A NEURAL NETWORK MODEL OF MEMORY
AND HIGHER COGNITIVE FUNCTIONS

David D. Vogel, Ph.D.
Ross University
Portsmouth
Commonwealth of Dominica

Correspondence:
David Vogel, Ph.D.
Ross University
Portsmouth
Commonwealth of Dominica

E-mail: dvogel@rossmed.edu.dm
Telephone: 1-767-445-5355 Ext. 287
Fax: 1-767-445-3457
A NEURAL NETWORK MODEL OF MEMORY
AND HIGHER COGNITIVE FUNCTIONS

ABSTRACT

I first describe a neural network model of associative memory in a small region of the brain. The model depends, unconventionally, on disinhibition of inhibitory links between excitatory neurons rather than long-term potentiation of excitatory projections. The model may be shown to have advantages over traditional neural network models both in terms information storage capacity and biological plausibility. The learning and recall algorithms are independent of network architecture, and require no thresholds or finely graded synaptic strengths. Several copies of this local network are then connected by means of many, weak, reciprocal, excitatory projections that allow one region to control the recall of information in another to produce behaviors analogous to serial memory, classical and operant conditioning, secondary reinforcement, refabrication of memory, and fabrication of possible future events. The network distinguishes between perceived and recalled events, and can predicate its response on the absence as well as the presence of particular stimuli. Some of these behaviors are achieved in ways that seem to provide instances of self-awareness and imagination, suggesting that consciousness may emerge as an epiphenomenon in simple brains.

Key words
cognition; conditioning; consciousness; memory; neural network; refabrication; R-net; sparse
A NEURAL NETWORK MODEL OF MEMORY AND HIGHER COGNITIVE FUNCTIONS

1. INTRODUCTION

Prima facie, trying to model the cerebrum at the neural network level is a fool’s game. The architecture and cellular function are only partially known. Moreover, any evolved control system is almost certain to be unnecessarily complex, and any model that represents such systems with methodological economy is unlikely to represent them accurately.

However, even models that are in fundamental respects inconsistent with the known cerebrum have provided insights into mechanisms that may contribute to actual cortical functioning. Marr’s (1971) model of the hippocampus has influenced research on that structure for 3 decades.

Moreover, it is extremely difficult for people to imagine how the workings of a machine can explain the existence of a mind. The allegedly transcendent mind is a correspondingly powerful object of the sort of mystification and special pleading that has long been inimical to scientific inquiry.

This paper details a biologically plausible neural network that is able to account for a variety of psychological phenomena including serial memory, classical and operant conditioning, secondary reinforcement, predictive fabrication, and refabrication of incomplete memories. Random excitation of the network allows it to discover sensory consequences of motor activities and to subsequently repeat motor activities by recalling their sensory consequences. The network distinguishes between recalled and perceived stimuli and is able to predicate behaviors, not only on the presence of some particular stimuli, but on the absence of others. Some of these properties are achieved in ways that seem to provide instances of self-awareness and imagination and suggest that consciousness may emerge as an epiphenomenon in relatively simple brains.

The network is “biologically plausible” in the sense that both its local architecture and its algorithms, while simplified, are consistent with those of the known brain. “Plausibility” in this sense is not the only matter of interest in brain models. Some models retain interest even though they are “neural networks” much more in the engineering than the biological sense of the phrase (e.g., Jani and Levine, 2000; Salum et al., 2000). Many other networks use large-scale block designs that are analogous to structures in the brain but do not specify biologically realistic connection matrices (e.g., Chechnik et al., 2001; Greenstein-Messica & Ruppin, 1998; Hasselmo & Schnell, 1994; Taylor & Taylor, 2000). Yoshida et al. (2002) provide interesting insights into aspects of recall without providing a mechanism of engram formation. Models whose local architectures are more representative of the real cerebrum are likely to be restricted to lower level problems (e.g., Fukai, 1999; Körner et al., 1999), but incorporate features that will likely prove useful in some as yet unrealized synthesis.

Many attempts at representing significant aspects of the cerebrum with realistic architectures
have involved models of associative memory derived from Marr's (1971) model of archicortex (e.g., Eichenbaum and Buckingham, 1990; Gibson, and Robinson, 1992; Treves and Rolls, 1994). These models depend on strengthening of excitatory synapses in a way that is thought of as analogous to long-term potentiation (LTP).

Section 2 describes a model of a small region of the brain that has been less fully developed in previous papers (Vogel, 1998; 2001). Memory, in this model, does not depend on LTP. Rather, it depends on the weakening of inhibitory links (disinhibition). Despite the small number of inhibitory neurons in the model, it provides greater storage capacity than realistic LTP based models. These model networks have previously been referred to as “R-nets” to distinguish their random architectures from those of mathematically more tractable “P-nets.” Individual R-nets are also referred to, here, as “regions.”

Section 3 describes a more complex network (a “C-net”) consisting of several R-nets variously regarded as sensory, motor, integrative or reinforcing on the basis of their positions in the network.

R-nets are sparsely connected (i.e., each neuron projects onto only a small fraction of the other neurons in the network). Information stored in sparsely connected networks must be sparsely coded (i.e., an “event” is represented by firing a small fraction of the neurons in the network). In real brains, primary sensory information is densely coded. The process of feature extraction presumably generates sparse coding in “higher” regions of real brains. All of the networks described, here, use information that has already been sparsely coded. Near relatives of these networks are able to convert dense to sparse coding, and I do not imagine that real brains make an exclusive distinction between mnemonic and encoding processes.

Analogies between particular regions of the network and particular regions of real brains are not drawn. The learning and recall algorithms used are suited to many different architectures, and many parts of the brain might be used to achieve functions like those modeled, here. As implemented, the architecture looks most like archicortex. However, it seems likely that evolution has conserved mechanisms, not only between archi- and neocortex, but between brainstem and cortex. The only architectural requirement of the model is that sufficiently many excitatory neuron pairs be linked by inhibitory neurons.

The model does not incorporate a number of conspicuous features of neocortex including its layered structure and cortical columns. Work that incorporates these features into models of such phenomena as rapid sensory perception (Körner et al., 1999) or spatial working memory (Tanaka, 1999) develops important, complementary ideas about neocortex.

