a dynamical setting. *Neural Networks, 9*(3), 427-434.

Clausen, H., McCrory, P., & Anderson, V. (2005). The risk of chronic traumatic brain injury in professional boxing: Change in exposure variables over the past century. *British journal of sports medicine, 39*(9), 661-664.

Colcombe S, K. A. (2003). Fitness effects on the cognitive function of older adults: A meta-analytic study. *Psychological Science, 14*(2), 125-130.

Colcombe, S. J., Kramer, A. F., Erickson, K. I., Scalf, P., McAuley, E., Cohen, N. J., et al. (2004). Cardiovascular fitness, cortical plasticity, and aging. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 3316-3321.

Collins, M. A., Zou, J. Y., & Neafsey, E. J. (1998). Brain damage due to episodic alcohol exposure in vivo and in vitro: Furosemide neuroprotection implicates edema-based mechanism. *Federation of American Societies for Experimental Biology, 12*, 221-230.

Corwin, J. V., Nonneman, A. J., & Goodlett, C. (1981). Limited sparing of function on spatial delayed alternation after two-stage lesions of prefrontal cortex in the rat. *Physiology and Behavior., 26*(5), 763-771.

Cotman, C. W., & Berchtold, N. C. (2002). Excercise: A behavioral intervention to enhance brain health and plasticity. *Trends in Neurosciences, 25*(6), 295-301.

Cotman, C. W., & Nieto-Sampedro, M. (1982). Brain function, synapse renewal, and plasticity. *Annual Review Psychology, 33*, 371-401.

Cowan, J. D. (1995). Fault tolerance. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 390-395). Cambridge, Massachusetts: The MIT Press.

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

Curtis, S. D., & Nonneman, A. J. (1977). Effects of successive bilateral hippocampectomy on drl 20 performance in rats. *Physiology and Behavior, 19*(6), 707-712.

DeCarli, C., Murphy, D. G., & Tranh, M. (1995). The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults. *Neurology, 45*, 2077-2084.

deCharms, R. C., & Zador, A. (2000). Neural representation and the cortical code. *Annual Review Neuroscience, 23*, 613-647.

DeKosky, S. T., & Scheff, S. W. (1990). Synapse loss in frontal cortex biopsies in alzheimer's disease: Correlation with cognitive severity. *Annals of Neurology, 27*, 457-464.

Desai, N. S., Rutherford, L. C., & Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience, 2*(6), 515-520.

Dolev, S. (2000). *Self-stabilization*. Cambridge, Massachucetts: The MIT Press.

Downs, D. S., & Abwender, D. (2002). Neuropsychological impairment in soccer athletes. *The Journal of sports medicine and physical fitness, 42*(1), 103-107.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Changes in gray matter induced by training. *Nature, 427*, 311-312.

Dudek, S. M., & Bear, M. F. (1992). Homosynaptic long-term depression in area ca1 of hippocampus and effects of n-methyl-d-aspartate receptor blockade. *Proceedings of the National Academy of Sciences of the United States of America, 89*, 4363-4367.

Elbert, T., & Rockstroh, B. (2004). Reorganization of human cerebral cortex: The range of changes following use and injury. *Neuroscientist., 10*(2), 129-141.

Elwood, P. C., Gallacher, J. E. J., Hopkinson, C. A., Pickering, J., Rabbitt, P., Stollery, B., et al. (1999). Smoking, drinking, and other life-style factors and cognitive function in men in the caerphilly cohort. *Journal Epidemiology Communnity Health, 53*, 9-14.

Erdös, P., & Rényi, A. (1959). On random graphs i. *Publ. Math. Debrecen, 6*, 290-297.

Eriksson, P. S., Perfilieva, E., Bjork-Eriksson, T., Alborn, A. M., Nordborg, C., Peterson, D. A., et al. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine, 4*(11), 1313-1317.

Erlanger, D. M., Kutner, K. C., Barth, J. T., & Barnes, R. (1999). Neuropsychology of sports-related head injury: Dementia pugilistica to post

concussion syndrome. *Clin. Neuropsychology, 13*(2), 193-209.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature, 407*, 630-633.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*(6958), 614-616.

Finger, S., & Stein, D. G. (1982). *Brain damage and recovery: Research and clinical perspectives*. New York: Academic Press.

Forster, M., Dubey, A., Dawson, K., Stutts, W., Lal, H., & Sohal, R. (1996). Age-related losses of cognitive function and motor skills in mice are associated with oxidative protein damage in the brain. *Proceedings of the National Academy of Sciences of the United States of America, 93*(10), 4765-4769.

Frankland, P. W., O'Brien, C., Ohno, M., Kirkwood, A., & Silva, A. J. (2001). Alpha-camkii-dependent plasticity in the cortes is required for permanent memory. *Nature, 411*, 309-313.

Friston, K. (2002). Beyond phrenology: What can neuroimaging tell us about distributed circuitry. *Annual Review Neuroscience, 25*, 221-250.

Fukui, K., Onodera, K., Shinkai, T., Suzuki, S., & Urano, S. (2001). Impairment of learning and memory in rats caused by oxidative stress and aging, and changes in antioxidative defense systems. *Annals New York Acadamy of Science, 928*, 168-175.

Gais, S., Plihal, W., Wagner, U., & Born, J. (2000). Early sleep triggers memory for early visual discrimination skills. *Nature Neuroscience, 3*, 1335-1339.

Gavin, M. R., & Isaac, W. (1986). Recovery of function of a conditioned avoidance response in rats with serial and single-stage bilateral occipital ablation. *Physiological Psychology, 14*(1-2), 31-35.

Geinisman, Y. (2000). Structural synaptic modifications associated with hippocampal ltp and behavioral learning. *Cerbral Cortex, 10*, 952-962.

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge: Cambridge University Press.

Glassman, R. B. (1987). An hypothesis about redundancy and reliability in the brains of higher species: Analogies with genes, internal organs, and engineering systems. *Neuroscience & Biobehavioral Reviews, 11*, 275-285.

Gold, D. P., Andres, D., Etezadi, J., Arbuckle, T. Y., Schwartzman, A., & Chaikelson, J. (1995). Structural equation model of intellectual change and continuity and predictors of intelligence in older men. *Psychology and Aging, 10*, 294-303.

Goldman, M. S. (1990). Experience-dependent neuropsychological recovery and the treatment of chronic alcoholism. *Neuropsychology Review, 1*, 75-101.

Gould, E., Beylin, A., Tanapat, P., Reeves, A., & Shors, T. J. (1999). Learning enhances adult neurogenesis in the hippocampal formation. *Nature Neuroscience, 2*(3), 260-265.

