BEHAVIORAL AND BRAIN SCIENCES (1995) **18**, 617–657
Printed in the United States of America

The Hebbian paradigm reintegrated: Local reverberations as internal representations

Daniel J. Amit
Racah Institute of Physics, Hebrew University, Jerusalem; Instituto di Fisica, Universita di Roma, Rome, Italy.
Electronic mail: *ilios.fiz.huji.ac.il*

Abstract: The neurophysiological evidence from the Miyashita group's experiments on monkeys as well as cognitive experience common to us all suggests that local neuronal spike rate distributions might persist in the absence of their eliciting stimulus. In Hebb's cell-assembly theory, learning dynamics stabilize such self-maintaining reverberations. Quasi-quantitive modeling of the experimental data on internal representations in association-cortex modules identifies the reverberations (delay spike activity) as the internal code (representation). This leads to cognitive and neurophysiological predictions, many following directly from the language used to describe the activity in the experimental delay period, others from the details of how the model captures the properties of the internal representations.

Keywords: active memory; associative cortex; attractor dynamics; content sensitivity; internal representations; learning; modeling

> então concluiremos que uma palavra, quando dita, dura mais que o som e os sons que a formaram, fica por aí, invisivel e inaudivel para poder guardar o seu próprio segredo . . .
> José Saramago, *A Jangada de Pedra* 1988
>
> We may conclude that a word, once said, persists longer than the vibration or the sound that formed it, it stays there, invisible and inaudible so as to preserve its very own secret . . .
> *The Stone Raft*

1. Introspective attractors

Neural network models of brain function can be roughly classified as feed-forward and feed-back or attractor networks. Very clear and informative evidence of the presence of attractors in temporal cortex of primates is presented in the recent studies of Miyashita and colleagues (Miyashita 1988; Miyashita & Chang 1988; Sakai & Miyashita 1991; see also Amit 1993). Here I will suggest that most of the results observed in these seminal experiments could be predicted on the basis of simple observations concerning common cognitive phenomena, without recourse to any specific model. One can then argue that such considerations probably underlay Hebb's (1949) postulation of *synaptic dynamics* as a means of stabilizing *reverberations* in *neural assemblies*. Such a reverbatory mechanism provides a potentially useful way of dissecting the complexity of the connection between sensory input and motor reaction.

Consider the familiar every-day situation in which one is given the task of translating a word from one language to another. The word is known in both languages, that is, both words are in memory. Suppose also that the task is communicated verbally (spoken). Both the task and the word in the first language are well understood. The acoustic stimulus disappears as soon as it has been presented. The response, the word in the second language, (1) might be produced very rapidly and correctly, or (equally commonly) (2) the first word might be recognized, together with a strong sense that the translated word is known, but the retrieval of the corresponding word might not be possible. One knows that one knows, yet one does not know. One can persist with the effort to retrieve the translated word for quite a while despite the absence of the eliciting stimulus. And not uncommon is a sequel in which the entire episode (word + task) fades away from our consciousness, only to surface resolved hours or days later.

During this long search period, conscious or unconscious, the originally presented word must have been available in a sense that goes beyond the fact that it was contained in our memory, that is, we knew it. After all, there are very many words we know in the language of the word presented for translation. What is special about this particular word, of course, is that it has been tagged by the stimulus – the sound of the spoken word.

This naive resolution of the problem posed by the familiar cognitive situation has quite significant implications. Somewhere in the skull, between the locus of the fully preprocessed stimulus and before the beginning of the generation of a response, there must be loci storing, *passively*, many memories. Those are the things we know and remember. They are, most likely, stored in the synaptic structure of each locus. Each of these loci must be able to *maintain* one memory out of the passive stock (the one tagged by the stimulus) in a special status, for relatively long times, and in a status which will make it available for future attempts to perform the task.

 0140-525X/95 $9.00+.10

This sort of consideration also leads to the conclusion that the number of such loci must be relatively high. One seems to be able to maintain several items *concurrently* in the tagged status. Those may be related to several items involved in a given task to items involved in different tasks which interweave during long time intervals of inattentive processing and also possibly to the active presence of the tasks themselves. Consequently, the cortex must be modular. Both anatomy and physiology provide hints as to the geometry (in terms of cortical volume, localization, and number of neurons).

2. The concept of attractors – active and passive memory

The attractor picture of a cortical module is, briefly, the following: The module consists of a relatively large ensemble of neurons (of the order of 10^5 cells) in which the probability that any two neurons are connected by a synapse is high (a few percent suffice). Such an ensemble of cells is geometrically localized in about $1mm^2$ of cortical surface. Experience (i.e., training) structures the set of synaptic connections in a Hebbian way. The resulting synaptic structure is such that when a stimulus activates a subset of the neurons in the module, raising their spike rates and then the stimulus is turned off, the activity of the neurons in the ensemble may, depending on the stimulus, either (1) decay rapidly back to spontaneous levels (stimulus ignored) or, for other classes of stimuli, (2) maintain a stimulus-selective subset of the neurons at elevated rates in the absence of the stimulus for long durations. In other words, the synaptic structure provides sufficient structured feedback so that the afferents (inputs) due to the distribution of activity among neurons with elevated spike rates keeps the same set of neurons active, leaving the others at spontaneous activity levels. The collective nature of this dynamical state of affairs (i.e., the fact that the elevated activity of each neuron is maintained by many others makes the attractors impressively immune to disorder in the synaptic structure as well as to dynamical noise, both of which are unavoidable in cortical conditions.

What determines which configurations of neurons in the assembly can collectively maintain each other in the elevated activity state is the synaptic matrix. This is passive memory. What determines which of the possible, self-maintaining configurations actually reverberates in the module is the stimulus. The short appearance of the stimulus "tags" one of the passive memories by activating the particular attractor associated with this stimulus. The activated configuration in the assembly is an attractor in the sense that each of these configurations is activated by a wide class of stimuli that are in some sense close to each other. The active configuration is then the representative of the class. This is what often goes under the name of *content addressable memory* or, less descriptively, *associative memory,* in contrast to the physical address referencing of memory used in digital computers.

The attractor type of memory activation contrasts with the computer in yet another sense. In the computer, when a piece of information is to be acted on, it is taken from its address and put in a special section of the processor for action. This is analogous to the activation of a passive memory. In the neural module, however, not only is the addressing done by content and not by physical address, but the activation leaves the item in the module. In some sense the module is both the memory and the register. Thus, whereas in the computer the activation of a memory item is signaled by the special location (the register) in the cortical module, we argue, the activated memory is distinguished by the activity. The place remains invariant. Moreover, whereas the computer register can hold any information configuration for processing, the attractor module will hold only the representatives of classes that have been trained into the synaptic structure.

I am not emphasizing these distinctions to imply a preference. It is simply that in a system like the cortex the register option is not available because the register must also be made of neurons, hence maintaining an item for later processing can depend only on synapses; so we are back to square one.

It is important to make a clear distinction at this point between the tagged persistent memories and concepts such as short, intermediate, and long term memory. The latter usually refer to the stock of passive memories, that is, to the duration of the synaptic programming or to the duration of its accessibility. They relate to the ability of the system to manipulate incoming tasks. The tagged persistent item introduces an additional "temporal" category: the persistent activity distribution excited by a stimulus. A memory of this type (sometimes referred to as *working memory,* see O'Keefe & Speakman 1987; Tanaka 1992) can belong to any of the three temporal categories. This basic distinction is sometimes overlooked, even by Hebb himself (see, for example, the second quote from Hebb & Donderi, 1987 and the discussion in sect. 4).

The simplest conceivable carrier of such a tagging signal is the persistent distribution of elevated spike rates (Hebbian reverberations) among the neurons in the module (Hebb's cell assembly). One may contemplate other stimulus-selective taggings of stored memories, but those would be much more difficult to observe. Since persistent spike distributions in the absence of the eliciting stimulus are governed mainly by the synaptic structure in the local assembly, the tokens maintained there during the prolonged performance of the task will be prototypic. It is likely that the structure of such reverberations will not depend on details of the stimulus, such as the tone, the pitch, or the modulation of the acoustic signal communicating the original word. In theoretical models the dynamics of multineuron systems, when maintaining activity distributions in the absence of the stimulus, gives rise to a global dependence on the stimulus: large classes of stimuli will elicit the same persistent spike distribution for all stimuli in a class. Stimuli which are different enough induce different persistent activities. In this sense the activity distribution in each reverberation can be considered as *representation* of the class of stimuli that elicit it.

It may be useful to clarify the role that is ascribed in the present context to the word internal "representation," given that it is at the center of so much debate in the cognitive science community. The computational situation described above, in which the performance of the task on the stimulus (the word to be translated) is to take place long after the stimulus has disappeared, seems to leave little choice. When the task is ultimately carried out, it must have an operand. That token, which survives somewhere in the cortex, is a representation of the set of equivalent stimuli. Such a token seems to be logically required and a candidate

for it is experimentally observed, as will be described in section 5.

The independence of the reverberation in the local assembly from the details of the stimulus does not imply that the internal representation does not depend on the task. Since the stimulus contains the word as well as the task, the persistent token may, in principle, depend on both. Yet the fact that the same word can be involved in many different tasks – rhyming, antonyms, synonyms and so on, suggests that the task may be represented likewise (or only) elsewhere. The linking of two separate representations is still an open problem, to be investigated both experimentally and theoretically.

3. The cortical processing cut

The presence of such reverberating stations in the cortex subdivides the feed-forward picture of cognitive performance into three categories. The partitioning takes place inside the cortex with the boundary lines lying at different distances from different sensory mechanisms. Roughly speaking, the three-way division is: (1) The *formation* (learning) and *function* of the organic system leading from the external world to the persistence stations: that is, the preprocessing of the input required to form different internal representations (the structure of the persistent taggings) for significantly different stimuli; (2) The *formation* and the *activation* of the *internal representations* (persistent taggings, Hebbian reverberations) by the preprocessed afferent stimuli, and the interaction between activated reverberations in different loci (modules); (3) The *organization* and the *functioning* of the *decoding* of the cortical reverberations into computationally driven reactions.

This 3-way partition of the cognitive computational system is advantageous for experimental as well as for theoretical study. The above analysis implies that the persistent tagging *represents* (at least for a while) some abstract feature of the stimulus involved in the task. Hence the neurophysiologist can search for these local modules (Hebbian assemblies) in the cortex of the performing mammal in the course of the performance of a behavioral task. The tentative acceptance of spike rate distributions allows an easy read-out for the neurophysiologist of the relevant representations. He can derive this from single unit recordings to be analyzed off-line. He can count on the fact that the *internal representations* he observes will not depend on particular details in the manifestation of the stimulus, provided the stimulus has been correctly classified (interpreted) by the animal. The latter can be monitored by the animal's response during the experiment.

From the point of view of cognitive science, one representation may be as good as any other. The opportunity provided by the 3-way Hebbian partition is that it allows a direct, empirical expression for the representations to replace a metaphorical one. As will be argued in section 5, the quantitative properties of the representations discussed here can be measured. A few tens of milliseconds following the disappearance of the stimulus, the contents of these attractors are the sole basis for the completion of the mental computation. Given a direct and measurable candidate for the representations, the hypotheses of cognitive science can be tested against a well-defined body of data on the neurophysiological level. The realistic neural substrate is required to inform the hypothesis about the potential as well as the limitations of the system.