### 2. A MODEL OF A LOCAL REGION (R-NET)

#### 2.1 Traditional Neural Networks

In models that follow the work of Marr (1971), excitatory neurons, analogous to principal
cells, are randomly connected to one another. In operation, a subset of the neurons in a network is activated. I will refer to this subset as a “training set” (though it would more commonly be called a “training vector”). Synapses between active neurons are then trained according to a “Hebbian” learning rule, which specifies that the strengths of excitatory synapses be increased when both the pre- and postsynaptic neurons are active at the same time. After training on some number of sets, a subset of a training set (a “recall set”) may be activated with the objective of reactivating the original training set (the “target set”). Because of previous synaptic training, members of a target set are likely to be more strongly activated by a recall set than non-target neurons. An activation threshold for firing may be computed stochastically such that most target neurons are above the threshold while most non-target neurons are not. As the number of stored training sets increases, the number of recall errors also increases. Errors may consist of active neurons that are not members of the target set (“spurious neurons”) or inactive members of the target set.

As the size of the recall set increases, the activation of all neurons increases, and the threshold that separates target from non-target neurons increases. Accordingly, traditional networks posit a uniform, inhibitory feedback onto all neurons that is proportional to the number of neurons in a recall set.

These traditional networks provide some insight into principles by which real brains may store information. However, they have several features that are biologically implausible.

First, the connectivity of the real cortex is at the lower edge of the connectivity required for successful information storage by these neural networks, which remain brittle with respect to training and recall set sizes. The relatively sophisticated network of Bennett, Gibson, & Robinson (1994) does not effectively store training sets either much larger or much smaller than 330 neurons.

Second, the networks are not robust enough to tolerate stochasticity in the inhibitory feedback.

Third, the mechanism of uniform inhibitory feedback is not made explicit, and no such mechanism seems to exist in the real brain. Single active inhibitory synapses appear to be extremely effective at silencing postsynaptic neurons (Buckmaster & Schwartzkroin, 1995; Conors, Malenka, & Silva, 1988; Miles et al., 1996). No finely graduated inhibitory cloud is present.

Fourth, synaptic strengths must be finely graduated and are dependent on architectural parameters such as connectivity. This seems to make traditional networks evolutionarily implausible because any change in architecture deoptimizes synaptic parameters.

Fifth, the brain is not randomly connected. While non-random connection matrices exist that are stochastically as smooth as random matrices (Vogel & Boos, 1997), they are relatively rare objects that are not likely to be discovered by evolutionary means.

Sixth, the analysis of the storage capacities of large versions of these networks is suspect. The problem arises in the use of DeMoivre’s Theorem to approximate a binomial distribution by means of a normal distribution (Marr, 1971). For small networks, equations for the number of errors during recall give results that closely approximate the actual number of errors found in simulations. It seems
natural to assume that as network size increases the normal approximation of a binomial distribution will only improve. However, it is not the network size that is the number of trials in the binomial distribution, it is the recall set size, and the recall set size typically remains constant in moving from simulations of small networks to analyses of large networks. It sometimes appears (Gibson & Robinson, 1992) that the number of errors in small networks is slightly larger than analysis predicts. If the number of neurons in a recall set remains constant, the number of spurious neurons will increase linearly with network size, and a small under-estimate of the number of errors in a small network becomes large in a large network.

2.2 The Local Network Model (R-net)

In R-nets, direct projections between excitatory neurons are ignored. Rather, there are only random projections of excitatory neurons (analogous to principal cells) onto a relatively small number of inhibitory neurons (analogous to interneurons), and recurrent projections from the inhibitory neurons onto excitatory neurons (Fig. 1). Pairs of primary neurons are then said to be “linked” by inhibitory pathways of length 2. Of the various kinds of interneurons present in the cortex, the basket cells are most common and bear the greatest resemblance to the inhibitory neurons of the model.

Coactivation of two excitatory neurons causes both synapses in any link between the neurons to become “trained.” The synapses are binary, being trained or not. Training of both synapses in an inhibitory link functionally “cuts” the link. Unless both synapses are trained, inhibitory links are absolutely inhibitory, excitatory neurons firing on the \( x \)th cycle if and only if they receive no inhibition from neurons firing on the \( (x-1) \)th cycle.

This disinhibition may be modeled in either of two ways. The obvious mechanism is simply to weaken both synapses so the pathway is no longer inhibitory. An interesting alternative is to strengthen both synapses noting that “inhibitory” GABA synapses are excitatory when sufficient \( \gamma \)-aminobutyric acid (GABA) is released. The second interpretation is implemented in the current studies.

It is important to note that no thresholds are required under either interpretation. Excitatory neurons are binary, firing if they are not inhibited. Inhibitory neurons have activities equal to the sum of their inputs.

Employed as local regions in C-nets, these R-nets have additional properties. To accommodate a noisy environment, neurons are allowed to accumulate inhibition that decays linearly with time. If a recall set is active along with spurious neurons, the spurious neurons will typically acquire more inhibition than target neurons. As the inhibition of all neurons decays, target neurons are likely to fire first and inhibit non-target neurons. In addition, in simulations of recall of the smallest training sets (20-neuron sets), trained inhibitory links have been made slightly excitatory. This improves storage somewhat by reducing the probability of a single spurious neuron silencing a large enough fraction of target neurons to prevent recall. Accordingly, the activation algorithm for the network is as follows.

The network is initialized by activating a small set of excitatory neurons. Inhibitory neurons are then synchronously updated according the following rule.
\[ a_{i,x} = \sum w_{i,e} a_{e,x} \]  

Eq. 1

where \( a_{i,x} \) is the activity of the \( i \)th inhibitory neuron on the \( x \)th cycle, \( a_{e,x} \) is the current activity of the \( e \)th excitatory neuron with possible values 0 or 1, and \( w_{i,e} \) is the strength of the projection of the \( e \)th excitatory neuron onto the \( i \)th inhibitory neuron with possible values of 1 (untrained) or 10 (trained). Excitatory neurons are then synchronously updated according to the following rule.

\[ I_{e,x} = I_{e,x-1} + \sum I_{i,x-1} + 1 \quad [I_{\min, e} < I_{e,x} = 0] \]  