Grady, C. L., & Craik, F. I. M. (2000). Changes in memory processing with age. *Current Opinion in Neurobiology, 10*, 224-231.

Greenough, W. T., Black, J. E., & Wallace, C. S. (2002). Experience and brain development. In M. H. Johnson, Y. Munakata & R. O. Gilmore (Eds.), *Brain development and cognition*. Oxford: Blackwell Publishing.

Hairston, I. S., & Knight, R. T. (2004). Sleep on it. *Nature, 430*, 27-28.

Hamilton, R. H., & Pascual-Leone, A. (1998). Cortical plasticity associated with braille learning. *Trends in Cognitive Sciences, 2*, 168-174.

Harman, D. (1956). Aging: A theory based on free radical and radiation chemistry. *J. Gerontology, 11*(3), 298-300.

Hasselmo, M. E. (1994). Runaway synaptic modification in models of the cortex: Implications for alzheimer's disease. *Neural Networks, 7*(1), 13-40.

Hasselmo, M. E. (1995). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioral Brain Research, 67*, 1-27.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425-2430.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hemmen, J. L. v. (1997). Hebbian learning, its correlation catastrophe, and unlearning. *Network: Computation Neural Systems, 8*(3), V1-V17.

Hess, G., Aizenman, C. D., & Donoghue, J. P. (1996). Conditions for the induction of long-term potentiation in layer ii/iii horizontal connections of the rat motor cortex. *Journal of Neurophysiology, 75*(5), 1765-1778.

Hess, G., & Donoghue, J. P. (1996). Long-term depression of horizontal connections in rat motor cortex. *European Journal of Neuroscience, 8*(4), 658-665.

Hill, R. D. (1993). The impact of long-term exercise training on psychological function in older adults. *Journal of Gerontology, 48*, P12-P17.

Hill, S., & Tononi, G. (2004). Modeling sleep and wakefulness in the thalamocortical system. *Journal of Neurophysiology, 93*, 1671-1698.

Hobson, J. A. (1989). *Sleep*. New York: Scientific American Library.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of ion currents and its application to conduction and excitation in nerve membranes. *Journal Physiology (London), 117*, 500-544.

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

Hopfield, J. J., Feinstein, D. I., & Palmer, R. G. (1983). 'unlearning' has a stabilizing effect in collective memories. *Nature, 304*, 158-159.

Horn, D., Levy, N., & Ruppin, E. (1998a). Memory maintenance via neuronal regulation. *Neural Computation, 10*, 1-18.

Horn, D., Levy, N., & Ruppin, E. (1998b). Neuronal regulation versus synaptic unlearning in memory maintenance mechanisms. *Network: Computation in Neural Systems, 9*(4), 577-586.

Huber, R., Ghilardi, M. F., Massimini, M., & Tononi, G. (2004). Local sleep and learning. *Nature, 430*, 78-81.

Hultsch, D. F., Hertzog, C., Small, B. J., & Dixon, R. A. (1999). Use it or lose it: Engaged lifestyle as a buffer of cognitive decline in aging? *Psychology and Aging, 14*, 245-263.

Ikeda, K., Tanihara, H., Tatsuno, T., Noguchi, H., & Nakayama, C. (2003). Brain-derived neurotrophic factor shows a protective effect and improves recovery of the erg b-wave response in light-damage. *Journal of Neurochemistry, 87*(2), 290-296.

Isaac, C. L., & Mayes, A. R. (1999). Rate of forgetting in amnesia: I. Recall and recognition of prose. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*(4), 942-962.

Izquierdo, I., Quillfeldt, J. A., Zanatti, M. S., Quevedo, J., Schaeffer, E., Schmitz, P. K., et al. (1997). Sequential role of hippocampus and amygdala, entorhinal cortex and parietal cortex in formation and retrieval of memory for inhibitory avoidance in rats. *European Journal of Neuroscience, 9*, 786-793.

Jacobs, K. M., & Donoghue, J. P. (1991). Reshaping the cortical motor map by unmasking latent intracortical connections. *Science, 251*, 944-947.

Jacobson, M. (1991). *Developmental neurobiology*. New York: Plenum Press.

Jones, E. G. (2000). Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annual Review Neuroscience, 23*, 1-37.

Kali, S., & Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience, 7*(3), 286-294.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.). (1991). *Principles of neural science* (3rd ed.). Connecticut: Appleton & Lange.

Karni, A., Meyer, G., Jezzard, P., Adams, M. M., Turner, R., & Ungerleider, G. (1995). Functional mri evidence for adult motor cortex plasticity during motor skill learning. *Nature, 377*, 155-158.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., & Sagi, D. (1994). Dependence on rem sleep of overnight improvement of a perceptual skill. *Science, 265*, 679-682.

Kavanau, J. L. (1996). Memory, sleep, and dynamic stabilization of neural circuitry: Evolutionary perspectives. *Neuroscience and Biobehavioral Reviews, 20*(2), 289-311.

Kavanau, J. L. (1997). Memory, sleep and the evolution of mechanisms of synaptic efficacy maintenance. *Neuroscience, 79*(1), 7-44.

Kempermann, G., & Gage, F. H. (1999). New nerve cells for the adult brain. *Scientific American, 280*(5), 48-53.

Kempermann, G., Kuhn, H. G., & Gage, F. H. (1998). Experience-induced neurogenesis in the senescent dentate gyrus. *The Journal of Neuroscience, 18*(9), 3206-3212.

Kempter, R., Gerstner, W., & Hemmen, J. L. v. (2001). Intrinsic stabilization of output rates by spike-based hebbian learning. *Neural Computation, 13*, 2709-2741.

Kilgard, M. P., Pandya, P. K., Vazquez, J., Gehi, A., Schreiner, C. E., & Merzenich, M. M. (2001). Sensory input directs spatial and temporal plasticity in primary auditory cortex. *Journal Neurophysiology, 86*, 326-338.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia for fear. *Science, 256*, 675-677.

Kleim, J. A., Jones, T. A., & Schallert, T. (2003). Motor enrichment and the induction of plasticity before or after brain injury. *Neurochemical Research, 28*(11), 1757-1769.

Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. Oxford: Oxford University Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59-69.

Kolb, B. (1995). *Brain plasticity and behaviour*. Hillsday, New Jersey: Lawrence Erlbaum.

Kolb, B. (1999). Synaptic plasticity and the organization of behaviour after early and late brain injury. *Canadian Journal of Experimental Psychology, 53*(1), 62-76.