The above properties are not common to all models of internal representations. For example, in a feed-forward description of computation, a particular pattern of neural activation persists only for as long as the stimulus is on. A given neural activity distribution, elicited during the presentation of the stimulus, cannot be supposed to be available for computation at a later time. Moreover, the activity distribution may be sensitive to the particular details of the stimulus. Thus the activity distribution may be richer than what is actually used for continuing the computational process. It may therefore not provide sufficient constraints on the computation. The attractor, in contrast, contains measurable information for long periods and that information is the same for all stimuli which are classified by the same attractor.

Moreover, the attractor dynamics distinguishes naturally between the course of an unfamiliar stimulus and a familiar one, the former being significantly different from the stimuli one has learned to classify. Attractor dynamics leaves all neurons in the module at very low activity levels, despite the fact that during the presentation of the stimulus as many neurons may be excited as with a familiar stimulus. This distinction is not naturally available for alternative paradigms of representation.

Such representations and their task-dependent patterns can provide invaluable information about how computation is organized in the cortex. Before proceeding with the elaboration of the experimental and theoretical account, I return to the subject of this essay's title, Donald Olding Hebb.

4. Multicomponent Hebbian paradigm

Allusions to Hebb abound in the preceding text. They have not been explicitly formalized because I have been trying to emphasize the intuitive appeal and the almost imperative nature of the local internal representations. Yet I believe that whatever is valid in this picture must have been clearly perceived by Hebb many years ago. Someone joining the field in the last decade finds innumerable references to Hebb's work, but the general tenor of these references is of a synaptic engineering type. Almost any type of synaptic learning in neural networks genuflects in Hebb's direction. Yet Hebb was not a neurochemist or a neurophysiologist. He was a psychologist searching for a neurally based substrate for psychology to supplement or replace the mythological one.

The Hebbian paradigm is multidimensional. It is composed of a prescription for synaptic modifications: synapses are modified by afferent stimuli in a way that tends to stabilize the pattern of activity generating the synaptic modifications. The stable neural activity distributions are excited in the local *assembly* by each of the learned stimuli. This is an *unsupervised learning* mode which aims at producing synaptic structures which can sustain a selected set of activity distributions in the *local assembly*, to use Hebb's language. The role of the resulting synaptic structure is to sustain the local activity produced by a stimulus in the absence of the eliciting stimulus. To maintain the activity evoked by the stimulus in the presence of the stimulus does not require any synaptic modification.

Hebb's paradigm is not about the activity generated in one assembly by the activity present in another. It is not a

feedforward picture, for better or for worse. It can be summarized as a process generating the feedback connectivity required for maintaining reverberations (persistent spike distributions) in a local network by the activity in the same network. The citations below make this point quite clearly.

> Let us assume that the persistence or repetition of a reverbatory 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. (Hebb 1949)

Elsewhere Hebb says "It seems that *short-term* memory may be a *reverberation* in the closed loops of the cell assembly and between cell assemblies, whereas *long term* memory is more structural, a *lasting* change of synaptic connections (Hebb & Donderi 1987, p. 110, emphases in the original)."

This is not intended as a hagiography of Donald Hebb, nor proof by dogma, nor a decomposition of texts. Rather, it is an attempt to salvage a profound idea from excessive fragmentation that has obscured the potential three-way partition of the cognitive system. In particular, there has been an underestimation of the phenomena necessary for mental processing, phenomena which can be observed relatively easily neurophysiologically and which can provide important information for deciphering some basic ideas pertinent to cortical computation.

Note that in the second quotation the distinction between long term memory and active memory is implied, yet the terminology is not adopted. Clearly, the reverberation is an active state of the assembly, whereas the structure of the synaptic organization is not. Moreover, the reverberation must be sustained by the underlying synaptic structure; hence it is a particular expression of the properties of the synaptic structure. A given synaptic structure may persist for short, intermediate, or long periods.

To conclude this speculative section, it may be of value to point out an additional bonus promised by this outlook. It is known anatomically, physiologically, and neurologically that as one proceeds along the elaboration path in the cortex, one always finds backward projections, as far back as into the primary sensory areas. On the other hand, it is a very familiar experience to have a given sensory capacity notably improved when the content of the observed stimulus is known. For example, when vision is impeded by distance or haze so that a given object cannot be discerned (or read), receiving a cue as to the nature of the object (or the written text) often produces a clear perception of the target. In other words, suppose that the sensory information concerning an object is not sufficient to produce recognition. Suppose further that other information about the object is provided (vocally, for example) sometime prior to the observation of the object. The information in the vocal signal has been recognized and hence has excited a reverberation in some module. The back-projections from this module to the more primary areas in which the visual signal is being processed may provide enough additional information to make recognition possible. The special role of the reverberation is to make the contingent information available long after the signal producing it has disappeared.

Hebb's idea about reverberation in cortical assemblies seems to have been motivated by observations of Lorente de Nó (1949), concerning neural diagrams of Golgi-stained cortical slices made by Ramon y Cajal. The neural diagrams discussed by Lorente de Nó contain a small number of neurons, because the staining method makes only a small fraction of the neurons in the slice visible. Hence the feedback circuits observed seem relatively simple and suggested simple flows. Hebb was aware of the fact that the circuits were too simple and would probably not be able to sustain a reverberation for a sufficiently long time, and that to ensure long living reverberations a much larger number of neurons would be required (see Hebb & Donderi 1987).

5. Experimental evidence for Hebbian reverberations and beyond

The fact that local reverberations are almost a logical necessity does not eliminate the need to test their existence empirically. This has been done in the last few years in a particularly convincing way by Miyashita (Miyashita 1988; Miyashita & Chang 1988; Sakai & Miyashita 1991) in a culmination of a program started some twenty years ago by Fuster (1973) and Niki (1974). In these experiments monkeys are trained to perform delayed image matching (delayed match to sample, DMS) of visual images. The images are assumed to be meaningless for the monkey and mutually uncorrelated in their geometrical structure. For this purpose they are generated by a computer using a graphic procedure with several stochastic components. Typical images are shown in Figure 1. Following training, testing proceeds according to the following protocol: (1) an image appears on the screen for a short period (200 msec); (2) The screen remains blank for a prolonged period (as long as 16 seconds); (3) a second image appears briefly; (4) the monkey should react selectively depending on whether the first and second image were the same or different.

Training is accomplished, prior to the insertion of recording electrodes, by presenting the monkeys with long sequences of pairs of images as described above, and rewarding them for correct responses. In different experiments the sequence of first images is varied. For example, the first images can be drawn at random from a store of generated images; a fixed sequence can be shown in a fixed order; or the set of images can be divided into pairs, the members of each pair following each other in fixed order in each trial during training, with the pairs selected at random. The second stimulus is always randomly selected with about 50% chance of being the same as the first. Note that the existence and the structure of the attractor depends only on the first stimulus. The second stimulus, the one used for matching, serves to maintain attention to the task. The

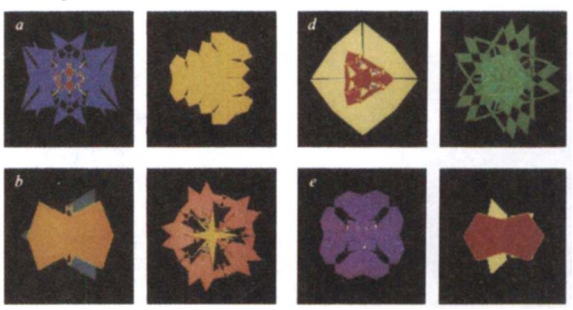

Figure 1. Several visual images used in the experiment. Reprinted with permission from *Nature* (Y. Miyashita, "Neuronal correlate of visual associative long-term memory in the primate temporal cortex." *Nature* 335:817–820, 1988). Copyright 1988 Macmillan Magazines Limited.

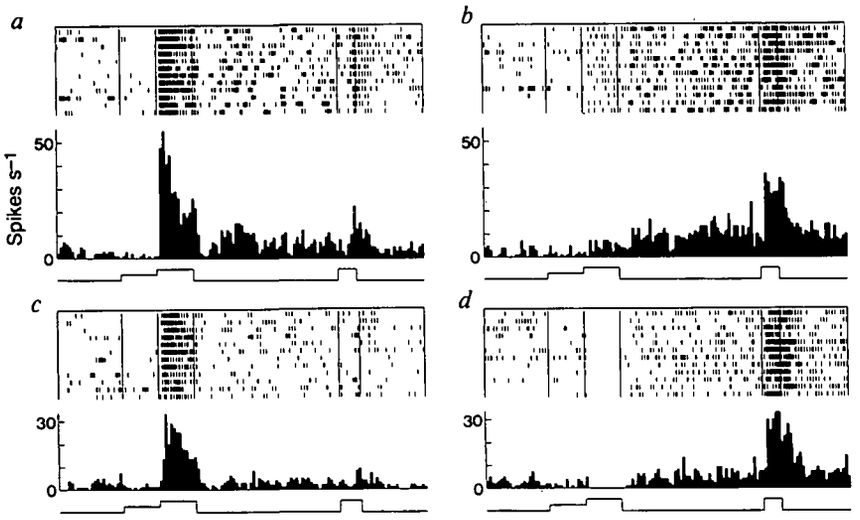

Figure 2. Reverberation dynamics. Four types of neuronal behavior observed in single units (from Sakai & Miyashita 1991, Fig. 3): At the top of each window, 12 spike rasters demonstrate the reproducibility of the delay activity on the single unit level, despite intervening presentations of other stimuli. Each row of dots is the representation of spike times recorded from a given neuron in a single trial, with the same image for the initial stimulus. The similar density of spikes in the delay period (the wide central interval) in all 12 rasters, despite the fact that other images intervened as initial stimulus between them, is the reproducibility underpinning the representation concept. Bottom: spike rate histograms. (a) Neuron active in presence of stimulus and persisting in its absence; (b) Same neuron unaffected by stimulus, active in delay period (a different stimulus leading to the same attractor); (c) Same neuron, third stimulus, neuron active in presence of stimulus, weakly active in reverberation, a different attractor; (d) Same neuron inhibited in presence of a fourth stimulus and weakly active in delay period. Under each window is the time course of the trial's protocol: prestimulus; warning; initial stimulus; delay; second stimulus. Reprinted with permission from *Nature* (Kuniyoshi Sakai & Yasushi Miyashita, "Neural organization for the long-term memory of paired associates," *Nature* 354:152–155, 1991). Copyright 1991 Macmillan Magazines Limited.

ordering of initial stimuli applies exclusively to the training phase. During testing, the first stimulus in each trial is drawn at random.

By impressive use of circumstantial evidence Miyashita and colleagues succeed in identifying a small part (about 1mm^2) of anterior ventral temporal (AVT) cortex where persistent stimulus-selective activity is manifested during the delay period, that is, in the interval between the presentation of the first and second images, when the stimulus is absent. The fact that the selective activity distribution can persist for as long as 16 seconds in a rather noisy environment is convincing evidence for the local maintenance of a reverberation by the feedback in the synaptic structure.

The main findings of these experiments for the theorist are:

1. Self-sustained, long-lasting stimulus-selective spike-rate distributions (reverberations) exist in the cortex.

2. The locus of the reverberations is locally modular, that is, within a given cortical region they appear to concentrate in one restricted portion. The module can accordingly be considered a spatially cohesive group of neurons, in contrast to a sparsely distributed collection. This fact accords well with the anatomical observations (e.g., Braitenberg & Schutz 1991), which indicate that the probability of synaptic connectivity between neurons in cortex falls off with separation. But within a range of 1mm that probability is still of the order of several percent. Thus, despite the fact that single synaptic contacts in cortex are typically weak (on the order of 100 presynaptic inputs are required to elicit a spike [McNaughton et al. 1981; Sayer et al. 1989]), this level of connectivity allows for the maintenance of robust attractors.[1]

3. The reverberations are reproducible on the single unit level, that is, the rates recorded on the same electrode from the same cell depend only on the stimulus leading to the reverberation. During testing, the first stimulus of each trial is drawn in random order, hence, between two presentations of the same image as first stimulus many other images intervene as first stimuli. Yet the delay activity is independent of this history. The reverberation can be considered as internal representations of the stimulus.