Eq. 2

\[ a_{e,x} = 1 \quad \text{if} \ I_{e,x} = 0 \]

\[ a_{e,x} = 0 \quad \text{elsewise} \]

where \( I_{e,x} \) is the inhibition of the \( e \)th excitatory neuron on the \( x \)th cycle, \( a_{e,x} \) is the activity of the \( e \)th neuron on the \( x \)th cycle, and \( I_{i,x-1} \) is a function of the \( i \)th inhibitory neuron given by

\[ I_{i,x-1} = -a_{i,x-1} \quad \text{if} \ a_{i,x-1} < 10 \]

\[ I_{i,x-1} = -1 \quad \text{if} \ a_{i,x-1} > 10 \quad \text{and the projection of the \( i \)th inhibitory neuron is untrained.} \]

\[ I_{i,x-1} = t \quad \text{if} \ a_{i,x-1} > 10 \quad \text{and the projection of the \( i \)th inhibitory neuron is trained.} \]

\( I_{e,x} \) is held to values between \( I_{\min, e} \) and 0. In all of the studies reported, here, \( I_{\min, e} \) is -20+rand[0,5]. The small variation of \( I_{\min, e} \) prevents oscillations after many neurons are simultaneously inhibited to \( I_{\min, e} \).

I am unable to provide a formal analysis of the R-nets employed, here, without resorting to the same improper use of DeMoivre’s Theorem described above. However, a series of papers (Vogel and Boos, 1997; Vogel, 1998; Vogel, 2001) formally demonstrates the substantial storage capacities of networks progressively approximating the R-nets used here. The worst-case networks modeled, here, have only half the connectivity of networks used previously. However, the consequences are easily understood.

The chief consequence of very sparse connections is that some neurons may have relatively few links to elements of a recall set, and these links may become spuriously trained. (The synapse of an excitatory neuron onto an inhibitory neuron may be trained by one set while the synapse of the inhibitory neuron onto an excitatory neuron is trained by another.) The number of neurons with few links to a recall set increases linearly with the size of the network, and very small sets are recalled poorly in large, very sparsely connected networks. In compensation, the probability of a non-target neuron having no untrained links to elements of a recall set may be seen to be the product of the probabilities for each element of the recall set. With increasing recall set size, this product gets small rapidly, and the recall set size needed to suppress spurious neurons increases much less rapidly than the network size.

A further consequence of this poor performance when small recall sets are activated is seen when R-nets are required to recall large target sets from very small recall sets. This difficulty is largely overcome by allowing the R-net to cycle repeatedly. In most simulations reported, here, the network is allowed to pass through 100 cycles before its performance is evaluated, though an isolated R-net excited with 10 members of a 40-neuron training set will typically converge in about 20 cycles.
2.3 Biological Plausibility of the Architecture

As the histologic data are scant, it is impossible to clearly interpret the architecture of any relevant part of the brain. The architecture of the R-nets described below is based on the following, worst case interpretation of hippocampal architecture.

The projections of interneurons onto principal cells have been more extensively studied than those of principal cells onto interneurons (e.g., Buckmaster & Schwartzkroin, 1995; Gulyás et al., 1993; Sik, Tamamaki, & Freund, 1993). In a review article, Freund and Buzsáki (1996) estimate that typical interneurons of the rat hippocampus synapse on 1000 to 3000 principal cells. Primate (Ribak, Seress, & Leranth, 1993) and especially human (Seress et al., 1993) cortex has larger numbers of inhibitory neurons than the rat, and some of these interneurons appear to have much more extensive connections. For the purposes of this discussion, the number of principal cells innervated by each interneuron is taken to be 1500. The ratio of interneurons to principal cells in human hippocampus is roughly 0.2.

Studies of the principal cell fanout onto interneurons are not common. However, Sik, Tamamaki, and Freund (1993) describe the complete axonal arbor of a single pyramidal cell of subfield CA3 of the rat hippocampus. This single example formed asymmetric (presumably excitatory) synapses on 322 parvalbumin staining neurons (presumably inhibitory interneurons). 322 is probably a considerable underestimate of the typical number of synapses on interneurons because the pyramidal cell studied had a total of only 15,295 synapses, and only synapses on parvalbumin staining cells were counted as synapses on interneurons. Perhaps as many as half the basket cells in the hippocampus do not stain with parvalbumin (DeFelipe, 1997). However, 322 is taken, here, as the number of inhibitory neurons innervated by each excitatory neuron.

Electrophysiological results of Miles (1990) indicate that about 30% of principal cell pairs have inhibitory links. If another 10% have links that have been disinhibited, then assumptions of random connectivity, and the numbers of projections described above lead to a calculation of the number of principal cells in a network of $10^6$.

All R-nets simulated for this paper are constructed so that (1) the number of inhibitory neurons is 20% of the number of excitatory neurons, (2) the total number of projections onto inhibitory neurons equals the number of projections onto excitatory neurons, and (3) there are sufficient numbers of projections so that 40% of excitatory neuron pairs are connected.

It is likely that electrophysiologic studies (Miles, 1990) seriously under-estimate the connectivity of the hippocampus. The hippocampus of the rat only contains about 180,000 neurons (Akdogan et al., 2002; West et al., 1991). For a subregion with 60,000 excitatory and 6,000 inhibitory neurons, the numbers of synapses described above could link 99.97% of principal cell pairs. Such connectivity would increase the total storage capacity of these R-nets for sets of all sizes, though it would decrease the information stored per synapse for larger sets (Vogel, 2001).

2.4 Biological Plausibility of the Disinhibition Algorithm

Much of what is routinely thought of as “LTP” may be attributed to disinhibition. “LTP” may
be demonstrated in the Schaeffer collateral projections to subregion CA1 either by tetanizing the projections or by paired-pulse induction in which the postsynaptic neuron is depolarized during a Schaffer collateral volley. The induction of LTP by the mere tetanic stimulation of many synapses does not have any obvious computational use. However, paired-pulse induction is analogous to Hebbian learning.

The chief component of increased excitability following paired-pulse stimulation is long-term transformation of GABA synapses from inhibition to excitation (Collin et al., 1995; Nathan & Lambert, 1991). This transformation includes both GABA\textsubscript{A} and GABA\textsubscript{B} receptors (Desai et al., 1994). As pointed out by Staley et al. (1995), the firing rates (Buzsáki et al., 1992) of a majority of GABA\textsubscript{A}ergic neurons in vivo appear to be sufficiently high to be excitatory. Additionally, Stelzer et al. (1994) find that LTP of CA1 pyramidal cells induced by tetanizing Schaeffer collaterals is accompanied by disinhibition of dendritic, but not somatic, inhibitory synapses.

