

# Selfrepairing Neural Networks: a Model for Recovery from Brain Damage

Jaap M.J. Murre<sup>1,2</sup>, Robert Griffioen<sup>1</sup>, and I.H. Robertson<sup>3</sup>

<sup>1</sup>University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

jaap@murre.com, griffioen@pobox.com

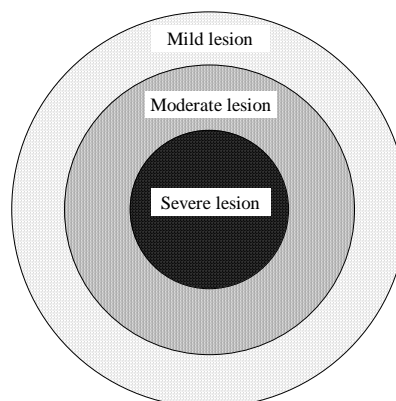
<sup>2</sup>University of Maastricht, The Netherlands

<sup>3</sup>Trinity College, Ireland

**Abstract.** We introduce selfrepairing neural networks as a model for recovery from brain damage. Small lesions are repaired through reinstatement of the redundancy in the network's connections. With mild lesions, this process can model autonomous recovery. Moderate lesions require patterned input. In this paper, we discuss implementations in three types of network of increasing biological plausibility. We also mention some results from random graph theory. Finally, we discuss the implications for rehabilitation theory.

## 1. Introduction

Brain damage incapacitates hundreds of millions of people and hundreds of thousands of professionals are involved in its treatment using a great variety of therapies. Despite this worldwide effort, the theoretical basis for treatment of recovery from brain damage is scant. We have recently developed a new framework for the study of recovery from brain damage [1], proposing a triage of recovery patterns that depend on the severity of the lesion (Fig.1). A mild lesion heals spontaneously through autonomous recovery, but there is no recovery from severe lesions; other brain areas must compensate for the loss of function. Moderate lesions show recovery, in particular with rehabilitative input. Since the work of Luria [2,3], theorizing in rehabilitation has mainly focused on fostering effective compensation strategies. We argue, however, that with the triage, rehabilitation through guided recovery is possible in case of moderate lesions. In this paper, we discuss an elaboration of our framework, using selfrepair in neural networks as a model of guided recovery. In



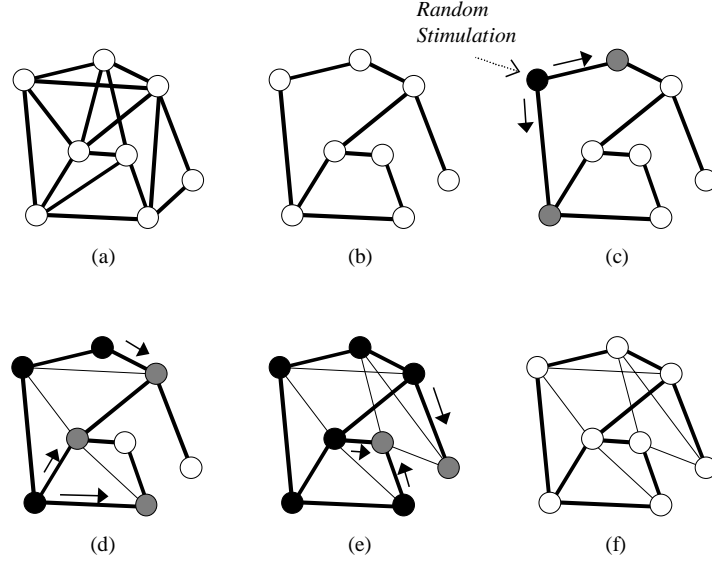
**Fig. 1.** Classification of different degrees of lesioning caused by brain damage.

these networks a repair mechanism is operative that maintains the redundancy present in the connections at a ‘safe’ level, despite perturbations of the network structure. We argue that such mechanisms underlie spontaneous recovery from brain damage and that guided recovery can be achieved by augmenting the spontaneous repair mechanism by suitable, patterned input.