4. The reverberations (representations) are *not* single neuron properties. Single neurons cannot maintain selective activity rates. They are most likely collective emergent properties of the local module (the assembly). These experiments demonstrate explicitly that neurons driven by the stimulus may not be active during the reverberation and vice versa: neurons that are unaffected during the presentation of the stimulus may be driven during the delay period. See Figure 2.

5. The internal representations are distributed: many neurons participate in the representation of each memory and different representations share neurons.

6. The local internal representations are attractors (as was mentioned in item 4 above); a whole class of similar stimuli leads to the same reverberation.

7. The representations in this particular module of cortex are prototypes, that is, they are blind to color, size, angular orientation. This seems rather typical for this region of cortex (Tanaka 1992).

8. In their spike rate distributions the representations code, among other things, for temporal correlations in the training phase (e.g., see Fig. 3). In other words, what is represented depends not only on what is learned but also on how it is learned. This is an embryo of context sensitivity on the neurophysiological level. To be more specific, each of the correlated attractors contains information (expressed in

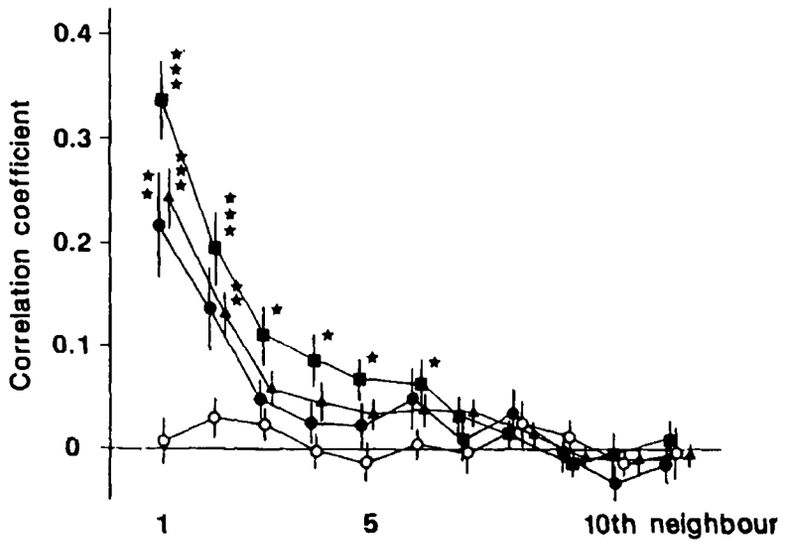

Figure 3. Correlated reverberations. Correlation coefficients (Kendall rank coefficients) of spike activities in a neural population in the delay period as a function of the positional separation of the stimuli exciting the reverberations in the training sequence. (From Miyashita 1988, Fig. 3c) Full circles represent correlations of delay activity distribution for learned images (used in training). Empty circles refer to activity distributions elicited by "new" images (not used in training). The different curves represent different samplings of neurons in the module selected for the computation of the correlations. The asterisks are irrelevant to the present discussion. Reprinted with permission from *Nature* (Y. Miyashita, "Neuronal correlate of visual associative long-term memory in the primate temporal cortex," *Nature* 335:817–820, 1988). Copyright 1988 Macmillan Magazines Limited.

neural activities) about the image activating it as well as about the images activating the attractors correlated with it. Hence, each attractor "knows" about its neighbors in the temporal training sequence (see, e.g., Amit et al. 1993).

9. In a different learning protocol, in which the sequence of first stimuli is presented as a set of permanently ordered pairs, the internal representation appear to become identical for both members of each pair. So does the behavior of the monkey in the behavioral paradigm. This does not imply that the monkey cannot detect the difference between the two pictures, but only that in the module used for generating the response for this specific task the two representations merge.

The foregoing empirical description requires several notes of clarification. On the one hand, we have emphasized the distributive and collective nature of the local internal representations and on the other, the reproducibility of recordings of single units. This apparent contradiction disappears if one keeps in mind that the enhanced, persistent activity of any individual neuron can only be sustained with the support of its fellow neurons participating in the same reverberation, via the local synaptic feedback. The activity of the single neuron therefore carries information about the stimulus only if it is accidentally selective between the different distributed reverberations. Among the active neurons there may be several that have the same activity in different representations. Also, the fact that the reverberations are dominated by the synaptic structure, and not by fine details of the stimulus, implies that every time the same stimulus is presented, the same reverberation is aroused. Consequently, the internal representations, for whatever they represent, may be perused and catalogued, recording one neuron at a time, as actually done in the experiments of Miyashita and colleagues.

This method precludes observing phenomena related to correlations in spike emission times if those manifest themselves in different neurons. To observe those, multiunit probes are required. Yet the representational, computational, and cognitive information contained in rate distribution is far from exhausted. On the other hand, the relative ease of accessing this type of information as well as analyzing and modeling it, compared to multielectrode data, makes it a very attractive subject of investigation.

6. Empirical and cognitive implications

A significant aspect of the generation of random images in the Miyashita experiments is that it is quite likely that the correlations between internal representations of stimuli are due only to their context dependence, that is, their frequent contiguous appearance in the training phase. The effort invested in the generation of the images is recompensed by the fact that in some sense the registered correlations between the attractors are the minimal correlations among internal representations: those due purely to the constraints imposed on the learning process by sequential training and the existence of attractors.

The attractor picture and the observed correlations it creates among internal representations have a rather universal feature. The dynamic tagging of a given memory may produce persistent activities in several modules in the cortex. The representation of one stimulus in different modules may encode different features of the stimulus, such as color, shape, and so on. Different modules may be involved in the generation of different types of responses. One gets the impression, from the experiments of the Miyashita group, that the observed module in AVT is directly related to the pair association task (Sakai & Miyashita 1991), while it is not as directly related to the matching task. The universality intended here is that this does not matter, in the sense that wherever the attractor related to the response is, it must represent similar correlations of the internal representations because they depend only on the fact that one is learning one attractor while reverberating in a previous one.

This observation suggests that the lessons of these experiments may be used to speculate about human cognitive psychology. The fact that correlations form between the attractors representing semantically meaningless, uncorrelated stimuli implies that priming phenomena should be observed among stimuli of this kind.[2] The only condition is that the stimuli must be presented in an ordered sequence during training, independently of whether the attractor cortical network involved in the cognitive task can be observed. That priming should take place in a situation of correlated attractors can be concluded intuitively, and is confirmed in model networks. It is the mere observation that if the assembly is in a given reverberation, due to the priming stimulus, the test stimulus to be recognized will find it easier (and hence faster) to induce a transition to another reverberation attractor, the more similar the activity distribution in the latter attractor is to the priming attractor.

Thus, purely on the phenomenological, pretheoretical level, given the observation of correlated attractor representations, one is led to consider, for example:

1. Testing priming effects in humans with semantically meaningless stimuli, essentially imitating the Miyashita experiments, but measuring reaction time changes upon priming. One would expect a decrease in the priming effect with the distance in the training sequence, that is, with a decrease in the correlation between the internal representations.

2. Investigating the assumption that false positives (Anisfeld & Knapp 1968) are also caused by attractor correlations.[3] This hypothesis can be tested most clearly on sequentially memorized nonsense stimuli.

3. Testing the effects of the intensity of temporal correlations on the internal attractor correlations. In other words, one can prepare a set of stimulus sequences with an increasing proportion of sequences of fixed order, the others being random. One should expect a threshold behavior in the priming effect as a function of this proportion.

Similarly, the attractor interpretation of the experiments suggests informative extensions of the experiments in primate "cognitive neurophysiology":

(a) An immediate consequence of the framework proposed is that simple attractors (internal representations) must be formed prior to the correlated ones, as an intermediate stage in forming correlated attractors. Hence, one should find a point in the training process at which internal representations exist, but express no correlations.

(b) An experiment related to the one above is to train at length with patterns presented in a random sequence. Internal representations should form, but they should be uncorrelated. Their correlations with the activity distributions driven by the stimuli themselves would be invaluable to the solidification of the modeling effort.

(c) One can perform on monkeys the priming experiments suggested for humans above, observing attractor transitions in the cortical module under investigation.

(d) One can extend the experiments on pair association (Sakai & Miyashita 1991), which have been interpreted as attractor fusing. This can be done by measuring the correlations between pair representations in situations of partial ordering of the pairs in the training phase. Specifically, in the experiment, first stimulus images belonging to a pair were always shown during training, contiguously and in the same order, while pairs were selected at random. One can keep the strict ordering within each pair and have each pair followed by a given pair with some probability p, choosing the subsequent pair at random with probability $1 - p$.

(e) It is important to establish the structure of the distribution of activity in the module following the response, which underlies the scenario for learning correlations.

(f) But most important would be the identification of the additional representation modules in the cortex. These would be in different cortical or even extracortical, regions, as different dimensions of a stimulus seem to be stored at different stations along the cortical elaboration path (see Damasio & Damasio 1991).

7. Toy models and realistic modeling

Modern physics has a powerful methodology in that very structured phenomena in a complicated system are investigated in toy models. In other words, phenomena like superconductivity, magnetism, liquid crystals, and so on, are not investigated by starting from the well established dynamical laws of systems of nuclei and electrons. Instead, some essential features of the elements, deemed relevant to the structured, emergent phenomenon under investigation, are represented in a simple, tractable model. If the dynamics of the toy model actually produce the expected structure, the robustness of the phenomenon to the reintroduction of the omitted complexity of the underlying elements is investigated. This iteration serves both to justify the toy as well as to study further details of the emerging structures.

The discussion in sections 1 and 5 above has left us with a double task: (1) to provide a theoretical framework in which a plurality of stimulus-selective persistent activities can be embedded in a single assembly of neuron-like elements, and (2) to demonstrate that in such a framework correlations (context sensitivity) emerge between the reverberation (internal representations) corresponding to uncorrelated learned stimuli.

The first task has been performed by the Hopfield model (Amit 1989; Hopfield 1982), which has served as the basic toy model in describing the emergence of a diversity of structured attractors, robust to many types of random damage and noise. This was done by using an explicit form for a synaptic matrix, one that has a learning flavor, in the sense that the set of synaptic efficacies is constructed, in an additive fashion, from the correlations of activities of neural pairs in afferent patterns that are to be the attractors of the network. This is in the grain of the Hebbian paradigm: External stimuli to be learned impose activity distributions on an assembly, which in the learning process develops a set of synaptic efficacies that can autonomously maintain such activity distributions as reverberations. It is, after all, the synaptic matrix, and only the synaptic matrix, that can maintain the delay activity in the absence of the stimulus. The possibility of generating a synaptic matrix which endows a network with a large variety of different robust attractors in a single module was the main achievement of this model. Moreover, the formulation of the model allowed for the detailed computation of many properties of the dynamical response of the assembly to external afferent stimuli.

Furthermore, this toy model has also served to generate metaphors for several psychiatric and neurological pathologies. Hoffmann (1987) has used it to describe a distinction between mania and schizophrenia. Virasoro (1988) has used its properties under a random destruction of synapses (a lesion) to capture phenomena such as prosopag-

nosia. Yet this model has left unanswered a whole set of questions of detail. A representative list includes:

1. The model predicts high spike rates in attractors, whereas recordings produce rates much below saturation (Miyashita 1988). In fact, the attractors in the model are activity distributions in which about half the neurons in the assembly are quiescent while the other half emits spikes at saturation rates. This allowed modeling in terms of binary discrete variables for the neurons.