I know of no equivalent demonstration in local connections. However, in depolarization-induced suppression of inhibition (DSI) a 1 second depolarization of CA1 pyramidal cells (Alger et al., 1996; Morishita et al., 1997) or cerebellum Purkinje cells (Vincent et al., 1992) produces a 1 minute suppression of GABA\textsubscript{A}ergic inhibition. This is not a long-term effect. However, as the induction procedure does not involve both synapses in a link, it is unlike the procedure suggested by the model or employed in the studies of long projection pathways discussed above. Induction of DSI does appear to involve a retrograde messenger (Alger et al., 1996), possibly glutamate (Morishita et al., 1997), that acts on the inhibitory neuron. Such a retrograde messenger is a necessary component of any mechanism that would train both synapses in an inhibitory path of length 2, it being necessary that synapses on the inhibitory neuron be informed that the target excitatory neuron is firing. Moreover, induction of DSI is strongly facilitated by cholinergic agonists (Martin & Alger, 1999). This kind of dependence is needed to accommodate the reinforcement-induced learning employed in Section 3.

The slow decay of inhibition relative to excitation in the model is consistent with the slow decay of summed inhibitory postsynaptic potentials elicited by spike trains (Thomson et al., 1996).

2.5 Simulations of Simple R-nets.

Storage capacities of R-nets (Figs. 2 – 5) were evaluated by activating randomly selected sets of neurons and training synapses according to the rules given above. After some number of sets were trained, recall sets consisting of half the elements of a training set were activated, and the network was allowed 100 cycles in which to converge on the target set. The number of training sets was increased until the mean number of errors per target set was 10% of the target set size.

The R-nets employed in Figs. 2 through 5 vary in size from 3000 excitatory neurons (each with 18 projections onto 600 inhibitory neurons each with 85 reciprocal projections) to 75,000 excitatory neurons (each with 87 projections onto 15,000 inhibitory neurons each with 435 reciprocal projections). As described previously, large networks cannot store small training sets. The 50,000-excitatory-neuron R-net was unable to store 20- or 30-neuron training sets with only 10% errors. However, as seen most clearly in Figs. 4 and 5, small increases in recall set size result in large
increases in the range of network sizes that will effectively store the sets, and large networks are very
broadly tuned. It may be observed that the number of 40-neuron training sets stored is almost a linear
function of the number of synapses in the R-net over the entire range of R-nets studied.

When the training sets are relatively small, errors are chiefly due to the firing of spurious
neurons that have stochastically few links to the recall set. When training sets are relatively large,
errors are chiefly due to over-training and are like those of more densely connected networks in which
the relationship between storage capacities and numbers of synapses is similarly linear (Vogel &
Boos, 1997; Vogel, 1998, 2001). It seems reasonable, on the grounds that the probability of a
neuron being spurious is a product of probabilities, to approximate the relationship between optimal
training set size and network size as a power function. Acknowledging the risks of extrapolation, such
a function extrapolated to the $10^6$-neuron network described previously gives an optimal training set
size of roughly 100 neurons. If it is assumed that the R-nets studied here scale up in a manner
analogous to the moderately more densely connected R-nets previously studied (Vogel, 2001), the
$10^6$-neuron network would store more than 130,000 100-neuron sets. It may be anticipated that this
assumption under-estimates the storage capacities of the very sparsely connected networks presently
employed. Over the range of R-nets studied, the storage capacities per synapse of very sparsely
connected networks for large training sets progressively outstrip those of more densely connected R-

The R-net with 50,000 excitatory neurons is comparable in size to subregions CA1 or CA3 of
the rat hippocampus (Akdogan et al., 2002), though the model is probably less densely connected
than the real rat hippocampus. While a more densely connected rat hippocampus might store more
than the 2000 50-neuron training sets of the model, the capacity of the model seems adequate for a
subregion of the rat hippocampus.

3. A MULTI-REGION NETWORK (A C-NET)

3.1 General Considerations for a 2-Region Network

As in the real brain, two or more R-nets may be connected by projections of the excitatory
neurons of one R-net onto both excitatory and inhibitory neurons of another. The inhibitory links thus
formed are governed by the rules already given. Projections onto excitatory neurons undergo LTP
whenever pre- and postsynaptic neurons are active at the same time. LTP produces a small decrease
in the number of cycles needed for elements of a training set in one region to induce the firing of
elements of the training set in another region. For multi-region networks, Equation 2 becomes

$$I_{e,x} = I_{e,x-1} + \sum w_{e,j}a_{j,x-1} + \sum I_{i,x-1} + 1$$  \hspace{1cm} \text{Eq. 3}$$

where $w_{e,j}$ is the strength of the projection of the $j$th excitatory neuron onto the $e$th excitatory
neuron, and $a_{j,x-1}$ is the activity of the $j$th excitatory neuron on the previous cycle.

The large number of weak excitatory projections causes regions in which no training set is
active to fire neurons in sequences that appear random when observed at intervals of much more than
10 cycles. (At shorter intervals, some neurons may be seen to fire several times in succession.) These pseudo-random sets of neurons are used, below, to form representations in one region of activity in another. When a training set of more than about 20 neurons is active in a 4000-excitatory-neuron R-net, it largely suppresses the random firing of neurons.

### 3.2 Architecture and Rules

The architecture of one useful C-net is illustrated in Fig. 6. Each R-net consists of 4000 excitatory neurons, each with 20 projections onto inhibitory neurons, and 800 inhibitory neurons, each with 100 projections onto excitatory neurons. This network includes two “channels.” Each channel is a sequence of four R-nets designated “early sensory,” “sensory,” “motor,” and “late motor.” During simulations, only neurons in the early sensory region are forced to fire from outside the C-net.

Regions are reciprocally connected, except for there being no recurrent projections to early sensory R-nets. The early sensory regions are meant to reflect only the perceived environment of the network.

Each excitatory neuron of one region has 400 projections onto excitatory neurons and 5 projections onto inhibitory neurons of any other R-net within the same channel to which it projects. There are weaker connections to regions in other channels consisting of 2 projections onto inhibitory and 160 projections onto excitatory neurons. This architecture may be thought of as a blocky version of a real brain in which neurons project most densely to a small region and less densely to surrounding regions.