Kolb, B., & Gibb, R. (1993). Possible anatomical basis of recovery of function after neonatal frontal lesions in rats. *Behavioral Neuroscience, 107*(5), 799-811.

Kolb, B., Holmes, C., & Whishaw, I. Q. (1987). Recovery from early cortical lesions in rats. Iii. Neonatal removal of posterior parietal cortex has greater behavioral and anatomical effects than similar removals in adulthood. *Behavioral Brain Research, 26*(2-3), 119-137.

Kolb, B., Stewart, J., & Sutherland, R. J. (1997). Recovery of function is associated with increased spine density in cortical pyramidal cells after frontal lesions and/or noradrenaline depletion in neonatal rats. *Behavioral Brain Research, 89*(1-2), 61-70.

Kopelman, M. D. (1989). Remote and autobiographical memory, temporal context memory, and frontal atrophy in korsakoff and alzheimer patients. *Neuropsychologia, 27*, 437-460.

Kovalchuk, Y., Hanse, E., Kafitz, K. W., & Konnerth, A. (2002). Post-synaptic induction of bdnf-mediated long-term potentiation. *Nature, 295*, 1729-1734.

Kritchevsky, M., & Squire, L. R. (1989). Transient global amnesia: Evidence for extensive, temporally graded retrograde amnesia. *Neurology, 39*, 213-219.

Kuhn, H. G., Dickinson-Anson, H., & Gage, F. H. (1996). Neurogenesis in the dentate gyrus of the adult rat: Age-related decrease of neuronal progenitor proliferation. *The Journal of Neuroscience, 16*(6), 2027-2033.

Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters, 87*(19), 1-4.

Liao, D., Zhang, X., Brian, R. O., Ehlers, M., & Huganir, R. L. (1999). Regulation of morphological postsypnaptic silent synapses in developing hippocampal neurons. *Nature Neuroscience, 2*, 37-43.

Linnarsson, S., Bjorklund, A., & Ernfors, P. (1997). Learning deficit in bdnf mutant mice. *European Journal of Neuroscience, 9*(12), 2581-2587.

Liu, J., Solway, K., Messing, R. O., & Sharp, F. R. (1998). Increased neurogenesis in the dentate gyrus after transient global ischemia in gerbils. *The Journal of Neuroscience, 18*(19), 7768-7778.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., et al. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature, 429*, 883-891.

Lumer, E. D., Edelman, G. M., & Tononi, G. (1997). Neural dynamics in a model of the thalamocortical system. I. Layers, loops and the emergence of fast synchronous rhythms. *Cerbral Cortex, 7*, 207-227.

Luskin, M. B. (1993). Restricted proliferation and migration of postnatally generated neurons derived from the forebrain subventricular zone. *Neuron, 11*(1), 173-189.

MacGregor, R. J., & Oliver, R. M. (1974). A model for repetetive firing in neurons. *Cybernetik, 16*, 53-64.

Magavi, S. S., Leavitt, B. R., & Macklis, J. D. (2000). Induction of neurogenesis in the neocortex of adult mice. *Nature, 405*, 951-955.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., et al. (2000). Navagation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences USA, 97*(8), 4398-4403.

Malsburg, C. v. d. (1973). Self-organization of orientation sensitive cells in the striata cortex. *Kybernetik, 14*, 85-100.

Malsburg, C. v. d. (1995). Self-organization and the brain. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*. Cambridge, Massachusetts: The MIT Press.

Maquet, P. (1995). Sleep function(s) and cerebral metabolism. *Behavioral Brain Research, 69*, 75-83.

Maquet, P. (2001). The role of sleep in learning and memory. *Science, 294*, 1048-1052.

Marshall, J. F. (1984). Brain functions: Neural adaptations and recovery from injury. *Annual Review Psychology, 35*, 277-308.

Martin, S. J., Grimwood, P. D., & Morris, R. G. M. (2000). Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review Neuroscience, 23*, 649-711.

Matser, J. T., Kessels, A. G., Jordan, B. D., Lezak, M. D., & Troost, J. (1998). Chronic traumatic brain injury in professional soccer players. *Neurology, 51*(3), 791-796.

Matser, J. T., Kessels, A. G. H., Lezak, M. D., & Troost, J. (2001). A dose-response relation of headers and concussions with cognitive impairment in professional soccer players. *Journal of Clinical and Experimental Neuropsychology, 23*(6), 770-774.

McAllister, A. K., Katz, L. C., & Lo, D. C. (1999). Neurotrophins and synaptic plasticity. *Annual Review Neuroscience, 22*, 295-318.

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

McEachern, J. C., & Shaw, C. A. (2001). Revisiting the ltp orthodoxy. In C. Hölscher (Ed.), *Neuronal mechanisms of memory formation*. Cambridge: Cambridge University Press.

McGraw, P. N., & Menzinger, M. (2003). Topology and computational performance of attractor neural networks. *Physical review. E, Statistical, nonlinear, and soft matter physics, 68*, 047102?

Mednick, S. C., Nakayama, K., Cantero, J. L., Atienza, M., Levin, A. A., Pathak, N., et al. (2002). The restorative effect of naps on perceptual deterioration. *Nature Neuroscience, 5*(7), 677-681.

Meeter, M. (2003). Control of consolidation in neural networks: Avoiding runaway effects. *Connection Science, 15*, 45-61.

Meeter, M., & Murre, J. M. J. (2004). Consolidation of long-term memory: Evidence and alternatives. *Psychological Bulletin, 130*, 843-857.

Meeter, M., & Murre, J. M. J. (2005). Tracelink: A model of amnesia and consolidation. *Cognitive Neuropsychology, 22*, 559-587.

Mendez, M. F. (1995). The neuropsychiatric aspects of boxing. *Int. J. Psychiatry Med., 25*(3), 249-262.

Muellbacher, W., Ziemann, U., Wissel, J., Dang, N., Kofler, M., Facchini, S., et al. (2002). Early consolidation in human primary motor cortex. *Nature, 415*(6872), 640-644.

Münte, T. F., Altenmüller, E., & Jäncke, L. (2002). The musicians's brain as a model of neuroplasticity. *Nat. Rev. Neurosci., 3*, 473-478.

Murre, J. M. J. (1994). A model for categorization and recognition in amnesic patients. In M. Gazzaniga (Ed.), *Proceedings of the cognitive neuroscience meeting 1994*.

Murre, J. M. J. (1996). Tracelink: A model of amnesia and consolidation of memory. *Hippocampus, 6*, 675-684.