(2) The recorded coding levels[4] in a given reverberation are much lower than the 50% implied by the toy model.

(3) The model produced auto-associative networks, that is, with attractors close to the memorized stimuli, whereas experiment did not (Miyashita 1988; Sakai & Miyashita 1991).

(4) A bi-modal distribution of rates among neurons is predicted in an attractor, contrary to the empirical observation.

Early criticisms of the Hopfield model have concentrated on the type of synaptic matrices used. Symmetric, fully connected matrices, with random distribution of excitatory and inhibitory synapses on the axons of each neuron, and infinite analog depth[5] for each synapse, have made the analysis by theoretical physicists easier. They are, however, unrealistic impositions, because it is unlikely that the cortex will generate symmetric synaptic structures; cortex is connected at most at a level of 10% (Braitenberg & Schutz 1991), excitation and inhibition find their places on different neurons – Dale's rule. Yet these criticisms have been shown to be relatively innocuous. The synaptic dilution and the limited analog depth of the synapses have been treated by Sompolinsky (1986) and shown to affect the performance of the network only mildly. In fact, in some cases the performance per remaining material resource (such as storage capacity per surviving synapse) was even found to improve in the less ideal system.

The question of the low coding levels has also been found to be tractable. Its resolution has brought to light the fact that if one looks for a network with uniform thresholds for the neurons, the behavior of the network is strongly dependent on the choice of the representation for the neural states. If one insists on representing neurons by two state variables, there is a clear advantage in representing these states by (0, 1) over (−1, 1) (see Buhmann et al. 1989; Tsodyks & Feigel'man 1988).

A more elaborate modification of the dynamics of the original toy model was required in order to account for the relatively low spike rates in the attractors observed experimentally. It required a more detailed treatment of the single neuron dynamics, arriving at a description of neurons in terms of coupled systems of afferent currents and efferent spike rates. This description has produced networks with attractors operating far below the saturation of the neurons composing the network (see Amit & Tsodyks 1991a; 1991b). The modified networks, with low (arbitrary) coding levels and low spike rates preserved the main features of the original toy model: robust diversity of attractors, classifying stimuli auto-associatively. Two outstanding issues remained: auto-associativity and bi-modality.

The difficult problem of modifying a network to form attractors with correlations of the Miyashita type from uncorrelated stimuli learned in a fixed order found a solution with a flavor of the Miyashita training scenario at the level of the toy model, again indicating the usefulness of such models as drawing boards for new ideas. Auto-associative ANNs (attractor neural networks) are based on the idea that the synaptic matrix codes for the correlations of activities of pairs of neurons as induced by a given afferent stimulus. The neural pair correlations in different stimuli are coded independently of each other. It was shown (Griniasty et al. 1993) that when synaptic modifications, induced by training on a sequence of stimuli presented in a fixed order *also* record the correlations of the activities of pairs of neurons induced by one stimulus with that of its immediate predecessor in the sequence, the resulting attractors display correlations of the Miyashita type. The resulting attractors, each classified by the uncorrelated stimulus that had been learned and that excites it, are correlated for as far as five apart in the training sequence. This theoretical result has manifested itself in a dramatic way, in that each persistent delay activity (attractor) has a finite similarity index, with exactly five of the nearest stimuli in the sequence (Cugliandolo 1994). It is just the approximate range of significant correlations observed in the experiment. But this surprising five was deduced in the rather artificial context of the toy model.

Note that two types of correlations enter this discussion, and they should not be confused: one is the correlations between the activities of pairs of neurons during the presentation of stimuli for learning. These drive the Hebbian learning. Then, when learning has generated the synaptic matrix, the network's dynamics is controlled by that matrix. In particular, this synaptic matrix determines the structure of the attractors (delay activity distributions). The correlations between these attractors are of the second type. They are the ones measured by Miyashita.

The significance of the result is that:

1. Synaptic information about single-neighbor contiguity in the training sequence (i.e., the inclusion in the synaptic efficacies [learning] of activity correlations of pairs of neurons, one active as driven by a given stimulus and the other by the preceding one) was sufficient to induce correlations of the corresponding internal representations to a distance compatible with experiment.

2. The attractor picture underlying the model makes the information about a preceding stimulus in a training sequence naturally available in the persistent attractor at the time of presentation of the current stimulus. Recall that the network is supposed to code for activity correlations in successive first stimuli (in each trial). Any two such stimuli are presented with a separation of many seconds; but the existence of attractors allows the information of the one, first stimulus to be around for as long as is needed for the subsequent stimulus to be presented.

3. The form of the correlation (magnitude of coefficients and its rate of decay as a function of separation in the training sequence) was found to be independent (in a certain range) of the value of the contiguity amplitude parameter, which is the relative strength of the contribution to the synaptic efficacy due to neighboring images in the training sequence to the contribution and due to each image separately.

The phenomenon persists when the formal neurons are replaced by quasi-realistic, integrate-and-fire, neurons (Amit et al. 1994). An ANN operating with a synaptic matrix containing information on temporal contiguity in the training process preserves the main features of the attractor correlations in the systems of discrete neurons. The correlation coefficients (Kendall rank coefficients, as used in the experiments) of the network of realistic neurons, which also

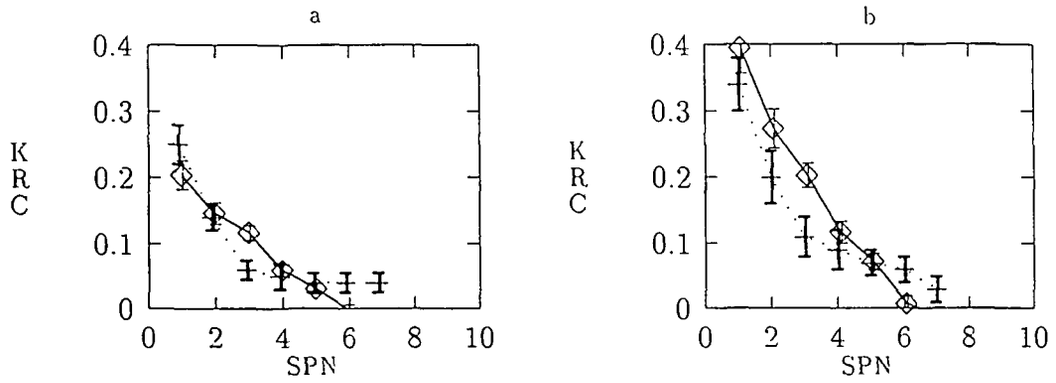

Figure 4. Confrontation with theoretical model (4000 neurons) in Amit et al. 1993: Kendall rank coefficients (as in Fig. 3) of attractors as a function of the separation of the stimuli in the training sequence. To each stimulus corresponds an attractor (reverberation): the one excited by it. The attractors are labeled by the serial position number (SPN) of the corresponding stimuli in the training sequence. The learned stimuli forming the synaptic matrix are uncorrelated, because the images presented for learning were uncorrelated; the activity distributions in the attractors are correlated. The error bars are standard errors in the sample of neurons, *a*. Correlations for regular sample second curve from top in Fig. 3, *b*. Correlations in enhanced sample, top curve in Fig. 3. Full curve, model results; dotted curve, experiment.

include reactive, separated inhibition, agree quantitatively quite well with the measured values (Fig. 4). Moreover, the more realistic model presents some additional features that brings it even closer to the biological experience:

The attractors expressing the Miyashita correlations do not exhibit a simple bi-modal distribution of spike activity among the neurons in the assembly (Amit et al. 1994). All previous ANN models produced sharp bimodality, either because the neurons were discrete (i.e., quiet or at saturation frequencies) or because the models implemented auto-associativity. Experiments manifest attractors, but not simple bimodality. The model predicts a large, stimulus-selective peak at very low spike frequencies and a wide distribution of rates among the active neurons. Consequently, the combination of realistic neurons and attractor correlations (i.e., the departure from auto-associativity) gives a potential response to the problem of the nature of the rate distribution.

To conclude this discussion we show one more measurement carried out on both the model and the performing monkey. In Figure 5 we show the distribution of activities produced in a given neuron in the reverberations provoked by the presentation of the complete set of learned stimuli. One sees the rate of spike emission by this neuron, in the delay period, for each of the stimuli plotted in the order in which they had been learned.

This is not the place to develop a more detailed discussion of the confrontation of theory with experiment (see Amit et al. 1994). The discussion has been opened here only to indicate that the insights gained by the interpretation of the Miyashita experiments in the language of attractor dynamics are accompanied by a candidate model which captures the experimental results to a very impressive degree of detail. Such a model can consequently serve as a starting point for the development of future paradigms in cognitive psychology.

8. Predictive theories

As a language and as a set of models, attractors offer a strongly predictive framework. They generate, as in section 6, several detailed experimental predictions even prior to the elaboration of detailed models. The models enlarge the predictive commitments of the approach. The detailed studies (Amit et al. 1994) imply the following:

(1) The appearance of the uncorrelated reverberations described in section 6, either due to short training periods

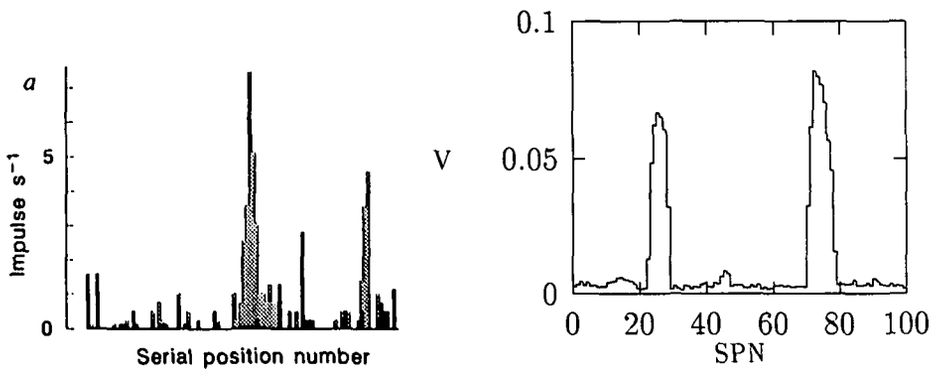

Figure 5. Average delay discharge rate versus serial position separation on a given cell. (a) Fig. 3a in Miyashita 1988. Reprinted with permission from *Nature* (Y. Miyashita, "Neuronal correlate of visual associative long-term memory in the primate temporal cortex," *Nature* 335:817–820, 1988). Copyright 1988 Macmillan Magazines Limited. (b) Model in Amit et al. 1994. This displays the level of activation of the particular neuron in the reverberation stimulated by each of the hundred stimuli in the learned sequence. The existence of two peaks indicates that this neuron participates in the representation of two uncorrelated stimuli. The side wings of each peak are due to the correlations of the attractors, developed in learning a fixed sequence.

or to training with random sequences of stimuli should be accompanied by narrower distributions of spike rates.

(2) The relation between the coding rate of the stimuli, the fraction of the module's neurons driven into high rates when the stimulus is present, and the coding rate in the corresponding attractor (Amit et al. 1993) is predicted.

(3) The statistics of the distribution of spike rates expressed on a given neuron by the entire set of stimuli memorized in the assembly (Amit et al. 1993) are predicted.

(4) If pair associations are formed by the training process as in Sakai & Miyashita 1991), then when the pair attractors become correlated by partially ordering the pairs (see end of sect. 5), the correlations of the pair attractors will be very high and will extend far down the sequence. This list can be continued.