The network has two “integrative” regions, designated the “reward” and “punishment” regions, each of which has reciprocal connections to the motor and sensory regions of both channels. Both regions also receive projections from the early sensory regions. Because these integrative regions connect to a large number of local regions, projections of a neuron to or from these regions reach only 1 inhibitory and 80 excitatory neurons.

Finally, there is a “reinforcement” sensory region that connects only to the integrative regions (in the same reciprocal pattern as connects regions within a channel). In it are two pretrained 20-neuron sets designated the “reward” and “punishment” sets. By mechanisms not made explicit in the model, activation of the punishment set inhibits all activity in the late motor region. Additionally, activation of the punishment set from outside the C-net (as opposed to recall by activity in an integrative region) causes training to occur in all regions of the network, except the reward integrative region, while activation of the reward set causes training to occur in all regions of the network, except the punishment integrative region. A reinforcement set is regarded as active when 18 of its 20 elements are firing. The reader will please be careful to distinguish between the reward and punishment sets in the reinforcement region and the reward and punishment integrative regions.

### 3.3 Formation of Training Sets

Suppose some activities in the late motor regions result in corresponding activities in the early sensory regions as, for example, motor activity produces corresponding proprioceptive responses. Such correlated activity may be trained by activation of a reinforcement set. Later activation of a recall
set anywhere in the channel will result in activation of the corresponding late motor target neurons (i.e., recall of a sensory event reproduces the motor output that originally caused it). Mechanisms for storing repeatedly correlated activity that do not depend on reinforcement are discussed in Section 4.

In the C-nets described below, regions with no active training set spontaneously fire up to 10 neurons per cycle (with means of 4 to 6 depending on network details). A 10-neuron training set is not sufficiently large to reliably suppress random activity. However, training events may be made to produce larger training sets by a variety of means, most of which are not discussed, here. In studies of information storage capacity, I have simply activated a randomly selected set of neurons of the appropriate size at the time of training. In behavioral studies of freely running networks, any neuron that is active during the 25 cycles prior to firing of a reinforcement set is trained as though currently firing.

3.4 Storage Capacity

The parameter space of the network has not been optimized. The effort to do so would not be rewarded by proportional results. Rather, these results emphasize the robustness of the network. In all of the studies of this subsection, $w_{e,j} = 0.1$ or 0.4 (untrained or trained), and $t = 0.4$.

The storage capacity of the network was evaluated by creating training sets encompassing all regions of the network. For some studies randomly selected sets of 15 to 30 neurons per region were activated and trained. For others, random sets of variable size were formed by initializing the network with 4 randomly selected active neurons in each R-net, and assigning other neurons $I_{e0}$’s randomly selected between 1 and 20. The network then ran for 300 cycles at which time training was triggered by activation of a reinforcement set. The network was then reinitialized and the process repeated.

After storage of sufficiently many sets, recall sets were activated in the early sensory regions and 100 cycles were allowed for the remainder of the network to converge on corresponding target sets. The percentage of errors indicated in Fig. 7 represents the ratio of errors to the total number of target neurons in all R-nets except the reinforcement region.

When the network was allowed to generate training sets by means of its own random firing, sets in the early sensory region ranged in size from 26 to 39 neurons (mean 30). In other local regions, set size ranged from 14 to 27 neurons (mean 23). The number of these “random” sets stored is not limited by errors, which never exceeded 1.9% with 30 sets stored. Rather, in each of 10 trials, there was a sharp transition over the range of 30 to 35 sets stored. When a small number of randomly selected neurons is activated in a C-net in which up to 30 sets have been trained, the network will continue to cycle for thousands of cycles without converging on one of the training sets. However, by the time 35 sets have been trained, the network begins to converge on some training set within 100 cycles, making it impossible to add additional training sets by means of the protocol described above.

3.5 Serial Memory

A model of serial memory that is more satisfactory than those I have discussed previously (Vogel, 1995, 2001) is produced by training projections from “recently” firing motor neurons onto “currently” firing sensory neurons. By “recently” and “currently” I mean within the last 150 and 25
cycles, respectively. (In other studies, I have trained recently firing integrative projections onto currently firing channel neurons.) One of two mechanisms is then used to cause sensory neurons to cease firing before motor neurons. In some studies, I have employed burst-firing neurons, giving neurons a shorter period of firing in the sensory region than the motor region. In others studies, I have added the rule that activation of either reinforcement set suppresses firing of any active neuron in the sensory region.

The following protocol now produces serial memory. Some number of sets is trained by occasional activation of a reinforcement set. These sets are “stitched together” into a series by first activating set 1 in the early sensory region. The C-net converges on the target set. When set 1 stops firing in the sensory region, set 2 is activated in the early sensory region. When the network converges on set 2, the network is again trained, and recently active neurons (of set 1) in the motor region are trained to currently active neurons (of set 2) in the sensory region. A third set is then activated in the early sensory region, and so forth.

At some later time, when set 1 is activated in the early sensory region, the network converges on set 1. Activity of set 1 sensory region neurons inhibits activity of set 2 sensory region neurons. However, when set 1 sensory neurons stop firing, set 1 motor neurons induce firing of set 2 sensory neurons. As set 2 appears in the sensory regions, its untrained projections to set 1 motor region neurons silence set 1, and the C-net converges on set 2, and so forth through the remainder of the series. A single channel of the C-net may be trained on two different series with common elements, or on one series with repeated elements, because the second channel can serve as context for the first.

If a series is activated in one early sensory region, and the other channel does not have an active training set, the series will be induced in both channels. If different series are activated in different early sensory regions, each channel will follow the series initiated in its early sensory region.

As described so far, serial learning requires each element of a series to be activated promptly so that activity of the previous set is not more than 150 cycles in the past when the network is trained. However, this restriction may be removed by inducing an oscillation in which prolonged firing of motor neurons rekindles firing of sensory neurons before motor neurons become silent, and firing of sensory neurons rekindles motor neurons. The obvious limitation is that the period of non-firing must be short compared to the period of firing. This problem may be ameliorated by allowing neurons that have ceased burst-firing to fire slowly (as is usual in the real cortex). This allows the sets of a series to be introduced at arbitrary intervals provided no new sets are activated in any region of the network. New recall sets will still induce convergence on their target sets.