Murre, J. M. J. (1997). Implicit and explicit memory in amnesia: Some explanations and predictions by the tracelink model. *Memory, 5*, 213-232.

Murre, J. M. J., Graham, K. S., & Hodges, J. R. (2001). Semantic dementia: Relevance to connectionist models of long-term memory. *Brain, 124*.

Murre, J. M. J., Griffioen, A. R., Dulk, P. d., & Robertson, I. (submitted). Selfrepairing neural networks.

Nadel, L., & Moskovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7*, 217-227.

Nadel, L., Samsonovitch, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging and neuropsychological results. *Hippocampus, 10*, 352-368.

Nakamura, H., Kobayashi, S., Ohashi, Y., & Ando, S. (1999). Age-changes of brain synapses and synaptic plasticity in response to an enriched environment. *Journal of Neuroscience Research, 56*, 307-315.

Nicolelis, M. A. L., Ghazanfar, A. A., Stambaugh, C. R., Oliveira, L. M. O., Laubach, M., Chapin, J. K., et al. (1998). Simultaneous encoding of tactile information by three primate cortical areas. *Nature Neuroscience, 1*(7), 621-630.

Nudo, R. J., Wise, B. M., Sifuentes, F., & Milliken, G. W. (1996). Neural substrates for the effects of rehabilitation training on motor recovery following ischemic infarct. *Science, 272*, 1791-1794.

Palm, G. (1982). *Neural assemblies: An alternative approach to artificial intelligence*. Berlin: Springer-Verlag.

Palm, G., & Sommer, F. T. (1996). Associative data storage and retrieval in neural nets. In E.Domany, J. L. v. Hemmen & K.Schulten (Eds.), *Models of neural networks iii (1996)* (pp. 79-118). New York: Springer.

Parent, J. M., Yu, T. W., Leibowitz, R. T., Geschwind, D. H., Sloviter, R. S., & Lowenstein, D. H. (1997). Dentate granule cell neurogenesis is increased by seizures and contributes to aberrant network reorganization in the adult rat hippocampus. *The Journal of Neuroscience, 17*(10), 3727-3738.

Pavlides, C., Greenstein, Y. J., Grudman, M., & Winson, J. (1988). Long-term potentiation in the dentate gyrus is induced preferentially on the positive phase of the theta-rhythm. *Brain Research, 439*, 383-387.

Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience, 9*, 2907-2918.

Petsche, T., & Dickinson, B. W. (1990). Trellis codes, receptive fields, and fault tolerant, self-repairing neural networks. *IEEE Transactions on Neural Networks, 1*(2), 154-166.

Phillips, L. H., & Sala, S. D. (1996). Aging, intelligence and anatomical segregation in the frontal lobes. *Learning and Individual differences, 10*(3), 217-243.

Porter, M., & O'Brien, M. (1996). Incidence and severity of injuries resulting from amateur boxing in ireland. *Clinical journal of sport medicine: official journal of the Canadian Academy of Sport Medicine, 6*(2), 97-101.

Quartarone, A., Bagnato, S., Rizzo, V., Siebner, H. R., Dattola, V., Scalfari, A., et al. (2003). Abnormal associative plasticity of the human motor cortex in writer's cramp. *Brain, 126*, 2586-2596.

Ramirez, J. J., Bulsara, K., Moore, S. C., Ruch, K., & Abrams, W. (1999). Progressive unilateral damage of the entorhinal cortex enhances synaptic efficacy of the crossed entorhinal afferent to dentate granule cells. *Journal of Neuroscience, 19*, 1-6.

Reid, C. A., Dixon, D. B., Takahashi, M., Bliss, T. V., & Fine, A. (2004). Optical quantal analysis indicates that long-term potentiation at single hippocampal mossy fiber synapses is expressed through increased release probability, recruitment of new release sites, and activation of silent synapses. *Journal of Neuroscience, 24*(14), 3618-3626.

Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B., & Davatzikos, C. (2003). Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain. *The Journal of Neuroscience, 23*(8), 3295-3301.

Ribot, T. (1881). *Les maladies de la memoire*. Paris: Germer Baillare.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1999). *Spikes: Exploring the neural code*. Cambridge, Massachusetts: The MIT Press.

Rioult-Pedotti, M.-S., Friedman, D., & Donoghue, J. P. (2000). Learning-induced ltp in neocortex. *Science, 290*, 533-536.

Robbins, T. E., & Everitt, B. J. (1996). Neurobehavioral mechanisms of reward and motivation. *Current Opinion in Neurobiology, 6*, 228-236.

Robertson, E. M., Pascual-Leone, A., & Miall, R. C. (2004). Current concepts in procedural consolidation. *Nature Reviews Neuroscience, 5*, 1-7.

Robertson, I. H., & Murre, J. M. J. (1999). Rehabilitation of brain damage: Brain plasticity and principles of guided recovery. *Psychological Bulletin, 125*, 544-575.

Robins. (1996). Consolidation in neural networks and in the sleeping brain. *Connection Science, 8*, 259-275.

Robins, A., & McCallum, S. (1998). Catastrophic forgetting and the pseudorehearsal solution in hopfield type networks. *Connection Science, 7*, 121-135.

Roffwarg, H. P., Musio, J. N., & Dement, W. C. (1966). Ontogenetic development of the human sleep-dream cycle. *Science, 152*, 604-609.

Rumelhart, D. E., & McClelland, J. L. (1981). An interactive activation model of the effect of context in perception: Part 1. *Psychological Review, 88*, 375-405.

Rutherford, A., Stephens, R., & Potter, D. (2003). The neuropsychology of heading and head trauma in association football (soccer): A review. *Neuropsychology review, 13*(3), 159-179.

Ryan, A. J. (1998). Intracranial injuries resulting from boxing. *Clinics in sports medicine, 17*(1), 155-168.

Saito, S., Kobayashi, S., Ohashi, Y., Igarashi, M., Komiya, Y., & Ando, S. (1994). Decreased synaptic density in aged brains and its prevention by rearing under enriched environment as revealed by synaptophysin contents. *Journal of Neuroscience Research, 39*, 57-62.

Salat, D. H., Kaye, J. A., & Janowsky, J. S. (1999). Prefrontal gray and white matter volumes in healthy aging and alzheimer disease. *Arch. Neurol., 56*, 338-344.

Salthouse, T. A., Berish, D. E., & Miles, J. D. (2002). The role of cognitive stimulation on the relations between age and cognitive functioning. *Psychology and Aging, 17*(4), 548-557.