9. Some provisos and defensive outlook

The above picture sounds too good. It may still have to undergo modifications. The main exposed flanks we perceive at this stage are:

(a) The Miyashita correlated reverberations might not be autonomous, in the sense that they are not really coded in the observed area, but reflect attractor activity in one or several other areas. Such modules could then drive the module observed afferently. This question should be given serious attention.

(b) The specific predictions follow from models with a prescribed synaptic matrix. This may be too restrictive. The models represent the hypothesized learning process (Griniasty et al. 1993) in a plausible way. What gives such synaptic matrix some credence is the relative robustness of the results to variations in it, provided the logic of learning a sequence of stimuli with fixed order is maintained. In particular, there is the robustness under variation of the contiguity amplitude parameter (e.g., items 1 and 3 in sect. 7, para. 10; Brunel, in press, presents a realistic learning dynamics based on an incoming flow of stimuli in a fixed sequence that leads to a synaptic matrix with attractors correlated as in Miyashita).

(c) The neural elements in the model networks may still be somewhat too idealized. This may be at the origin of some systematic differences in the details of the attractor correlation coefficients and rate distributions.

Even if some modifications are required, the reverberation picture is too fertile to be lightly discarded. We received it from Hebb aged some fifty years. It has only recently been given a direct empirical (neurophysiological) dimension (Miyashita 1988; Miyashita & Chang 1988; Sakai & Miyashita 1991) and has been endowed with a precise mathematical model (Griniasty et al. 1993; Hopfield 1982). One is tempted to start drawing a host of speculative conclusions from the new framework exposed by Miyahita's monkeys. The context provides fertile ground for adventures in cognitive psychology and even for some aspects of linguistics ranging from binding, which may be related to syntax, to priming, which may be extrapolated to semantics. It may even suggest a substrate for psychology itself.

Yet the lessons learned from these experiments include the one which advises restraint. It is just these experiments which indicate that our imagination concerning brain computation is still too much constrained by formal mathematics, by computer languages, and by artificial intelligence. In this connection it is well worth recalling the wisdom of John Von Neumann, writing 40 years ago:

> Thus the outward forms of *our* mathematics are not absolutely relevant from the point of view of evaluating what the mathematical or logical language *truly* used by the central nervous system is . . . the above remarks about reliability and logical and mathematical depth prove that whatever the system is, it cannot fail to differ considerably from what we consciously and explicitly consider as mathematics. (Von Neumann 1954, p. 82; author's emphases)

It is most likely that attending a while longer to the details of the contact between modeling and experiment will keep open options which a premature harvest of speculation would foreclose.

ACKNOWLEDGMENTS

I am indebted to Peter Hillman for a critical reading of an earlier version of this manuscript and to an anonymous *BBS* referee who has helped me improve the paper significantly.

NOTES

1. Some studies (Mason et al. 1991) find a small fraction of strong synapses. They are ascribed to multiple contacts between certain pairs of neurons. If that fraction is significant, it may lead to some interesting internal dynamics inside the attractor.

2. Priming is the experimental observation that the time for recognition of an incomplete pattern is shortened if the presentation of the stimulus to be recognized is preceded by the presentation of a cognitively related stimulus.

3. False positives is an experiment in which a subject is required to identify whether a given stimulus belongs to a subset of stimuli or not. The effect is that when the test stimulus is correlated (cognitively) with one of the items in the subset, the number of wrong "yes" answers increases.

4. The proportion of neurons in the assembly with elevated spike rates.

5. The ability to maintain with high precision a large number of different, closely spaced values.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Are single-cell data sufficient for testing neural network models?

Ehud Ahissar

Department of Neurobiology, The Weizmann Institute, Rehovot 76100, Israel. **bnehud@wiccmail.weizmann.ac.il**

Abstract: Persistent activity can be the product of mechanisms other than attractor reverberations. The single-unit data presented by Amit cannot discriminate between the different mechanisms. In fact, single-unit data do not appear to be adequate for testing neural network models.

There is considerable confusion in current brain research concerning the absence of working models for neural codings and internal representations. Thus, Amit's elegant attempt to explain neurophysiological data with the attractor neural network model is praiseworthy. However, I disagree with some of his basic claims.

His approach is based on attractors being visible via neuronal persistent activities, since persistent activities could only be maintained by network reverberations. I contend that persistent activ-

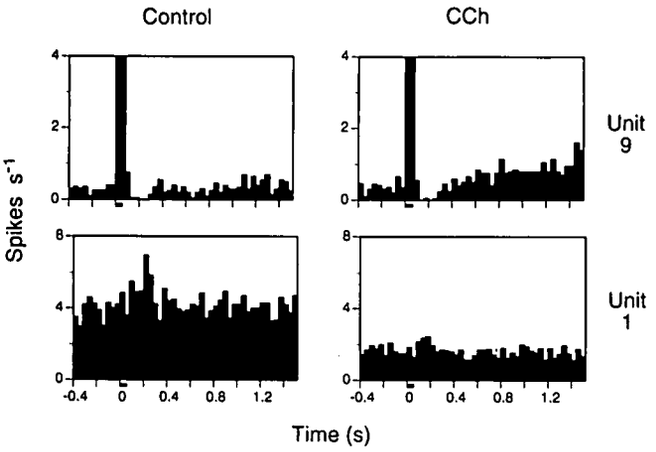

Figure 1 (Ahissar). Neuromodulator-induced persistent activity. Two units were simultaneously recorded via 2 different electrodes from the auditory cortex of an anesthetized guinea pig (see Haidarliu et al. 1995 for methods). PST histograms (bin = 40 msec) were computed from 4 blocks, each containing 280 presentations of tone bursts (8430 Hz, 50 msec duration) intermingled with other auditory stimuli. During the second and fourth blocks an acetylcholine agonist, carbachol (CCh, 1M), was iontophoretically applied from two electrodes: 50 nA from the electrode that recorded unit 1 and 150 nA from a third electrode. Distances between electrodes were 0.3–0.5 mm. While none of the units showed persistent activity in the control state (first and third blocks, left column), persistent activity was induced selectively in Unit 9 during CCh application. The depression of Unit 1 by this application was probably independent of the stimulus. The ordinate in the histograms of Unit 9 was truncated to emphasize the persistent activity. Thick horizontal bars near time = 0 denote stimulus duration.

ities can be maintained by other mechanisms, thus do not necessarily indicate network reverberations (attractors). For example, selective persistent activity of cortical neurons could be the result of a selective activation of diffused ascending systems such as the cholinergic and noradrenergic ones. The neuromodulators of these systems influence evoked activities in a stimulus specific manner (McKenna et al. 1989; Metherate & Weinberger 1989). Furthermore, acetylcholine can selectively induce persistent neuronal activity (Fig. 1). Thus, an alternative model to explain Miyashita's results could be that during the delay period ascending systems are activated to maintain persistent neuronal activity. Whether such neuromodulator-induced persistent activities are stimulus-specific is yet to be tested.

Testable predictions are required for discriminating between different models. Both critical predictions (whose rejection leads to a rejection of the model) and unique predictions are desirable. Of the predictions presented for the attractor model, none was defined as critical and most are probably not unique. When critical testing is not possible, as appears to be the case with neural network models and current experimental tools, models can still be tested by their degree of consistency with available data, as Amit did in his paper.

The reported consistency between Miyashita's data and the attractor model is indeed impressive. However, the same data also lead to inconsistencies, at least with the straightforward interpretation of the attractor model as presented by Amit. For example:

(1) Under "realistic" neuronal time constants, Amit's model cannot explain persistent activity at firing rates lower than 50 s^{-1} (Amit & Tsodyks 1992). (Amit & Tsodyks suggested that reverberating firing rates lower than 50 s^{-1} could be the result of longer time constants. However, Amit does not supply the time constants required to explain Miyashita's very low rates [usually less than 10 s^{-1}]; time constants which might actually be too long.)

(2) If the persistent activity of a recorded single-unit indicates the same attractor as the stimulated activity, as Amit suggests, the two activities should be correlated. However, this was not the case (Miyashita & Chang 1988).

(3) The monkeys' performance was not impaired when new stimuli (not previously seen by the monkey) rather than learned ones were presented (Miyashita 1988). The new stimuli evoked persistent activities that were not statistically different from the responses to learned stimuli, and thus should also indicate attractor reverberations according to Amit's view. In that case, when would the attractors for the new stimuli have been formed? If new attractors were not formed (i.e., single-unit persistent activity does not necessarily indicate attractor reverberations), then attractors are probably not required for performing the task.

(4) Amit argues that the neuronal associations between two different *first stimuli* had been formed as a result of activity correlations between the persistent previous attractor activity and the new stimulated activity. This is unlikely because (a) the persistent activity was weak and often not correlated with the stimulated activity, (b) in some paradigms persistent activity was significantly reduced already within the delay period (Sakai & Miyashita 1991)) and probably further during the inter-trial interval, and (c) the network was probably driven into a different attractor in at least half of the cases where the *match stimulus* differed from the *first stimulus*.

Amit claims that single-unit data is sufficient for observing attractors since the reverberations are reproducible on the single unit level, that is, the single-unit activity depends *only* on the stimulus leading to the reverberations (sect. 5, finding 3). From the data presented (e.g., Amit's Fig. 2), picking up single trials (single raster lines) at random clearly does not allow identification of the stimulus that evoked them. The source for this response variability is not evident from the data; however, factors other than the visual stimulus clearly affect these activities ("reverberations").

If one assumes that attractor neural networks operate in the cortex, one can describe single-unit activities by the same terminology. However, the aim of neurophysiological experiments is to allow discrimination between different models. Amit's target article does not explicitly provide critical tests for his model and does not address consistencies of the presented single-unit data with other optional models. I contend that single-unit data can supply comparable amounts of consistencies and inconsistencies for many possible models (see my Fig. 1) and as such are not suitable for conclusive testing of models. Current technology allows simultaneous recording of a few tens, and even few hundreds of single-units (Haidarliu et al. 1995; Kruger 1991; Wilson & McNaughton 1993). It is not yet clear whether these techniques can supply definite answers. Use of these techniques requires the expression of the predictions of theoretical models in terms of co-activation of several single-units.

ACKNOWLEDGMENTS

I thank Sergio Serulnik for providing the data for Figure 1 and for helpful discussions; B. Schick for editing. The experiments had been supported by the Israel Science Foundation administered by The Israel Academy of Sciences and Humanities.

Where the adventure is

Elie Bienenstock and Stuart Geman

Division of Applied Mathematics, Brown University, Providence RI 02912.
elie@dam.brown.edu and **geman@brownvm.brown.edu**

Abstract: Interpreting the Miyashita et al. experiments in terms of a cell-assembly representation does not adequately explain the performance of Miyashita's monkeys on novel stimuli. We will argue that the latter observations point to a *compositional* representation and suggest a dynamics involving rapid and reversible binding of distinct activity patterns.

Amit has interpreted Miyashita's remarkable experiments with clarity and precision. This provocative target article leaves us with

much to ponder. Among the many fascinating results reported in Miyashita's three *Nature* papers, we would like to highlight a small-print sentence at the end of the *Methods* section of the legend of Figure 1 in Miyashita (1988): "The monkey's performance level did not differ significantly for "learned" and "new" stimuli (85–100% correct)." Considering that this is so, that these "new" stimuli do *not* elicit the high-frequency delay discharge (7 or 8 spikes/s) found for "learned" stimuli, and, further and quite remarkably, that the monkeys' performance in this delayed matching-to-sample task is virtually invariant under a number of transformations affecting stimulus size, orientation, and color (Miyashita & Chang 1988), one may wonder: Why did Miyashita and coworkers fail to find any trace of the "attractors" corresponding to these new stimuli? Are these attractors located in parts of the brain different from the region of anterior ventral temporal cortex studied by Miyashita? Or are attractors necessarily the result of repetitive learning? How is it that the representation of these new stimuli, whatever that representation may be, serves the task of invariant matching? Could it be that the neural representation of a new stimulus, even when maintained for 16 seconds in what Amit calls the "tagged status," is *not* a Hebbian-type cell assembly , but is instead something more than a collection of neurons demonstrating a persistent elevated spiking activity?