In the C-net of Subsection 3.7, asynchronous firing and hyperexcitability in recently burst-firing neurons also contribute to stability during intermittent firing. Asynchronous firing is produced by the rule: \( a_{e,x} = 1 \) only if \( \text{mod}(x+e)/p = 0 \), where \( p \) is the period of firing. This rule is biologically improbable, but any rule would do. By “hyperexcitable” I simply mean that recently burst-firing neurons will fire intermittently if \( I_{e,x} \) is greater than some value less than zero.

These rules have unwanted consequences when trying to start from a tabula rosa condition by intermittently reinforcing randomly evolved states. As soon as one set is trained, the network is
trapped in versions of that state, and additional states cannot be trained. Numerous, interesting, biologically possible solutions exist. However, in these simulations I simply require a neuron to stop firing after 1500 cycles without inhibition.

3.6 Serial Storage Capacity

Serial memory was evaluated by modifying the network of Subsection 3.4 so that sensory and motor region neurons burst-fired for 100 and 130 cycles, respectively, followed by 100 cycles of firing on every twentieth cycle. No procedure was implemented to produce asynchronous firing (i.e., uninhibited, previously burst-firing neurons fired if mod[x/p] = 0). Post-burst neurons were not hyperexcitable.

Some number of sets was trained by alternant activation of reward and punishment sets. A training set was activated in the early sensory regions and the entire C-net converged on the set over 100 cycles. Sometime during the next 1000 cycles, another training set was activated in the early sensory regions, and after convergence, a reinforcement set was activated. Again, at some time in the next 1000 cycles, a third set was activated, and so forth.

Recall of a series was considered successful if, after activation of the first set in the early sensory regions, the late motor elements of each set in the series were either recalled or suppressed in the proper order with no more than 10% errors on recall.

In a series of 10 trials, the network stored an average of 19 sets in series (range 15 – 21). Failure almost always resulted from a failure of the two channels to recall synchronously, resulting in simultaneous activation of both integrative regions and failure of either reinforcement set to become active.

3.7 A Conditioned Learning Example

An extended example of conditioned learning is illustrated in Figs. 8 through 13. In this example, 3 training sets are formed. For convenience, and to help clarify the analogy between the behavior of the network and some cognitive behaviors, I refer to the sets as birdfeeder, pounce, and Ms-Grandebico (the owner of the birdfeeder). I refer to the early sensory subsets as see-the-birdfeeder, see-the-bird, and see-Ms-Grandebico, and to the late motor subsets of the first two sets as run-to-the-birdfeeder and pounce-on-the-bird.

I do not mean to imply that this network can actually implement a rule such as “Run to the birdfeeder; think about a bird until the real bird appears; then pounce if and only if Ms Grandebico is not present.” The model does not attempt to compute a sequence of actions needed to run to the birdfeeder. None-the-less, the network does learn to behave in ways that seem analogous to some of the cognitive processes required for such a behavior.

In this particular example, the rate of readout of a series is increased by causing activation of a reinforcement set to suppress burst-firing in the sensory region. Neurons of all regions burst-fire for 100 cycles. This change requires a number of additional, biologically possible network properties. Firing of a reinforcement set also suppresses firing in the reward and punishment regions of any neuron for which $E_{x} < -7$. (A set that has been trained by both reward and punishment in different contexts
may activate neurons in both reward and punishment regions with the context-appropriate region converging first. Deep inhibition of the alternative region prevents occasional, subsequent emergence of the context-inappropriate reinforcement set. The suppression happens, incidentally, to be of the same length of time as intermittent firing. Firing of the reward or punishment sets also results in self-inhibition. In this case, each active neuron is simply given an inhibition of $I_{\text{min,e}}$, permitting the set to re-emerge rapidly enough to follow the emergence of a series of rewarded (or punished) sets in a channel.

The C-net may sometimes converge on the reward or punishment set before all elements are active in the late motor region. The inactive elements may then become active after the punishment set has self-inhibited. Accordingly, the effects of the reward and punishment sets are delayed until they have been active for 20 cycles. Finally, activity in the early sensory regions should be a somewhat evanescent, reflecting a changing environment. Accordingly, firing of a reinforcement set also suppresses active neurons in the early sensory region so that these regions may fire randomly. Intermittent firing in the integrative regions can cause rapid re-emergence of a reinforcement set, which may suppress a newly emerging set of a series while it is only active in the sensory regions. Accordingly, intermittent firing is not implemented in the integrative regions. Biologically possible alternatives to all these strategies exist.

The strength of untrained direct excitatory projections has been increased to 0.2. This change makes the sizes of sets more uniform from R-net to R-net; it does not affect the ability of the network to perform the task. The post-burst period of intermittent firing is extended to 150 cycles, requiring a decrease in the asynchronous firing period to 5 cycles and a post-burst excitability such that neurons fire if $I_{e,x} > -5$.

The C-net is initialized as described above except 7 neurons per R-net are activated. The behavior of one channel during the first 302 cycles is shown in Fig.8, and is representative of the behavior of the second channel. This initial condition does not appear to be a state the network would be likely to reach on its own. It leads to an oscillation that is damped in all except the late motor region by the end of 10 cycles. In the absence of small variations in $I_{\text{min,e}}$, such an oscillation may be uncontrolled.

On cycle 302 the reward set is activated, and the network is trained forming birdfeeder. The next 270 cycles are described in Fig. 9. Throughout this period the early sensory and punishment regions fire at random following suppression by the reward set. During the first 100 cycles of this period, the sensory region fires small subsets of the newly formed training set, burst-firing having been suppressed by the reward set. The motor and late motor regions converge on birdfeeder, the neurons that were firing on cycle 302 serving as a recall set for other neurons of birdfeeder that were active during the previous 25 cycles. Toward the end of the first 100 cycles, burst-firing ends in the motor and late motor regions.

Approximately 50 cycles later, at cycle 459, elements of the sensory region begin to resume burst-firing. 100 cycles after that, the motor and late motor regions resume burst-firing as the sensory region ceases burst-firing. After cycle 513, slow firing in the motor regions and burst-firing in the
sensory region is sufficient to re-excite the reward region, so that by cycle 567 the reward set has been firing for 20 cycles, and burst-firing in the reward region is, again, suppressed. Alternant firing of sensory and motor regions continues in this way until the network is 1500 cycles beyond the training event at cycle 302.

At 2000 cycles, a second training set, *pounce*, is created in the same way with comparable results (not shown).