Selfrepair-through-maintenance is ubiquitous in nature. It can be found in nearly all organisms on many different levels. Indeed, the very fundament of evolution, the double helix structure, allows DNA to repair itself [4]. A striking example of the power of selfrepair is the *Dienococcus radiodurans*, a radiation-resistant bacterium that is able to survive under conditions of starvation and oxidative stress. Its DNA selfrepair and genetic redundancy enable the organism to withstand severe ionizing and ultraviolet irradiation effects [5]. There is strong evidence that neurons in the brain are also under constant attack from damaging influences, causing synapses to be destroyed on a large scale [6]. Without some type of repair process, the cumulative effects of such minute lesions would soon cause our memories and cognitive capacities to vanish. Though representations in the brain may have considerable redundancy, this by itself does not ensure a long lifetime. Let’s take a simple analogy. Suppose ten agents are each guarding a copy of some crucial document (‘neural memory representation’). Once a month, they meet and lost copies are replaced (the ‘repair cycle’). This process may continue until some month, by chance, all copies are found to be lost. Even with a 50% loss chance per copy, the combined monthly survival probability is  $1 - 0.5^{10} \cong 0.999$  and the expected lifetime of the document is 85 years (1023 months). Without the monthly ‘repair’ session, however, the expected lifetime drops to a mere 4 months, despite the document’s ten-fold redundancy. In neural networks, redundancy resides in the synaptic connections of the network, each supporting a fragment of a representation. Any repair must, therefore, focus on reinstating lost connections. We have analyzed such processes in artificial neural networks, demonstrating that the effects of gradual lesioning can be fully undone if repair cycles follow the lesion (Fig. 2). If the lesions are too large, some memory representations will be beyond repair. We have analyzed the conditions for this using the theory of random graphs.

## 2. Simulations with Attractor Networks

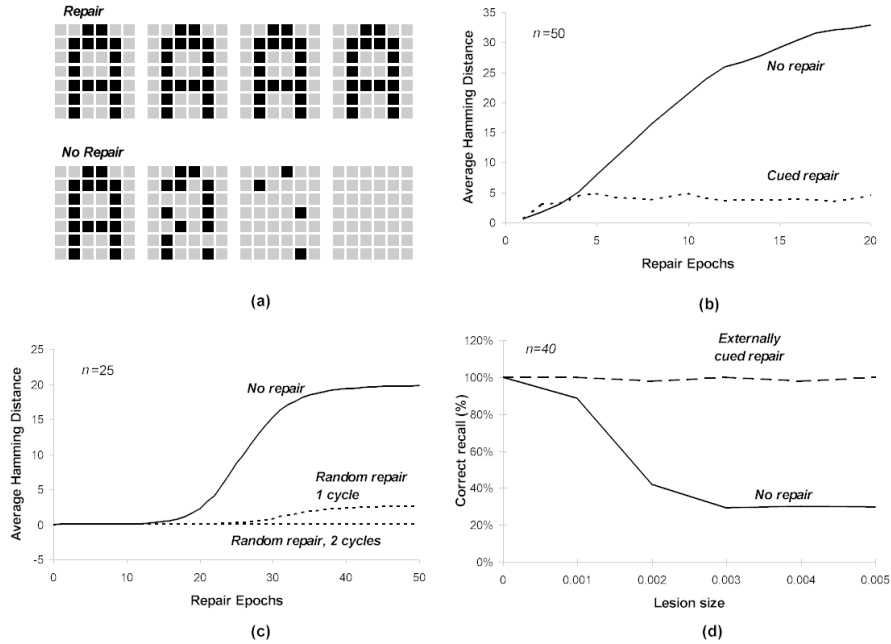
Neural networks are fault tolerant. This means that even if all synapses are randomly perturbed, an incomplete cue may still lead to perfect recall. In a Hopfield network [7], the weight on a synapse from neuron  $j$  to  $i$  is formed by several increments  $T_{ij}^s$ , each coding the co-occurrence of binary signals in a specific pattern  $S$ . If we have stored some patterns in a neural network, small lesions can be undone in the following manner. At each time step  $t$ , a perturbation  $e_t$  is added to every synapse, and this lesion is followed by a repair cycle during which we use some (partial or random) cue to recall a pattern  $\tilde{S}$ . We undo part of the perturbation by storing  $\tilde{S}$  again, incrementing the existing synapses with the Hopfield (Hebbian-type) learning rule:



**Fig. 2.** Schematic illustration of autonomous 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 vulnerable to further lesioning, compared to its pre-repair state (b).