Salvador, R., Suckling, J., Coleman, M. R., Pickard, J. D., Menon, D., & Bullmore, E. (2005). Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral Cortex, 15*, 1332-1342.

Sanes, J. N., & Doneghue, J. P. (2000). Plasticity and primary motor cortex. *Annual Review Neuroscience, 23*, 393-415.

Scarmeas, N., & Stern, Y. (2003). Cognitive reserve and lifestyle. *Journal of Clinical and Experimental Neuropsychology, 25*(5), 625-633.

Scheff, S. W., Wright, D. C., Morgan, W. K., & Bowers, R. P. (1977). The differential effects of additional cortical lesions in rats with single- or multiple-stage lesions of the visual cortex. *Physiological Psychology, 5*(1), 97-102.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience, 5*(7), 688-694.

Sejnowski, T. J. (1977). Storing covariance with non-linearly interacting neurons. *Journal Mathematical Biology, 4*, 303-321.

Sejnowski, T. J., & Destexhe, A. (2000). Why do we sleep? *Brain Research, 1*, 1-16.

Shadmehr, R., & Holcomb, H. H. (1997). Neural correlates of motor memory consolidation. *Science, 277*, 821-825.

Simard, D., Nadeau, L., & Kröger, H. (2005). Fastest learning in small-world neural networks. *Physics Letters A, 336*(1), 8-15.

Singer, W. (1990). Ontogenetic self-organization and learning. In J. L. McGaugh, N. M. Weinberger & G. Lynch (Eds.), *Brain organization and memory: Cells, systems, and circuits* (pp. 211-233). Oxford: Oxford University Press.

Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics, 2*(2), 145-162.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99*, 195-231.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology, 5*, 169-175.

Squire, L. R., Cohen, N. J., & Nadel, L. (1984). The medial temporal region and memory consolidation: A new hypothesis. In H. Weingarter & E. Parker (Eds.), *Memory consolidation* (pp. 185-210). Hillsdale, NJ: Lawrence Erlbaum.

Squire, L. R., Haist, F., & Shimamura, A. P. (1989). The neurology of memory: Quantitative assessment of retrograde amnesia in two groups of amnesic patients. *Journal of Neuroscience, 9*, 828-839.

Star, E. N., Kwiatkowski, D. J., & Murthy, V. N. (2002). Rapid turnover of actin in dendritic spines and its regulation by activity. *Nature Neuroscience, 5*(3), 239-246.

Stephens, R., Rutherford, A., Potter, D., & Fernie, G. (2005). Neuropsychological impairment as a consequence of football (soccer) play and football heading: A preliminary analysis and report on school students (13-16

years). *Child neuropsychology: a journal on normal and abnormal development in childhood and adolescence, 11*(6), 513-526.

Steriade, M. (2001). Impact of network activities on neuronal properties in corticothalamic systems. *Journal of Neurophysiology, 86*, 1-39.

Steriade, M., McCormick, D. A., & Sejnowski, T. J. (1993). Thalamocortical oscillations in the sleeping and aroused brain. *Science, 262*, 679-685.

Sterr, A. (1998). Changed perception in braille readers. *Nature*, 134-135.

Stickgold, R., Hobson, J. A., Fosse, R., & Fosse, M. (2001). Sleep, learning, and dreams: Off-line memory reprocessing. *Science, 294*, 1052-1057.

Stickgold, R., James, L. T., & Hobson, J. A. (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience, 3*(12), 1237-1238.

Stoyan, D., Kendall, W. S., & Mecke, J. (1997). *Stochastic geometry and its applications*. Berlin: John Wiley & Sons.

Taub, E., Uswatte, G., & Elbert, T. (2002). New treatments in neurorehabilitation founded on basic research. *Nat. Rev. Neurosci., 3*, 228-236.

Tchernev, E. B., Mulvaney, R. G., & Phatak, D. S. (2005). Investigating the fault tolerance of neural networks. *Neural Computation, 17*, 1646-1664.

Toni, N., Buchs, P. A., Nikonenko, I., Bron, C. R., & Muller, D. (1999). Ltp promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature, 402*, 421-425.

Tononi, G., & Cirelli, C. (2003). Sleep and synaptic homeostasis: A hypothesis. *Brain Research Bulletin, 62*.

Trachtenberg, J. T., Chen, B. E., Knott, G. W., Feng, G., Sanes, J. R., Welker, E., et al. (2002). Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. *Nature, 420*, 788-794.

Tropea, D., Caleo, M., & Maffei, L. (2003). Synergistic effects of brain-derived neurotrophic factor and chondroitinase abc on retinal fiber sprouting after denervation of the superior colliculus in adult rats. *The Journal of Neuroscience, 23*(18), 7034-7044.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience, 4*(8), 832-838.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., & Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature, 391*, 892-896.

Turrigiano, G. G., & Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current Opinion in Neurobiology, 10*, 358-364.

Turrigiano, G. G., & Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci., 5*(2), 97-107.

Urano, S., Sato, Y., Otonari, T., Makabe, S., Suzuki, S., Ogata, M., et al. (1998). Aging and oxidative stress in neurodegeneration. *Biofactors, 7*(1-2), 103-112.

Uylings, H. B. M., West, M. J., Coleman, P. D., Brabander, J. M. d., & Flood, D. G. (1999). Neuronal and cellular changes in aging brain. In J. Trojanowski & C. Clark (Eds.), *Neurodegenerative dementias. Clinical features and pathological mechanisms*. N.Y.: McGraw-Hill.

Villablanca, J. R., Carlson-Kuhta, P., Schmanke, T. D., & Hovda, D. A. (1998). A critical maturational period of reduced brain vulnerability to developmental

injury. I. Behavioral studies in cats. *Brain Research Developmental Brain Research, 105*(2), 309-324.

Voronin, L. L., & Cherubini, E. (2004). "deaf, or mute and whispering" silent synapses: Their role in synaptic plasticity. *Journal of Physiology, 557*(Pt 1), 3-12.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*, 43-48.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small world' networks. *Nature, 393*, 440-442.

Webbe, F. M., & Ochs, S. R. (2003). Recency and frequency of soccer heading interact to decrease neurocognitive performance. *Applied neuropsychology, 10*(1), 31-41.

Weiss, S. (1999). Pathways for neural stem cell biology and repair. *Nature Biotechnology, 17*, 850-851.

White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., et al. (1999). Genome sequence of the radioresistant bacterium deinococcus radiodurans r1. *Science, 286*, 1571-1577.