At cycle 4000, *see-the-birdfeeder* is activated (Fig. 10). The network may be seen to converge on *birdfeeder*, sequentially, in the sensory, motor, and late motor regions. Note that the reward region does not converge on *birdfeeder* until after it is firing in both sensory and motor regions. At cycle 4060, the reward set has been active for 20 cycles and suppresses burst-firing in the sensory and reward regions.

At cycle 4110, *see-the-bird* is activated, and the network converges on *pounce* with the reward set firing. At this time, the reward set is also activated from outside the C-net causing a training event in which projections of recently firing *birdfeeder* motor neurons onto currently firing *pounce* sensory neurons are trained.

At cycle 6000, *run-to-the-bird-feeder* is again activated in the early sensory region (Fig. 11). The network converges on *birdfeeder*, and on cycle 6059, activation of the reward set suppresses burst firing in the sensory region. Continued firing of *birdfeeder* in the motor region induces firing of *pounce* in the sensory region.

By cycle 6219 the network has converged on *pounce* and the reward set has fired. This time, however, the punishment reinforcement set is activated from outside the C-net, and training, for the first time, includes the punishment region. Note, that on the previous occasion, when the network was rewarded, *see-the-bird* was active. On this occasion, *pounce* has been recalled by *birdfeeder* rather than the “perception” in the early sensory region of *see-the-bird*. This punishment event adds three neurons to the sensory subset of *pounce*. These neurons are not trained to *see-the-bird* neurons, which were not active, or the 20 reward region *pounce* neurons, but are trained to a 26-neuron set in the punishment region.

On cycle 8000, *see-the-birdfeeder* is again activated, and by cycle 8058 the network converges on *birdfeeder* (Fig. 12). By cycle 8170 the network converges on *pounce*. During this time, *pounce* neurons in the punishment region are activated before those in the reward region, and *pounce-on-the-bird* is suppressed.

On cycle 8270, shortly before the anticipated resumption of burst-firing of *pounce*, *see-the-bird* is activated. Now, as burst-firing resumes, the reward set becomes active and *pounce-on-the-bird* remains active.

It is commonly necessary to punish the network more than once to get it to reliably suppress *pounce-on-the-bird* when *see-the-bird* is not active. Such training depends on the addition of small numbers of spuriously firing neurons to relevant training sets.

On cycle 12,000, a third set, *Ms Grandebico*, is trained by activating the reward set. On
cycle 14,000, *see-the-bird* and *see-Ms Grandebico* are activated in left and right channels, respectively. Each channel converges on its respective set, and the network is trained by activating the punishment set (not shown).

On cycle 16,000, *see-the-bird* is activated (Fig. 13). After activation of *run-to-the-birdfeeder*, *pounce* is recalled. As *see-the-bird* is not active, the punishment set becomes active, and *pounce-on-the-bird* is suppressed. On cycle 16,169, *see-the-bird* is activated in the left channel only, and on cycle 16,217 the network converges on the reward set. However, *pounce-on-the-bird* is still suppressed by the previous activation of the punishment set. Activity of *see-the-bird* is maintained, and the C-net reconverges on cycle 16,410 with the reward set active, and *pounce-on-the-bird* firing.

At 18,000 cycles, *see-the-bird* is activated in the left channel while *see-Ms Grandebico* is activated in the right channel (Fig. 13). The left channel promptly converges on *pounce* while the punishment set formed during the previous joint appearance of *Ms Grandebico* and the bird is activated suppressing *pounce-on-the-bird*.

Earlier in the simulation, at cycle 16,169, *see-the-bird* was activated on the cycle after convergence on *pounce* and the punishment set. The network subsequently converged on *pounce-on-the-bird* and the reward set. Further simulations demonstrate that, if activation of *see-the-bird* is delayed by 25 cycles, the previously trained links between *pounce* in the left channel and *Ms Grandebico* in the right will recall *Ms Grandebico* in the right channel and cause activation of the punishment set even though *see-the-bird* is active in the left channel. In other words, the memory of *Ms Grandebico* is sufficient to stop *pounce-on-the-bird*. However, when *Ms Grandebico* stops burst-firing, activity of pounce in the left channel will activate *pounce* in the right channel, and if *see-the-bird* is still active, the reward set will become active unless *Ms Grandebico* is active in the early sensory region (not shown). In other words, the memory of *Ms Grandebico* may make the network hesitate, but only the presence of *Ms Grandebico* can stop the pounce. If *see-the-bird* is active in both channels, the network will pounce even in the presence of the memory of *Ms Grandebico*.

Each of 3 protocols was repeated 10 times. Each protocol started with *see-the-birdfeeder* through convergence on *pounce* and the punishment set. This was followed by prompt activation of 1) *see-the-bird* in both channels, 2) *see-the-bird* in one channel, or 3) both *see-the-bird* and *see-Ms-Grandebico*. Each replication had the outcome described above except for one trial in which protocol 2 led to convergence on the punishment set.

### 3.8 Parameter space

The space of parameters that permit a particular behavior is too complex to be readily described, and depends on the behavior being modeled. However, the behavioral example given in this subsection and other behaviors described in Section 4, but undocumented, have been simulated using both the C-net of Subsection 3.6 and alternatives to the C-net of Subsection 3.7 in which (1) there is no LTP of excitatory projections onto excitatory neurons, (2) the number projections per excitatory neuron onto inhibitory neurons of R-nets of different channels ranges from 1 to 3, or (3) the number projections per excitatory neuron onto inhibitory neurons between integrative and channel R-
nets ranges from 1 to 3. Most of the behavioral examples have been demonstrated in C-nets with up to 4 channels (but reduced R-net size and no late motor regions in order to accommodate serial computer memory constraints). The additional channels were added in a modular fashion with no change in architectural or synaptic parameters. Increasing the connectivity between the integrative and channel regions increases the rate of convergence in the integrative regions and requires an increased delay in activating the effect of reinforcement regions. No other adjustments were needed.

4. DISCUSSION

4.1 Memory and Conditioned Responses

Extremely sparse, powerful, local, inhibitory projections combined with many, weak, distant, excitatory projections are a common theme in the brain. It may be that these models reveal something of the function of this theme. The models appear to support mammalian size memory capacity, are not subject to severe evolutionary constraints, and permit the modular construction of behavioral solutions to significant problems.