$\tilde{T}_{ij}^{\tilde{S}} = (2\tilde{V}_i^{\tilde{S}} - 1)(2\tilde{V}_j^{\tilde{S}} - 1)$ , where  $\tilde{V}_i$  is the  $i$ -th element in  $\tilde{S}$ . This is done for all stored patterns. For a specific pattern  $S$  we now have for each synapse two increments plus perturbation:  $T_{ij}^S(t) + \tilde{T}_{ij}^{\tilde{S}}(t) + e_i$ . Repair is completed by normalizing the synapse strength (in this case by division by 2). With perfect recall,  $\tilde{S} = S$  and  $T_{ij}^S(t) = \tilde{T}_{ij}^{\tilde{S}}(t)$ , giving  $T_{ij}^S(t+1) = \{T_{ij}^S(t) + \tilde{T}_{ij}^{\tilde{S}}(t) + e_i\} / 2 = T_{ij}^S(t) + \frac{1}{2}e_i$ . In other words, a single repair cycle with normalization will reduce the effect of a perturbation by 50%. Multiple repair cycles can diminish perturbations to arbitrarily low levels.

In Fig. 3, we see that accumulated noise rapidly degrades the performance of the non-repaired network, but the repaired network continues to function. In Fig. 3.b, Hamming distance to target pattern is plotted against repair epochs. In Fig. 3.c, we use an asymmetric variant of the Hopfield learning rule with bounded weights  $T_{ij}^S = \max\{\min(\sum_S V_i^S [2V_j^S - 1], 1), -1\}$ . The network was trained with five non-overlapping patterns and ‘lesioned’ at each time step by setting 10% of the connections to 0. Contrary to the cued-approach of Fig. 3.b, here repair was initiated by a *random* cue. Repeating the cue-repair cycle twice within one epoch led to longer lifetimes, because it increased the probability that each pattern was repaired at least once (this is not ensured with randomly cued repair). Tests with patterns that



**Fig. 3.** Neural networks simulations of selfrepair ( $n$  gives the number of replications). (a) Snapshots from two networks in a single replication such as described in (b) but smaller. (b) Uniform noise in  $[-2.0, 2.0]$  was added at regular time steps to two neural networks with 100 neurons, trained with five randomly generated patterns. The self-repairing network was cued with patterns with 10% distortion and repaired as described in the text. (c) Autonomous repair (i.e., with random ‘cues’) in a modified Hopfield network. (d) Repair in a model of long-term memory and amnesia (see text).

overlapped about 18% also gave stable selfrepair as long as lesions did not exceed 1% (not shown). Such lesions disintegrate unrepaired patterns in about 350 lesion-repair cycles. In Fig. 3.d, we have recall after 20 lesion-repair epochs in the ‘cortex’ part of the TraceLink model of long-term memory and amnesia [8]. Here, representations were cued for repair from another network (not lesioned), simulating how one healthy area in the brain can cue a damaged one for repair. This model has non-negative weights, sparse activations, stochastic nodes, and Hebbian learning. Selfrepair was halted when the summed net input over all neurons exceeded a preset threshold. The process was stable if the lesion size remained below 0.15.

### 3. Connectivity Analysis

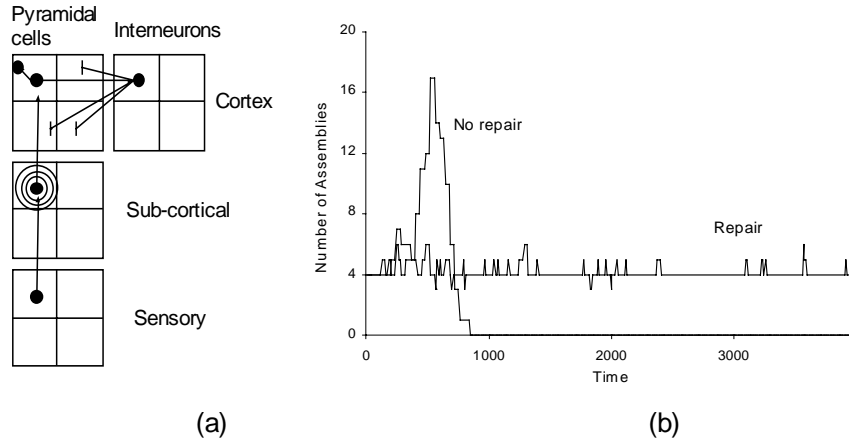
Successful repair hinges on near-perfect retrieval. We have analyzed the limits of perfect retrieval with theory of random graphs. An intact memory representation can be viewed as a connected random graph with  $N$  nodes (i.e., a path exists between each pair of nodes in the graph). Random deletion of connections (‘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 and Hebbian-type learning can randomly add connections. For large graphs, the probability of a graph being connected,  $p$ , is nearly a step function of the probability,  $f$ , that there is a connection between any given pair of nodes:  $p = \exp[-\exp\{-(fN - \log N)\}]$  [9, 10]. Small graphs require a much higher minimal  $f_c$  to ensure being connected (e.g., for  $N=100$  and  $f = 0.10$   $p \approx 1.000$ , but for  $N=10$  and  $f < 0.163$ ,  $p \approx 0$ ). If as a result of lesioning  $f$  falls below  $f_c$ , the representation is beyond repair: repair will be incomplete or even maladaptive, leading to dysfunctional circuits [1]. One implication of the graph theoretical model for brain theory is that large representations may survive gradual lesioning much longer than smaller ones.