We would tend to favor the latter hypothesis. We would further like to suggest that the monkeys' performance on new stimuli might make use of the *compositional* structure of these pictures, that is, their representation as a set of specific relationships – geometrical in this case – between subparts. Whereas the animals have no prior experience with the particular compositions that constitute the novel stimuli, they are most certainly familiar with many of the subparts and types of relationships, including edges, curves, junctions, symmetries, and so on. Each picture can in fact be viewed as the apex of a *compositional hierarchy* in which each part, such as a straight-line segment constituting a portion of a boundary, is itself a specific *relational* composition of more elementary subparts, such as local discontinuities. In principle, such hierarchies offer a *computational* framework for understanding perceptual invariances and context-sensitive recognition (cf. Bahl et al. 1983; Biederman 1987; Bienenstock & Geman 1995).

The representation of these hierarchies in neural machinery would evidently require the ability to bind component parts into a relationally-specific composite. Furthermore, this binding must be *dynamic,* inasmuch as many compositions are previously unfamiliar, as is the case with Miyashita's "new" stimuli. Amit acknowledges the need to bind activity patterns: "The linking of two separate representations is still an open problem, to be investigated both experimentally and theoretically" (final paragraph of sect. 2). We would like to suggest that this will require an interpretation of neural activity that goes beyond activation of subpopulations of neurons. The existence of specific relationships between parts must be specifically represented, and this cannot be done by merely co-activating populations of neurons that signal the respective pieces.

While these issues have been hotly debated in the last few years, in the context of perception as well as language (Fodor & Pylyshyn 1988), a biologically plausible set of ideas has been proposed some years ago by von der Malsburg (1981): the brain might achieve hierarchical/recursive compositionality by arranging the spiking times of interacting neurons into complex spatio-temporal patterns defined with millisecond accuracy (see also Bienenstock 1995; von der Malsburg 1987; von der Malsburg & Bienenstock 1986).

Miyashita-style experiments are typical in that they are not concerned with fine-temporal structure of neural activities. (This may be changing, amidst findings of correlations among spike-train activities across neuronal populations – see for instance Abeles et al. 1993.) As Amit himself points out, the representation of some entities as persistent spike-rate distributions is "almost a logical necessity" (beginning of sect. 5). We might surmise that such representations would be used for entities which – like faces, hands, bananas, or overlearned but otherwise nonsensical fractal pictures – are manipulated often and in an atom-like fashion, that is, without necessarily paying attention to their inner structure and/or syntactical relationships with fellow entities. The study of such neural representations, compatible with models involving Amit-style attractors, reflects a conception of the stuff of mind which one might call *atomic,* or *discrete,* or *all-or-nothing.*

But brains, and in particular human brains, constantly manipulate stuff that can hardly be construed as atom-like. When constructing the meaning of a sentence, we on the fly put together, in a highly content- and context-sensitive manner, fragments of meaning which, more often than not, had never been assembled before in our brain into just that composite structure. The same goes for much of nonverbal cognition. Furthermore, the very fragments assembled into a composite construct are hardly discrete entities themselves: they are, more often than not, composites too. Hebbian cell assemblies, and their modeling as point attractors in the dynamics of a neural network, do not account for this aspect of the complexity of mind. If there is an "adventure" today at the crossroads of neuroscience and cognitive science (Amit, sect. 9), it is rather, in our eyes, in the exploration of such *limitations* of the Hebbian associative-memory framework.

ACKNOWLEDGMENT

EB is on leave from CNRS, Paris, France. This work is supported by Army Research Office contract DAAL03-92-G-0115 to the Center for Intelligent Control Systems, National Science Foundation grant DMS-9217655, Office of Naval Research contract N00014-91-J-1021, and Advanced Research Projects Agency grant MDA972-93-1-0012.

Reverberation reconsidered: On the path to cognitive theory

Eric Chown

Department of Computer Science, Oregon State University, Corvallis, OR 97331. **chown@research.cs.orst.edu**

Abstract: Amit's work addresses a critical issue in cognitive science: the structure of neural representations. The use of Hebbian cell assemblies is a positive step, and we now need to consider its role in a larger cognitive theory. When considering the dynamics of a system built out of attractors, a more limited version of reverberation becomes necessary.

Amit's target article is a welcome look at work which, as he points out, has been too often ignored by psychologists and connectionists alike. The Hebbian cell assembly concept has the potential to fill a gap in cognitive science by forming a kind of bridge between the symbols used in artificial intelligence on the one hand, and the neural elements used by connectionists on the other. However, while Amit shows how a toy model can stimulate some experimental evidence he does not take the next step to consider the dynamics of a cognitive system consisting of countless numbers of attractors.

As Amit points out, perhaps the most important role of the cell assembly is that it keeps a stimulus "in the head" even when that stimulus is no longer present. Once this point is accepted it raises the following questions: How long does the representation of the stimulus remain active? What eventually causes the representation to become inactive?

The duration of reverberation. Both Hebb and Amit favor long lasting reverberation. One of the important reasons for this would appear to be its use in providing an explanation for short term memory. However, such a conceptualization raises at least as many problems as it solves. Many of these problems are related to learning. In a system where reverberation serves as short term memory at any given time there will be a host of active representations. Since Hebbian learning is based upon correlated activity all of these active representations will tend to become more strongly connected with each other. There are many reasons to believe that

learning is much more selective than this. Furthermore, as new inputs continue to come into the system and more attractors are activated each new active representation will in turn prime a large number of other representations both active and inactive. Representations can and will become active without external stimuli. Management of activity becomes of paramount importance in such a system. Scaling this type of system to deal with the constant barrage of widely varying inputs would seem to be an unmanageable task.

Fortunately, short term memory is easily achievable in a cell assembly without long reverberation times. One simple way to do this is by a temporary change in synaptic efficiency among the cell assembly's neurons. Such a change means that the representation can be easily reactivated after reverberation has ceased. Post tetanic potentiation (PTP) is an example of a well known neurophysiological process which could provide this type of functionality (Magleby 1987). A key advantage of shorter reverberation times is simplicity; only a few cell assemblies are active at any one time. Learning in such a system would be much more selective and focused with a strong bias towards associating representations which were activated closely in time. In other words no special modifications are needed to the Hebbian paradigm. Control is also simpler in such a system; with only a few active representations there would be far less priming of attractors.

What is a reasonable length of time for a cell assembly to reverberate? There is strong evidence from the field of memory consolidation that learning for an event is generally completed during a five second time period immediately following the event (Miller & Marlin 1984). If learning and activity are indeed correlated as Hebb proposed, this implies that cell assemblies are active for approximately five seconds. At a rate of about one newly active cell assembly per second this means that there will be in the neighborhood of five active cell assemblies at any given time, a figure which corresponds nicely with Amit's own magic number for how far apart elements in a sequence should correlate. It also matches the average number of elements found to be held in working memory (Mandler 1975).

The termination of reverberation. One of the problems with an attractor model is that the feedback built into the system, which was necessary to form the attractor in the first place, will tend to keep the attractor active indefinitely. This problem becomes exacerbated by the improved synaptic efficiencies associated with short term memory. Some mechanism is needed to ensure the timely transition from one set of active representations to the next. Note that it is not enough to rely on inhibition generated by the system's response to a new stimulus since the flow of consciousness continues in the absence of new stimuli. It is not clear from the target article that Amit has considered this issue as yet.

One mechanism that has been proposed for this purpose is neural fatigue (Kaplan et al. 1991). While neural fatigue has not been generally accepted there is growing evidence to support it (Artola & Singer 1993; Atwood & Nguyen 1990; Ito 1992). Fatigue can serve several purposes in the dynamics of the cell assembly (Kaplan et al. 1991; Chown 1994). In this context the most important of these is that the intense activity of an attractor naturally fatigues the neurons involved. This in turn makes them relatively ineffective at firing each other, thus reducing feedback in the attractor, and eventually causing the activity of the attractor to cease. The result is that cognition will naturally move from one active set of representations to the next without the need for elaborate control mechanisms.

Although Amit's effort to bring the cell assembly back into the mainstream of cognitive science is to be applauded, care must be taken to ensure that the problems inherent in using the cell assembly as a basis for a model of cognition are not ignored when using it to model specific problems. Amit's work represents a small portion of the rich space afforded by Hebb's work, a space that needs to be more thoroughly explored.

What's in a cell assembly?

G. J. Dalenoort and P. H. de Vries

Department of Psychology, University of Groningen, 9700 AB Groningen, The Netherlands. **g.j.dalenoort@ppsw.rug.nl**

Abstract: The cell assembly as a simple attractor cannot explain many cognitive phenomena. It must be a highly structured network that can sustain highly structured excitation patterns. Moreover, a cell assembly must be more widely distributed in space than on a square millimeter.

In his target article, Amit draws attention to the importance of the cell assembly, a concept that has only scarcely received attention in the literature since its introduction by Hebb (1949), in contrast to the learning rule usually named after Hebb, but already proposed in 1893 by E. Tanzi (see Dalenoort 1982). In fact, the importance of the idea of cell assemblies is comparable to that of the learning rule and both are indispensable in models of the brain.

Amit provides an interpretation of neurophysiological data from Miyashita et al. 1988a; 1988b; 1991) within the theoretical framework originated by Hebb, and this is an important contribution. It shows the need for the existence of cell assemblies, although one might also consider the lesion experiments of Lashley (e.g., 1951) as a demonstration of their existence. In fact, Lashley's results would point to a more distributed form of cell assemblies than Amit envisages. Following the experimental results of Miyashita and his colleagues (Miyashita 1988; Miyashita & Chang 1988; Sakai & Miyashita 1991) he seems to localize a cell assembly in a cortical surface of about 1 mm^2. Neither Miyashita nor Amit are very explicit: Are there no neurons outside that region belonging to a given cell assembly? How many cell assemblies may have neurons within that region? The answers to such questions are vital for the more general question of the biological basis of the huge capacity of human long-term memory. Amit says little about the distributedness of cell assemblies, for example, point 5 in sect. 5. In our own model we go much further. We assume that the biological basis of a memory trace must be a robust and identifiable excitation pattern, with much more structure than is given in a single attractor. Also, a concept can play very different roles in different contexts. This makes the physical picture of attractors insufficient to bring out the intricacies of human cognition. To use a metaphor: it would be like the claim that a complete description of a conversation would be possible in terms of "the language of sound waves." But no one believes that the gist of a conversation (the meaning) can be also adequately described at the physical level (more about levels in Dalenoort 1990).

Amit's square millimeter may correspond to a nucleus within a cell assembly. Such a cluster of neurons might have to be strongly localized in order to obtain a clear point in space for interactions with the sensory and motor systems. For almost any interconnected set of neurons there must be a critical threshold. When a sufficient proportion of the neurons are excited, the overall excitation level will continue to rapidly grow autonomously to its maximum. For the overall activity in the network to remain within bounds, there must be a mechanism to restrict the number of fully excited cell assemblies. This leads to the assumption that the super-threshold excitation of a cell assembly corresponds to processing that at the psychological level requires attentional effort, whereas subthreshold excitation corresponds to automatic processing.