Whitty, C. W. M., & Zangwill, O. L. (1977). Traumatic amnesia. In C. W. M. Whitty & O. L. Zangwill (Eds.), *Amnesia* (pp. 118-135). London: Butterworths.

Wierenga, C. J., & Wadman, W. J. (2003). Excitatory inputs to ca1 interneurons show selective synaptic dynamics. *Journal of neurophysiology, 90*(2), 811-821.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 255*, 676-679.

Wilson, R. S., Barnes, L. L., & Bennett, D. A. (2003). Assessment of lifetime participation in cognitively stimulating activities. *Journal of Clinical and Experimental Neuropsychology, 25*(5), 634-642.

Withers, G. S., & Greenough, W. T. (1989). Reach training selectivity alters dendritic branching in subpopulations of layer ii-iii pyramidals in rat motor-somatosensory forelimb cortex. *Neuropsychologia, 27*, 61-69.

Xerri, C., Merzenich, M. M., Petersen, B. E., & Jenkins, W. (1998). Plasticity of primary somatosensory cortex paralleling sensorimotor skill recovery from stroke in adult monkeys. *Journal of Neurophysiology, 79*, 2119-2148.

Xing, J., & Gerstein, G. L. (1996a). Networks with lateral connectivity. I. Dynamic properties mediated by the balance of intrinsic excitation and inhibition. *Journal of Neurophysiology, 75*(1), 184-199.

Xing, J., & Gerstein, G. L. (1996b). Networks with lateral connectivity. Ii. Development of neuronal grouping and corresponding receptive field changes. *Journal of Neurophysiology, 75*(1), 200-216.

Yikoski, A., Erkinjuntti, T., Raininko, R., Sarna, S., Sulkava, R., & Tilvis, R. (1995). White matter hyperintensities on mri in the neurologically nondiseased elderly: Analysis of cohorts of consecutive subjects aged 55 to 85 years of living at home. *Stroke, 26*, 1171-1177.

Zhang, L., Heier, L. A., Zimmerman, R. D., Jordan, B., & Ulug, A. M. (2006). Diffusion anisotropy changes in the brains of professional boxers. *AJNR American journal of neuroradiology, 27*(9), 2000-2004.

Zhang, W., & Linden, D. J. (2003). The other side of the engram: Experience-driven changes in neuronal intrinsic excitability. *Nat. Rev. Neurosci., 4*, 885-900.

Dutch Summary

Introductie

Hersenletsel en in het bijzonder dementie is een van de meest voorkomende kwalen in de westerse samenleving. De gemiddelde levensverwachting wordt steeds hoger, maar hiermee nemen ook het aantal (verschillende typen) hersenbeschadigingen toe. Een overzichtsartikel van Robertson en Murre (1999) onderscheidt verschillende typen van hersenbeschadigingen op basis van herstel, namelijk hersenbeschadigingen die nooit herstellen, hersenbeschadigingen die gedeeltelijk herstellen en hersenbeschadigingen die uit zichzelf herstellen. De twee laatstgenoemde beschadigingen laten zien dat de hersenen een vorm van zelfreparatie hebben. Dit proefschrift probeert door middel van theoretische modellen meer inzicht te krijgen in de zelfreparerende eigenschap van het brein. Een tweede doel is om een basis te leggen voor modellen voor hersenherstel na beschadiging en revalidatie.

Zelfreparatie: onderhoud van redundantie

De hypothese van dit proefschrift is dat de zelfreparerende eigenschap van het brein wordt gegeven door onderhoud van redundantie. Redundantie is een overschot of reserve aan informatie, zodat (kleine) beschadigingen niet direct tot fouten in gedrag leiden. Onderhoud van redundantie is de voortdurende correctie of reparatie van de kleine beschadigingen. Een goed voorbeeld van het voordeel van regelmatig onderhoud in vergelijking met af en toe repareren is de correctie van een tekst, waarbij fouten ontstaan door het wegvallen van de tekst. Stel je hebt de zin “we rijden met de auto op weg”. In het geval van onderhoud corrigeer je voor een kleine fout, zoals “we r$den met de auto o$ de we$. De fouten ‘$’ kunnen nu nog gemakkelijk worden hersteld aan de hand van overgebleven informatie in de rest van de zin. Als er geen onderhoud wordt gepleegd, dan kan bijvoorbeeld de zin “ we $$$$n $$t de $$$$ o$ de e” door fouten ontstaan. Het is duidelijk dat het veel moeilijker is om deze zin te corrigeren.

Redundantie komt op verschillende manieren voor in een systeem. Een eenvoudige vorm van redundantie is informatie in de vorm van kopieën. Essentieel voor redundantie is dat er informatie aanwezig moet zijn om het oorspronkelijke systeem te kunnen reconstrueren. In hoofdstuk 2 van het proefschrift wordt uitgegaan van de hypothese dat er verschillende vormen van redundantie in het brein te vinden zijn. Zowel op synaptisch niveau als op neuron niveau is er redundantie in de vorm van kopieën: na beschadiging zijn er reserve synapsen die

de beschadigde synapsen vervangen en worden er (een groter aantal) nieuwe neuronen geboren.

Redundantie is ook in een andere vorm aanwezig in het brein, namelijk als netwerk-redundantie. Bij beschadiging van een deel van het netwerk kan dit met behulp van informatie uit andere delen van het netwerk hersteld worden. Om een voorbeeld te geven: Stel dat je een netwerk hebt met knopen en verbindingen waarin er een verbinding is tussen knoop A en knoop B. Verbindingen kunnen alleen worden versterkt als twee knopen min of meer tegelijkertijd even actief zijn. Als de verbinding tussen A en B dermate zwak is, dat knoop A knoop B niet meer kan activeren, dan kan de verbinding niet meer versterkt worden door alleen het activeren van A. Netwerk redundantie houdt in dat knoop A verbonden is met een knoop C, die op zijn of haar beurt weer verbonden is met knoop B. Via knoop C kan knoop A toch knoop B activeren en hierdoor kan de verbinding tussen A en B weer sterker worden. Dit proefschrift is voornamelijk op netwerkredundantie gericht.