We now extend the example of the security agents 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  $f$ . 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). The results of this simple, but enlightening exercise indicate that the lifetime of the repair process is nearly indifferent to a reduction in connection density  $f$  until a critical limit is approach. A small decrease in  $f$  will then bring down the lifetime to a very low value, or indeed to zero (i.e., instant loss of the representation).

#### 4. Simulations of Repair in Somatosensory Cortex

To study whether selfrepair in more neurobiologically informed networks is plausible we implemented a model of somatosensory cortex with selfrepair. The model consists of a sensory-, a sub-cortical-, and a cortical map, each with 100 neurons (Fig. 4.a). The cortical map has 100 excitatory neurons and 100 interneurons. The model neuron is a simplification of the MacGregor neuron [11], which in turn is derived from the Hodgkin-Huxley neuron [12]. A Singer-Hebb learning rule [13] was used: when the postsynaptic neuron is active at time  $t+1$ , a weight increases or decreases with amount  $\mu$  ( $0 \leq \mu \leq 1$ ) if the pre-synaptic neuron is active or inactive at time  $t$ .



**Fig. 4.** Architecture (a) and simulation (b) of the selfrepair in a model of somatosensory cortex.

The network was seeded with four distinct assemblies. The stochastic nature of the model neuron hampers repair because random neurons may become active at any time. Stable long-term repair could be achieved if the level of damage was in equilibrium with the amount of repair.

Fig 4.b shows the number of clusters for two simulations: a simulation of selfrepair with a learning rate of 0.002 and a lesion size of 0.005, and a simulation without selfrepair but with a lesion size of 0.008. In the first simulation, the network is very active in the beginning; the number of assemblies changes often, but after about 1500 time steps, the network stabilizes. The network retains its initial number of assemblies throughout the simulation. In short, for these parameters the network exhibits successful selfrepair. We also ran the network for 40,000 lesion-repair cycles, during which it remained stable, further reinforcing the viability of this approach to selfrepair. In the second simulation, without selfrepair and with a lesion size of 0.008, all assemblies of the network eventually disappear, as expected.

## 5. Discussion

Hopfield, Feinstein and Palmer [14], and Crick and Mitchison [15] have postulated that REM sleep may function to clean up memories in the brain by removing unwanted associations through unlearning. Here, we demonstrate how the exact opposite process can function to safeguard the integrity of neural memory representations. Our approach to neural repair is similar to consolidation in long-term memory, often assumed in models of retrograde amnesia [8,16,17,18]. For the type of long-term memory consolidation in these models, there is evidence that it occurs in the slow-wave (i.e., non-REM) sleep phase [19]. We can, thus, hypothesize that the

brain has unlearning (cleaning) and repair phases and that these occur in specific sleep phases.

Our modeling work demonstrates that it is feasible that processes of lesioning and repair go on continuously in a healthy brain. We also believe that the repair mechanism has an important function during recovery from brain damage. For example, our neural network simulations and the graph theoretical analyses both predict that if selfrepair processes are active, many small lesions will result in much less performance loss compared to a single large lesion which size equals the cumulative size of the small ones. This effect has indeed been found and is known as the *serial lesion effect*. [20] review how several consecutive lesions, administered over a relatively long time period, have less dramatic behavioral consequences than one lesion of the same size. This effect has been observed both in experimental animals and in human patients. One explanation for this effect is the serial recovery hypothesis, which states that repair processes compensate damage in between lesions. The serial lesion effect has been observed frequently in human cancer patients since 1836, when the French physician Dax observed that sudden damage to the left hemisphere was far more likely to produce aphasic symptoms than slowly developing damage. In general, a slow growing tumor damages the patient's neural tissue but without initial behavioral consequences compared to patients with acute brain damage of equal size.