A cell assembly has a complex organization. Amit quotes Miyashita in relation to the context sensitivity of cell assemblies: the occurrence of a reverberation is correlated with that of reverberation that did develop in the same context. This observation does lead to the assumption that within a cell assembly various subensembles of neurons can be distinguished that correspond to the different contexts in which the assembly can play a role. The activation of such a subassembly thus triggers a context-specific activation of another cell assembly. The existence of these subassemblies accounts for the finding that priming depends on context. One can conclude that there does not exist a clear

borderline between two connected cell assemblies. One should speak of neurons not as belonging to a cell assembly, but as participating in the activity pattern corresponding to a memory trace. For the stability of the interaction between cell assemblies, inhibitory mechanisms are necessary. Such mechanisms may be implemented in the subensembles of neurons through which two assemblies interact. One necessary interaction is what we have come to call "backward inhibition": a highly active cell assembly – at a level above that of the critical threshold – inhibits the cell assemblies from which it received excitation. Such mechanisms presuppose a differentiated internal structure which is hard to incorporate in the models based on attractors; they seem to be necessary to explain serial processes.

A further internal structuralization of a cell assembly can be derived from the following observation from psychology: It is possible for a human being to associate instantaneously two arbitrary memory traces, to remember them, and to reproduce them in an association later on. At the neural level, this implies that a temporal connection can occur between any two cell assemblies provided they are activated in the same context. The neural representation of context is a global excitation pattern. So far, few hypotheses have been put forward concerning the way context can play a role in temporally connecting two cell assemblies. The only possibility might lie in resonances between spike trains, an idea that can be found in recent literature in various forms. This mechanism would also provide a solution to the binding problem in connectionism, as mentioned by Amit.

Dependent on the context and the psychological task, a cell assembly can sustain different temporal connections. Several subensembles of neurons must therefore participate in the cell assembly; each of these produces the spike resonance characteristic of a temporary connection. These subensembles exist next to the ones that implement the permanent connection of a cell assembly.

In the foregoing paragraphs we have described various interactions between cell assemblies. Such interactions are necessary if one takes Hebb's approach seriously and aims to represent psychological tasks in networks of cell assemblies. The complex transitions of excitation patterns that occur in such networks may not be easily represented by means of attractors. In such a representation only a single attractor is distinguished for each task or problem (Hopfield 1982). Tasks involving several cell assemblies and activation transitions between them, are therefore problematic.

How representation works is more important than what representations are

Shimon Edelman

Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel.
edelman@wisdom.weizmann.ac.il

Abstract: A theory of representation is incomplete if it states "representations are *X*" where *X* can be symbols, cell assemblies, functional states, or the flock of birds from *Theaetetus*, without explaining the nature of the link between the universe of *X*s and the world. Amit's thesis, equating representations with reverberations in Hebbian cell assemblies, will only be considered a solution to the problem of representation when it is complemented by a theory of how a reverberation in the brain can be a representation of anything.

It is possible (and, according to some, necessary) to distinguish two problems about mental representation (Cummins 1989). The first of these, the Problem of Representations, is basically empirical and has to do with the nature of the representations used by a given cognitive system (such as the brain of a monkey). Amit's thesis, equating long-term memory representations in the monkey visual system with reverberations in certain cell assemblies in the inferotemporal (IT) cortex, outlines a possible answer to the problem of Representations, mobilizing in support of the proposal an impressive battery of data from neurophysiology and from the mathematics of attractor systems. The target article, however, does not take a stand on the second problem, called by Cummins the problem of *Representation* (singular). Here, the central question is how, in principle, can a mental state, as realized, for example, by a reverberation in IT, *refer* to anything at all in the world (that is, what makes a given brain event the representation of, say, apple).

Let us consider the comparative value of theories addressing one but not the other of the two aspects of the problem of representation mentioned above. Note first that one can formulate a useful (i.e., predictive) theory of mental processes while treating the individual representations as black-box entities causally related to events in the world (and to each other). In the philosophy of mind, for instance, some of the doctrines that account for behavior in terms of the agent's beliefs and desires treat internal representations as such unanalyzable entities. The great practical value of folk psychology as an explanatory and predictive tool attests to the possibility of a methodological separation between the question of what representations are and what they are good for. In other words, a solution to the psychophysical problem – getting meanings into and out of the head – can stand on its own, no matter what is, precisely, the meaning of "meaning."[1]

Whereas solving the psychophysical problem would constitute a most significant advance on the way to understanding Representation, unraveling the mechanism (such as dynamics of cell assemblies) whereby individual representations may be realized is, at best, a small step (albeit in the right direction). To mix a Fodoresque metaphor with a simile borrowed from Harnad's paper on symbol grounding, seeking out and cataloguing representations whose causal relations with the world are unknown or ill understood amounts to trying to learn Mentalese from a Mentalese/Mentalese dictionary alone. Granted, a dictionary of representations is better than nothing, but without grounding at least some of the entries in reality it is difficult to make progress (Harnad 1990).[2]

It is interesting to note that experimental neurobiologists appear to take the psychophysical aspect of the problem of representation much more seriously than some theoreticians. One example in support of this observation can be found in Sakai et al. (1994), where the researchers describe an ingenious method of making sure that the stimulus associated with the response of a certain unit in monkey IT cortex is indeed the optimal one for that unit. In one of the experiments there, the stimulus shape was varied parametrically, leading invariably to a decrease in the unit's response. Under certain conditions on the shape space, this may be interpreted as evidence in support of the original stimulus being the optimal one. (Note that such evidence, if corroborated, would constitute a solution to the psychophysical problem for that particular neuron.)

Another example of progress in this direction is the groundbreaking technique of stimulus reduction, developed by Tanaka and his group (Fujita et al.1992; Tanaka 1992). Once identified as effective for a given neuron in IT, a stimulus shape in a typical experiment undergoes successive simplification until it no longer elicits a response from that neuron. The last stimulus in the reduction sequence that is still effective is the elementary feature that lies in the intersection between the set of features present in the original stimulus and the shape-space "receptive field" of the neuron. The representation inherent in the firing of this neuron is thus grounded in the external world. Note that by narrowing down the range of stimuli to which a particular representational event may refer, the reduction technique offers a partial relief, at a reductionist neurophysiological level, of the predicament of indeterminacy of radical translation, illustrated by Quine's celebrated Gavagai example (Quine 1960).

Stretching somewhat the sense of "footnote," one may consider this emerging neurophilosophy of representation as another foot-

note to Plato. In *Theaetetus* (360 B.C., Socrates offers (and criticizes) a number of metaphors for the concept of knowledge, one of which is a flock of birds within the cage of one's mind. To solve the problem of representation, it is not enough to understand the dynamics of flocking: one must also find out how to govern the flight of the birds from without the cage.

NOTES

1. In fact, meaning does not have to be in the head at all (Putnam 1988, p. 73), provided that whatever *is* in the head obeys a well-defined causal relationship with what is "out there" in the world (Edelman 1995; Locke 1690).

2. As an edifying example, consider Ijon Tichy's attempt to learn the meaning of the Ardrite world "scrupt," and its consequences, as recounted in the fourteenth voyage of the *Star Diaries* (Lem 1985, p. 103).

The Hebbian paradigm reintegrated: Local reverberations as internal representations

Walter J. Freeman

Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720-3200. **wfreeman@garnet.berkeley.edu**

Abstract: Recurrent excitation is experimentally well documented in cortical populations. It provides for intracortical excitatory biases that linearize negative feedback interactions and induce macroscopic state transitions during perception. The concept of the local neighborhood should be expanded to spatial patterns as the basis for perception, in which large areas of cortex are bound into cooperative behavior with near-silent columns as important as active columns revealed by unit recording.

Reexcitation is not reverberatory and a pixel is not a picture: Amit's "local neighborhood" neglects the hierarchical organization of cortex.

Recurrent excitation (positive feedback among excitatory neurons) has been found in olfactory cortex (Freeman 1967; 1975) due to synaptic interaction among pyramidal cells; it is essential to account for the long and variable time constants of cortical populations and for changes in evoked potential wave forms when animals are trained to respond to the electrical pulses used as conditioned stimuli (Emery & Freeman 1969; Freeman 1968). Mutual excitation in the olfactory bulb manifests a nonzero point attractor with the property that increased activity evoked by an electrical pulse decays faster to the baseline; this enables neural populations to provide stable excitatory biases and to maintain oscillatory populations having negative (inhibitory feedback) within near-linear dynamic ranges (Freeman 1979). Long-range feedback between populations sustains self-organized, stable background activity, which is aperiodic and probably chaotic (Freeman 1987). Such systems have multiple stable attractors, which are accessed itinerantly by variations in the strength of excitatory biases provided by sensory input (Freeman 1992) and by internal centrifugal, regenerative feedback under neuromodulatory (Freeman 1993). Amit's concept of background neural activity sustained by reentrant synaptic excitation is therefore well supported by extant data.

A flaw I see in Amit's formulation is the "local" origin of the self-sustained activity in a modest 100,000 neurons, in which connectivity is "on the order of a few percent." Anatomical measurements (Braitenberg & Schüz 1991; Sholl 1956) show that cortical connection density is 100 times more sparse. Each neuron sends to and receives from thousands of others, but packing density is very high, so each neuron interacts with its neighborhood, not with other single neurons. That neighborhood provides a local mean field intensity from weak but widespread synaptic interactions, which is a macroscopic property. Amit's choice of the word "reverberation" aptly conveys his thinking, because it is a repetitive pulsing like the echoes of a thunderclap or the clacking of pool balls on first break. Cortical neurons don't interact that way. Their mean pulse rates are low, and their pulse interval histograms are close to Poisson with a dead time, changing to Gamma distributions of order one half as the mean rate increases. Rarely are they periodic.

These are not the histograms from "integrate-and-fire" neurons like sensorimotor neurons in electric fish, which are highly periodic. Our models indicate that background activity is a macroscopic property (Freeman 1975; 1992) from a point attractor that maintains each neuron in a population just below its threshold. The "ripple" on the 10-100 pulses each millisecond on its dendrites, with distributed delays from surrounding cells, acts like thermal noise to trigger pulses almost at random, independently from its neighbors. The background bias serves to create a field of white noise, against which low levels of spatially coherent activity stand out clearly under spatial integration, by divergence in cortical efferent pathways (Freeman, in press). "Reverberatory" is unsuitable for pseudorandom pulse distributions and "1/f" spectra of associated dendritic currents (Freeman & Barrie 1994). "Self-organized steady-state" or "oscillatory background" activity is preferable.

Local neighborhoods of the size of cortical columns have neurons showing common patterns of unit activity, in terms both of mean firing rates and the frequency of modulation of probability of firing (Freeman 1975) – not surprisingly giving vigorous activity in clusters of neurons as reported by Miyashita. However, cortical patterns in perception cover areas that are sampled at a neighborhood with each electrode. "Silent" columns are as necessary as "noisy" columns, though the latter get the credit. Changing the "code" from one neuron to a local neighborhood enlarges the pixel size, but it doesn't get the picture in a spatial "code." Black, white, and gray are needed, but units only show white spots.

Leaving aside my reservations concerning "representations" (Freeman 1995; Skarda & Freeman 1987), I agree with Amit's caveat concerning the limitations of computation in modeling brain function. I would go beyond von Neumann, however, in saying that brains don't compute at all (just as a lens doesn't compute a Fourier transform) because they don't use numbers. The control of macroscopic state transitions between learned attractors by microscopic sensory inputs holds the essence of perception. This hierarchical organization is lacking in Amit's Hebbian networks, with resultant puzzling discrepancies between his model's outputs and Miyashita's observations.