In hoofdstuk 2 worden een aantal mechanismen besproken die onderhoud of reparatie kunnen uitvoeren aan netwerk redundantie. Deze mechanismen, die op synaptisch en neuraal niveau plaatsvinden, worden vooral geassocieerd met veranderingen die plaatsvinden in een gezond individu, zoals veranderingen veroorzaakt door leren. Een hypothese van dit proefschrift is dat er voordurend reparatie plaatsvindt: processen die repareren, zoals leren, vinden frequent plaats. Een andere hypothese is dat zelfreparatie niet alleen plaatsvindt bij kinderen maar ook bij volwassenen, omdat de neurale en synaptische processen ook in het brein van een volwassene worden gevonden. Niet alleen op neuraal en op synaptisch niveau zijn er data die zelfreparatie in volwassenen ondersteunen, tevens zijn er data op gedragsniveau die deze hypothese ondersteunen. Al deze data zijn gecentreerd rond twee begrippen, namelijk “use-it-or-lose-it” en het “serial-lesion” effect.

Het principe van “use-it-or-lose-it” stelt dat het gebruik van de hersenen de duur van het goed functioneren verlengt. Dit blijkt onder andere uit data van mensen die verzameld zijn over een tijdspanne van enkele jaren (zogenaamde longitudinale studies). Hierbij correleren bepaalde intellectuele (zoals schaken of het bijwonen van cursussen), sociale of fysieke bezigheden met langzamer cognitief verval. Het probleem met statistische correlaties van variabelen is dat het causale verband tussen de variabelen niet duidelijk en moeilijk bewijsbaar is. In dit geval hoeft het niet zo te zijn dat activiteiten ervoor zorgen dat het brein gezonder blijft, maar kan het ook zo zijn dat mensen met een gezonder brein juist beter in staat zijn allerlei activiteiten te ontplooien. Buiten longitudinale studies die een causaal verband aantoonden in beide richtingen, zijn er ook allerlei mens- en dierstudies die een effect

van activiteit op hersenanatomie laten zien en in het bijzonder op de hersenconnectiviteit. Dit komt overeen met de zelfreparatie gedachte dat activiteit of stimulatie effect hebben op de netwerkverbindingen in de hersenen. En dit heeft op zijn beurt weer invloed op het functioneren van het brein.

Het "serial-lesion" effect is de bevinding dat een serie van kleine laesies met tussen elke laesie een tijdsperiode, minder schade berokkenen dan een grote laesie op één moment van dezelfde grootte als de som van de serie kleine beschadigingen. Afhankelijk van allerlei variabelen, zoals de plaats waar de laesies in de hersenen worden toegediend, is het een robuust effect gebleken. Het effect ondersteunt zelfreparatie, omdat het de gedachte ondersteunt dat kleinere laesies makkelijker zijn te herstellen: er is bij kleine beschadigingen meer informatie aanwezig voor reparatie. Bovendien impliceert het effect dat er reparatie is na beschadiging. Deze hypothese wordt verder ondersteund door anatomische data waarbij tijdens een serial-lesion experiment is gemeten dat de connectiviteit verandert in tussenliggende herstelperioden, hetgeen er op wijst dat er inderdaad reparatie plaatsvindt na beschadiging.

Neurale netwerk modellen

We hebben in dit proefschrift geen nieuwe dierproeven of experimenten met mensen gedaan Er is voornamelijk onderzocht of zelfreparatie in de neurale netwerken van het brein zou kunnen werken en hoe dit dan precies zou werken. Het proefschrift gaat ervan uit dat hersenen werken volgens het principe van connectionistische of neurale netwerk modellen. Binnen deze klasse van modellen hebben we voor neurale netwerken gekozen die een groot aantal biologische eigenschappen delen met de hersenen van mensen en dieren. Binnen deze modellen zijn er ook weer verschillen in neurobiologisch detail en plausibiliteit.

Het model van zelfreparatie is als volgt: Er is een neuraal netwerk met een aantal geheugenrepresentaties. We nemen aan dat het netwerk onderhevig is aan laesies die de verbindingen tussen de netwerkknopen beschadigen. Zelfreparatie bestaat uit een stimulus die het netwerk activeert en een leerregel die de verbindingen in het netwerk kan veranderen. Zelfreparatie is succesvol als de veranderingen ten gevolge van laesies ongedaan worden gemaakt.

Het bovenstaande model is eerst getest met eenvoudige neurale netwerken (hoofdstuk 3). Uit dit onderzoek komen twee belangrijke parameters van het model naar voren: 1) het type stimulus waarmee het netwerk wordt geactiveerd en 2) de leerregel. De twee uiterste type stimuli waarbinnen het onderzoek is gedaan, zijn een type stimuli die gebruikt waren om

geheugen mee op te slaan en een type stimuli die geproduceerd werden door een gerandomiseerd proces. In het eerste geval is er een zeer grote waarschijnlijkheid dat er een correcte geheugenrepresentatie uit het geheugen wordt opgehaald. We noemen dit “gerichte” zelfreparatie. In het tweede geval heb je geen controle op welk geheugen wordt opgehaald en ook niet of het geactiveerde geheugen een correct geheugen is. We noemen deze vorm “autonome” zelfreparatie.

In het proefschrift hebben wij voornamelijk de Hebbian leerregel gebruikt. Dit is neurobiologisch gezien de meest plausibele leerregel in vergelijking met andere neurale netwerk leerregels. Relevant aan de Hebbian leerregel is dat er knopen worden versterkt als ze gelijktijdig aanstaan. Binnen deze leerregel zijn er variaties. In hoofdstuk 3 is het onderzoek gestart met een eenvoudige Hebbian leerregel. Deze leerregel bevat geen condities over een maximum of minimum aan de waarden van de sterkte van de verbindingen tussen de knopen. Uit het onderzoek in dit hoofdstuk blijkt dat deze leerregel alleen tot correcte netwerkreparatie leidt als er stimuli worden gebruikt die veel overeenkomst vertonen met stimuli waarmee de geheugenrepresentaties waren opgeslagen. Indien er een random stimulus wordt gebruikt, dan leidt dit tot een netwerk waarin maar één of enkele geheugenrepresentaties worden gerepareerd. In hoofdstuk 3 is tevens aangetoond welk type Hebbian leerregel nodig is om via random stimulatie een correcte netwerkreparatie te krijgen.

Het onderzoek van dit proefschrift heeft zich verder gericht op reparatie met random stimuli. De belangrijkste reden om autonome reparatie verder te onderzoeken is dat autonome zelfreparatie in neurale netwerken moeilijker is dan gerichte zelfreparatie. Als kan worden aangetoond dat autonome zelfreparatie mogelijk is, dan is gerichte zelfreparatie zeker mogelijk. Autonome zelfreparatie is moeilijk doordat een eenmaal geselecteerd geheugenrepresentatie voor reparatie een steeds grotere kans heeft op verdere selectie wat resulteert in een geheugenrepresentatie met een hele grote kans op selectie (“runaway repair”). Hierdoor worden er maar één of enkele geheugenrepresentaties in het netwerk gerepareerd. Met gerichte stimulatie heb je dit probleem niet, omdat je dan kan bepalen hoe vaak welke geheugenrepresentatie moet worden gerepareerd.