The example of Subsection 3.7 contains elements of classical and operant conditioning in a network that distinguishes between perceived and recalled stimuli. Activation of a motor output such as *pounce-on-the-bird* requires both the presence of one stimulus and the absence of another.

Other interesting behaviors are easily elicited. Secondary reward training appears when a set in one channel is associated with other sets in other channels that have been rewarded.

Behavioral examples analogous to either predictive fabrication (planning) or refabrication of memory depend on a common principle. If set *y* is trained to set *x* during event 1, and set *z* is trained to set *x* during event 2, then if sets *y* and *z* are active during event 3, set *x* will be activated in preference to any other set to which *y* or *z*, but not both, are trained. Sets *y* and *z* may appear at the same time in separate channels to induce *x* in a third channel, or set *x* may have followed set *y* in one channel and set *z* in another. This principle will fill in possible future events, or memories that are only partially encoded, in ways that are likely to be adaptive.

Once the idea of disinhibition with massive, weak, distant excitatory projections is put to use, the number of biologically possible options for constructing useful networks becomes large. All of the mechanisms introduced for Subsection 3.7 have alternatives, and none of them is needed in the network of Subsection 3.6. Even without provision for serial memory, the C-net of Subsection 3.4 can perform operant conditioning and secondary reinforcement tasks.

The chief limitation of these models seems to lie in the need for reinforcement as each set is added to a series. It is not difficult to imagine that coincidence firing induces a state in a link that may lead to training if reward or punishment occurs in the near future. The start of such a series might, for example, be defined by the onset of novelty (marked by the absence of convergence in an integrative region).

Alternatively, I have made preliminary progress on development of an algorithm that (1) trains
synapses on every cycle, (2) requires many cycles of correlated firing for the training to become more or less permanent, (3) reverses training when firing is uncorrelated. The algorithm creates sets when activity is repeatedly correlated, and it may be expected to permit training of series without the necessity of first training single sets.

More elaborate versions of this network may reasonably be expected to produce more complex psychological results. Integrative rather than early sensory regions may initiate recall, and integrative regions may be supervised by still other R-nets.

4.2 The Mystical Brain

The first and most obvious problem in discussing such matters as memory and consciousness lies in having some notion of what it is one intends to discuss. Characterizations of consciousness commonly involve notions of self-awareness or imagination. Crick (1979) writes, “There is more than a suspicion that such phenomena result from the computation pathways acting in some way on themselves…” Allusions to imagination are less frequent, but Gregory (1977) seems to refer to something like imagination when he remarks, “Perceptual reality - consciousness - may be a feature of the simulated world or the brain on which we act predictively.”

Consider the following tentative definition: “Consciousness” is the ability of an information processing system to (1) use information about how the system is processing information adaptively (self-awareness), and (2) provide parts of the system that are capable of processing information about events outside the system with information about events that might occur (imagination).

Compare this definition with the behavior of the network of Subsection 3.7. Sets do not become active in the integrative regions until they have become active in both motor and sensory regions. In other words, these regions harbor representations of how information is being processed. This information may be acted upon adaptively in a process that involves recall in both the sensory and reinforcement regions of training sets that represent events that might occur.

There are, of course, objections to the view that this C-net harbors a nanospark of consciousness. None-the-less, granting every such objection, if there are correlates of mind in the activities of the brain, then finding behaviors of the network that are like correlates of memory and consciousness still argues for the point of view that minds are not mystical objects, and some magnitude of consciousness may arise as an epiphenomenon in small brains organized to perform simple conditioning tasks.

REFERENCES


hippocampal CA1 cells. J. physiol. 496, 197-209.


LEGENDS FOR FIGURES

Fig. 1. A representation of an R-net. Smooth neurons are excitatory. Spiked neurons are inhibitory. The ratios of excitatory to inhibitory neurons and synapses shown are not the same as those in the model.
Fig. 2. The number of training sets stored as a function of the number of R-net excitatory neurons. The number of inhibitory neurons is 20% of the number of excitatory neurons, and the numbers of synapses are adjusted so that the probability of a pair of excitatory neurons being linked is 0.4 for all R-nets.

Fig. 3. The number of training sets stored as a function of the total number of R-net synapses.

Fig. 4. An expansion of the origin of Fig. 3 meant to make clear the rapid fall off of storage for 20- and 30-neuron training sets compared to 40-neuron training sets.

Fig. 5. Information per synapse as a function of R-net size.

Fig. 6. (a) Each box represents a local R-net. No inhibitory projections extend outside a box. (b) A C-net. Arrows indicate the directions of excitatory projections between R-nets.

Fig. 7. Percent errors (spurious and missing neurons compared to target set size) as a function of the number of sets stored in the C-net discussed in Subsection 3.4.

Fig. 8. Pseudo-random firing of the C-net described in Subsection 3.7 for 302 cycles following initialization. Activity in the right channel is not shown, but is similar to that shown for the left channel.

Fig. 9. The C-net of Subsection 3.7 shown for 270 cycles following the training of *birdfeeder*. In this and the remaining figures, diamonds denote numbers of active neurons that are members of no set. Stars, hexagons, and spirals generally denote numbers of active neurons that are members of *birdfeeder*, *pounce*, and *Ms Grandebico*, respectively. In the Reinforcement region (last frame) stars and hexagons denote numbers of active reward neurons and punishment neurons, respectively. Numbers in parentheses denote the number of neurons in a training set.

Fig. 10. The “stitching together” of *birdfeeder* and *pounce* into a series. Meanings of symbols are as in Fig. 9.

Fig. 11. “Punishment” of *pounce* when *see-the-bird* is not active. At the end of the sequence, the network has recalled the reward set. However, the punishment set is activated from the outside the C-net.

Fig. 12. Recall of *birdfeeder* followed by *pounce*, which is first recalled with the punishment set (when *see-the-bird* is not present) and then with the reward set (when *see-the-bird* is present).

Fig. 13. Behavior of the trained C-net. Both channels are shown with integrative and reinforcement regions repeated beneath each channel for ease of comparison. In the Punishment region, stars denote the number of active neurons in the set formed at the end of Fig. 11 when *pounce* without *see-the-bird* was punished.
Fig. 3

Fig. 4
Fig. 5

Fig. 6
Fig. 7

Fig. 8
Fig. 9
Fig. 11
Fig. 12
Cognition Model 32

Fig. 13