We also explored the limits of the selfrepair process (not all reported here). If the lesion size is too large, for example, maladaptive repair may occur and faulty rewiring may spring into existence and persist. The graph theoretical analysis above provides guidance to when this can be expected. Also, if the learning and cueing processes are not well calibrated, one pattern may become much stronger than the others and take over all resources. With proper constraints on cueing and induced weight changes this problem can be avoided. Above, we used bounded weights for this reason.

In Robertson and Murre [1] we develop a more encompassing framework for the study of rehabilitation from brain damage, in which selfrepairing networks are complemented with other ways to stimulate recovery of damaged circuits: providing targeted bottom-up and top-down inputs, maintaining adequate levels of arousal, and releasing inhibition by—often contra-lateral—competitor circuits that may suppress activity. The most important clinical implication is that residual capacity can lead to true recovery, but only with early diagnosis and suitable therapy that avoids compensatory strategies.

## References

1. Robertson, I.H., Murre, J.M.J.: Rehabilitation of brain damage: brain plasticity and principles of guided recovery. *Psychological Bulletin* 125 (1999) 544-575
2. Luria, A. R.: *Restoration of function after brain injury*. Pergamon, Oxford (1963).
3. Luria, A.R., Naydin, V.L., Tsvetkova, L.S., Vinarskaya, E. N.: Restoration of higher cortical functions following local brain damage. In Vinken, P.J.,

- Bruyn, G.W.: Handbook of clinical neurology vol. 3. Elsevier, New York (1975) 368-433.
4. Ayala, F.J., Kigger, J.A.: Modern genetics. Benjamin/Cummings, Menlo Park, CA (1982)
  5. White, O.J., Eisen, A., Heidelberg, J.F., Hickey, E.K., Peterson, J.D., Dodson, R.J., Haft, D.H., Gwinn, M.L., Nelson, W.C., Richardson, D.L., Moffat, K.S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, P., Lam, L., McDonald, T., Utterback, C., Zalewski, K.S., Makarova, J.J., Aravind, L., Daly, M.J., Minton, K.W., Fleischmann, R.D., Ketchum, K.A., Nelson, K.E., Salzberg, S., Smith, H.O., Venter, J.C., Fraser, C.M.: Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* 286 (1999) 1571-1577
  6. Cotman, C.W., Nieto-Sampedro, M.: Brain Function, synapse renewal, and plasticity. *Annual Review Psychology* 33 (1982) 371-401
  7. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA*. 79 (1982) 2554-2558
  8. Murre, J.M.J.: TraceLink: a model of amnesia and consolidation of memory. *Hippocampus* 6 (1996) 675-684
  9. Erdős, P., Rényi, A.: On random graphs I. *Publ. Math. Debrecen* 6 (1959) 290-297
  10. Bollobás, B. *Random Graphs*. Academic Press, London (1985)
  11. MacGregor, R.J., Oliver, R.M.: A model for repetitive firing in neurons. *Cybernetik* 16 (1974) 53-64
  12. Hodgkin, A.L., Huxley, A.F.: A quantitative description of ion currents and its application to conduction and excitation in nerve membranes. *Journal Physiology (London)* 117 (1952)
  13. Singer, W.: Ontogenetic self-organization and learning. In McGaugh, J.L., Weinberger, N.M., Lynch, G. (eds.): *Brain organization and memory: cells, systems, and circuits*. Oxford University Press, Oxford (1990) 211-233
  14. Hopfield, J.J., Feinstein, D.I., Palmer, R.G.: 'Unlearning' has a stabilizing effect in collective memories. *Nature* 304 (1983) 158-159
  15. Crick, F. and Mitchison, G.: The function of dream sleep. *Nature* 304 (1983) 111-114
  16. McClelland, J.L., McNaughton, B.L., O'Reilly, R.C.: 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 (1995) 419-457
  17. Nadel et al. (2001)
  18. Alvarez, R., Squire, R.L.: Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of National Academy of Sciences (USA)* 91 (1994) 7041-7045
  19. Wilson, M.A., McNaughton, B.L.: Reactivation of hippocampal ensemble memories during sleep. *Science* 255 (1994) 676-679
  20. Finger, S., Stein, D. G.: *Brain damage and recovery: Research and clinical perspectives*. Academic Press, New York (1982)