Hoofdstuk 4 biedt meer inzicht in autonome zelfreparatie door middel van een wiskundig kansmodel. Dit kansmodel drukt random stimuli en geheugenactivering uit in kansen. Het model bestaat uit twee lagen. Eén laag met invoerknopen en één laag met opgeslagen geheugens. Activatie in de invoerlaag wordt beregeld door een kansmodel waarbij elke knoop eenzelfde kans heeft om geactiveerd te worden. Hoe hoger de activatiekans, hoe hoger de intensiteit van de stimulus (uitgedrukt in het aantal knopen dat aangaat). Een

geheugenrepresentatie in de tweede laag is correct geactiveerd als er één of meerdere knopen van eenzelfde geheugenrepresentatie aanstaan. Staan er knopen aan van verschillende geheugenrepresentaties, dan wordt dit als een incorrecte activering beschouwd. Volgens de Hebbian leerregel worden knopen versterkt die gelijktijdig aanstaan. In het geval van incorrecte activering worden knopen uit verschillende geheugenrepresentaties met elkaar verbonden, hetgeen dus incorrecte reparatie is. Een resultaat van dit model is dat de intensiteit van de random stimulus laag moet zijn. Ditzelfde resultaat wordt ook gevonden in de complexe netwerken van de daarop volgende hoofdstukken. Het belangrijkste resultaat van hoofdstuk 4 is echter dat de kans op runaway repair niet groot is in een netwerk waarin zich veel geheugenrepresentaties bevinden. Met andere woorden: zelfreparatie in een brein, waar zeer veel geheugenrepresentaties opgeslagen zijn, lijkt mogelijk te zijn.

In hoofdstuk 5 wordt zelfreparatie in een meer neurobiologisch gedetailleerd netwerk onderzocht. Dit netwerk modelleert de sensorische invoer, thalamus en een deel van de cortex. Zelfreparatie vindt plaats in het corticale deel van het netwerkmodel. Een belangrijk verschil met de eenvoudige neurale netwerken van hoofdstuk 3 is dat er geen scheiding is tussen een fase waarin beschadigd wordt en een fase waarin wordt gerepareerd: beschadiging en reparatie vinden aldoor plaats en vinden daarmee in eenzelfde fase plaats. Opnieuw blijkt dat er een evenwicht is tussen zelfreparatie en beschadiging. Deze hangt af van allerlei parameters, zoals de leersnelheid van het netwerk en de laesiegrootte. Als een van de twee domineert, zal de structuur van het netwerk veranderen. Het belangrijkste resultaat van dit onderzoek is dat zelfreparatie in een zeer neurobiologisch plausibel netwerk mogelijk is.

Onze hersenen zijn afgestemd op de omgevingsstimuli door evolutie en ontwikkeling. Ook de stimuli die ons aangeboden worden tijdens leren zijn meer gestructureerd dan de random stimuli zoals gemodelleerd in dit proefschrift. In hoofdstuk 6 stellen we desalniettemin voor dat er zelfreparatie in het brein door middel van random stimuli kan plaatsvinden. Dit idee wordt ondersteund door een demonstratie van autonome zelfreparatie in een neuraal netwerk slaapmodel. Verder literatuuronderzoek laat zien dat slaaptheorieën de ideeën van redundantie en het algoritme van zelfreparatie delen met de theorie van zelfreparatie. In het bijzonder vertonen de slaaptheorieën van geheugenonderhoud (van fylogenetische geheugenrepresentaties) en van geheugenconsolidatie (van nieuw geheugen) veel overeenkomsten. Experimentele data die een van de twee slaaptheorieën ondersteund, ondersteunen tevens de theorie van zelfreparatie. Het is moeilijk uit de data op te maken welke theorie het meest waarschijnlijk is: alle drie de theorieën delen dezelfde algoritmische eigenschappen maar hebben een ander doel. Vanuit evolutionair oogpunt lijkt het niet vreemd

dat het doel zelfreparatie is, omdat zelfreparatie als een immuunsysteem voor het geheugen kan werken. Met dezelfde mechanismen kan later als een (zeer gunstig) bijeffect ook onderhoud aan het fylogenetische geheugen worden gepleegd en/of nieuw geheugen worden geconsolideerd.

Conclusies en belangrijkste bijdragen van het proefschrift

Een eerste doel van dit proefschrift was om aan te tonen dat zelfreparatie mogelijk is in de neurale netwerken van het brein. Er wordt aangetoond dat autonome zelfreparatie mogelijk is in artificiële neurale netwerken. Daar autonome zelfreparatie moeilijker is dan gerichte zelfreparatie, tonen we tevens aan dat gerichte zelfreparatie mogelijk is in artificiële neurale netwerken. In het brein vindt er hoogstwaarschijnlijk autonome zelfreparatie plaats door iets dat lijkt op gerichte zelfreparatie in artificiële neurale netwerken. Ervan uitgaande dat de artificiële neurale netwerken van dit proefschrift de essentiële of relevante bestanddelen van een “echt” brein bevatten, dan is met de simulaties in dit proefschrift ook aangetoond dat autonome zelfreparatie in het brein mogelijk is.

Een tweede doel van dit proefschrift was om een basis te leggen voor neurale netwerkmodellen van revalidatie. De zelfreparatie netwerken die zijn onderzocht in dit proefschrift hebben een aantal essentiële parameters van een echt brein. In hoofdstuk 2 is het type stimulus waarmee gerepareerd wordt onderzocht, alsmede het type leerregel. Andere belangrijke parameters voor zelfreparatie in het brein, die zijn onderzocht, zijn de grootte van de laesie, de frequentie van reparatie, de frequentie van beschadiging, etc. In dit proefschrift is een specifieke opzet met bepaalde parameterwaarden onderzocht. Deze opzet had een hoge frequentie van reparatie en beschadiging, waarbij er sprake was van kleine beschadigingen. Deze opzet komt het meest overeen met een model voor zelfreparatie in een gezonde proefpersoon (hoofdstuk 2). Andere typen van zelfreparatie en beschadiging kunnen gemodelleerd worden door de parameterwaarden van het in dit proefschrift gepresenteerde model van zelfreparatie te veranderen, hetgeen de basis legt voor modellen van revalidatie. Hierrmee is ook het tweede doel van dit proefschrift verwezenlijkt.