Not the module does memory make – but the network

Joaquín M. Fuster

Department of Psychiatry and Brain Research Institute, School of Medicine, University of California at Los Angeles, Los Angeles, CA 90095. **joaquin%chango.dnet@loni.ucla.edu**

Abstract: This commentary questions the target articles inferences from a limited set of empirical data to support this model and conceptual scheme. Especially questionable is the attribution of internal representation properties to an assembly of cells in a discrete cortical module firing at a discrete attractor frequency. Alternative inferences are drawn from cortical cooling and cell-firing data that point to the internal representation as a broad and specific cortical network defined by cortico-cortical connectivity. Active memory, it is proposed, consists in the sustained activation of the component neuron populations of the network.

The main contribution of Amit's target article is to reaffirm two plausible principles: (1) An internal representation is defined by a connective structure in the cortex; (2) Active memory, that is, the activation of an internal representation, consists in the reverberation of impulses within its connective structure. The article is less successful in its reductionistic attempt to limit the representation to a discrete cortical module and active memory to a discrete attractor frequency. Our empirical evidence, some of which is summarized below, supports neither of these limitations. Instead, it expands the internal representation to a much broader connec-

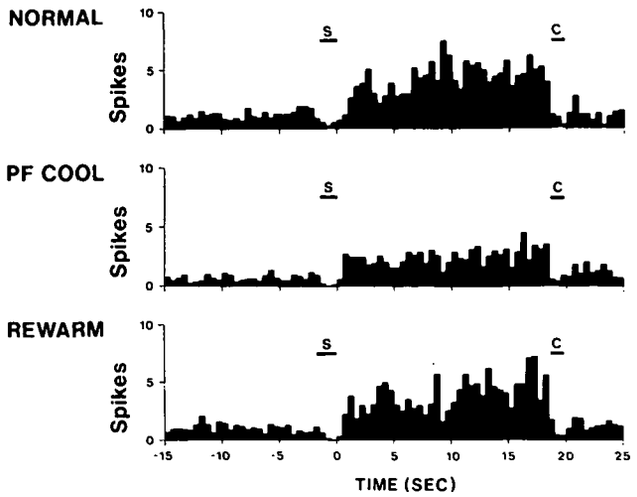

Figure 1 (Fuster). Average firing of an inferotemporal cell in delayed matching to sample, at normal cortical temperature, during prefrontal (PF) cooling (20°C), and after rewarming to normal temperature; 20 trials in each condition. *S* marks the sample period and *C* the choice (match) period. Prefrontal cooling reversibly attenuates the normal activation of the cell during the 18-sec retention period (delay). From Fuster et al. 1985.

tive structure that straddles several cortical areas, in other words, to a cortical network. According to this view, the activation of the internal representation, in behavior as in the cognitive domain, would simply consist in the sustained neuronal firing of the widely dispersed network nodes above certain levels or thresholds.

While a monkey is waiting to make a manual choice contingent on a prior sensory stimulus (e.g., the sample stimulus in a delayed matching task), many cells, in many cortical areas, undergo sustained elevated firing. If the stimulus is visual, such *memory cells* can be found in various areas of inferior temporal cortex, including AVT (Fuster 1990; Fuster & Jervey 1981; 1982). In addition, during the same task and responding to the same stimulus, visual memory cells can be found in prefrontal cortex (Fuster et al. 1982; Quintana et al. 1988). In this cortex, to be sure, there are fewer such cells, and they discriminate the stimuli less sharply than inferotemporal cells. Furthermore, also in prefrontal cortex, there are many memory cells that are attuned to the impending motor response, not the stimulus. Nonetheless, the point I wish to make is that visual memory cells, in a visual memory task, are not circumscribed to a discrete portion of temporal cortex. Rather, they seem part of a wide network, however *finite* and *specific* for the task, that extends as far afield as the frontal lobe. Apparently, that network ties together neuron assemblies in all the cortical regions representing the physical or associated attributes of the stimulus-memorandum: color, brightness, shape, location, motor responses (ocular or manual), and reinforcement (reward). Thus, in the delay period of the memory task, by virtue of associative links with all those components of its internal representations, the stimulus elicits sustained activity in all their respective areas of representation.

Other evidence indicates that the sustained cell discharge within any area or module of cortex is neither autonomous nor encoded by a given attractor frequency, as Amit postulates. Such is the evidence of functional relationships between distant cortical areas during active memory (Fuster et al. 1985; Quintana et al. 1989). Consider the inferotemporal cell in Figure 1. After inhibition during the sample stimulus, the cell shows sustained elevated discharge during the delay. Since the two stimuli of the task, red and green, induce just about the same degree of activation of the cell during memorization, one could speculate, with Amit, that the two lead to the same "attractor," one presumably encoding an attribute common to both stimuli. Note, however, that the cooling of dorsolateral prefrontal cortex, a region critically involved in any memory task (Fuster 1989), significantly and reversibly attenuates the sustained firing of the inferotemporal cell, while causing the monkey to make errors of performance. There are two reasonable explanations for the effect of prefrontal cooling on the inferotemporal neuron. Both are mutually compatible and probably true. One is that the prefrontal cortex is normally responsible for maintaining or at least modulating the mnemonic discharge of inferotemporal cells, which are centimeters away, probably by way of fibers in the uncinate fasciculus. The other is that the prefrontal cortex is in the reverberating loop that sustains the activity in two distant locations of the activated cortical network, one frontal and the other temporal. In any case, the frequency of discharge of the cell during memorization is susceptible to prefrontal influences – depressed by prefrontal cooling – and not solely determined by either the physical properties of the stimulus or the properties of a local attractor circuit.

Further evidence against mnemonic encoding by specific attractor frequencies comes from the close examination of the firing pattern of any given cell in response to a stimulus-memorandum. That pattern is far from uniform. It varies widely from trial to trial with the same stimulus (see Fig. 2 of target article), as well as within the delay of any given trial (Fuster 1990). Observe the delay activity of the cell in my Figure 2. This one clearly distinguishes the memory of red from the memory of green. Yet note that it does so only during the first few seconds of the delay. In fact, after the red memorandum the cell shows, *on the average across trials*, a sharp peak of excitation and then a gradual descent to baseline level of discharge. A closer analysis of the cell's discharge, at the millisecond level, reveals wide fluctuations of firing between two or more frequencies. The hidden units of one of our models of recurrent (i.e., reverberating) network (Zipser et al. 1993) mimic with astonishing similarity those fluctuations between attractor frequencies. (The fact that our computer model had undergone supervised training – by backpropagation – is irrelevant to the argument on the behavior of the fully trained network with established synaptic weights.)

Aside from the just-mentioned evidence, which casts doubts on local attractors encoding specific cues with local reverberations, one wonders why those attractors would be needed at all. Amit acknowledges the possibility that the empirical evidence he adduces for his argument and for his model could be explained within a conceptual framework that would be more in line with mine (Fuster 1995) – to wit, his first point under "provisos and defensive outlook" (sect. 9). Indeed, an internal representation can be understood as one of many widely distributed and overlapping cortical networks, each constituted by the neuronal populations that encode *in their connective structure* all the associated elements of the representation, including context. The activation of a memory, "content-addressed" by one of those elements (e.g., a sensory stimulus), can be maintained by reverberating impulses through reentrant circuits within and between those neuronal populations. In terms of the behavioral or cognitive efficacy of the activated memory, firing changes in the course of time are to be expected anywhere in the network, for its internal dynamics must adjust to the changing temporal relevance of network components – and the concomitantly changing *focus of representation*. More minute fluctuations around a certain average, at certain scales of time and of cortical tissue, may not matter. For, in the aggregate, the variable and probabilistic discharge at the cellular and modular levels translates into stability and determinancy of representation at the level of the network.

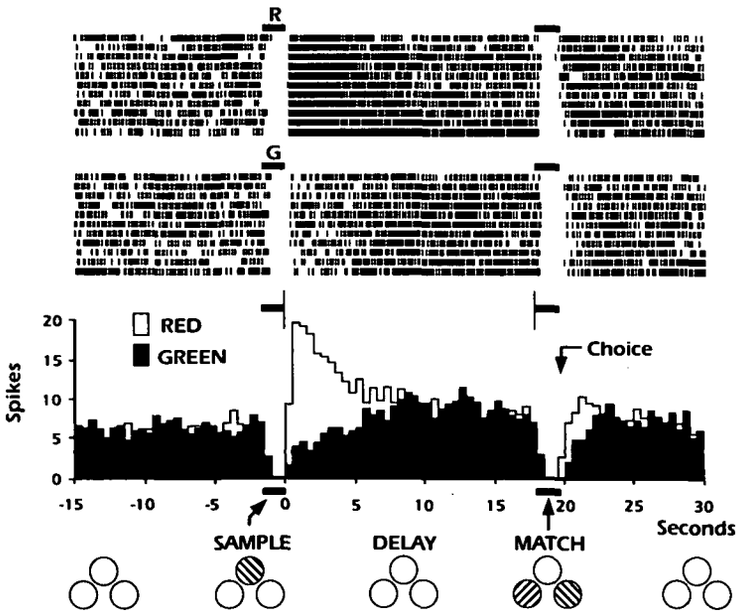

Figure 2 (Fuster). Rasters and average frequency histograms from an inferotemporal cell during delayed matching to sample with two patterned stimuli differing in color, one red and the other green. The monkey performs the task – diagrammed at the bottom of the figure – on a triangular array of translucid stimulus-response buttons. Each trial begins with presentation of the sample stimulus, red or green, in the top button; after a period of delay (about 18 sec), two stimuli appear simultaneously in the lower buttons; the monkey must choose the one matching the sample in pattern or color. Sample and location of correct choice change at random from trial to trial. The cell is inhibited in both sample and match periods, but activated during the memorization–delay–period; this activation is color-differential, higher for red than for green. From Fuster 1990.

Mathematics of Hebbian attractors

Morris W. Hirsch

Department of Mathematics, University of California, Berkeley CA 94720.
hirsch@math.berkeley.edu

Abstract: The concept of an attractor in a mathematical dynamical system is reviewed. Emphasis is placed on the distinction between a cell assembly, the corresponding attractor, and the attractor dynamics. The biological significance of these entities is discussed, especially the question of whether the representation of the stimulus requires the full attractor dynamics, or merely the cell assembly as a set of reverberating neurons. Comparison is made to Freeman's study of dynamic patterns in olfaction.

Computational modelling "is the core and raison d'être of the entire discipline of nonmedical neuroscience" (Amit 1994, p. 429). Numerical simulation of a precisely defined model may reveal unexpected, biologically meaningful properties. But only mathematical analysis can rigorously confirm them.

An attractor in a dynamical system is a set A of states such that any trajectory starting in A stays in A, and any trajectory starting sufficiently near A limits at a subset of A. Stable fixed points and limit cycles are the best understood attractors, and practically all neural network models are restricted to these types; but they are mathematically very special, and biologically implausible in many situations (Skarda & Freeman 1987). The alternatives – "chaotic" or "strange" attractors – have disgracefully little rigorous mathematical theory.

In the dynamical systems implicit in this article (as in Zeeman 1962), there is one state variable for each of the N neurons in a module: the neuron's firing rate, modeled by a continuous function of time. A state of the system is specified by a vector listing these rates, so that the state space is the N-dimensional vector space $\mathbf{R}^N$. As time proceeds the firing rates describe a curve in state space, called a trajectory of the system. These trajectories are assumed to be solutions of a system of differential equations.

A brief stimulus to some of the neurons in a module sets the initial firing rates. Amit argues persuasively that there must be persistent activity for a comparatively long time period after the stimulus has been shut off. He postulates that during this time the trajectory has approached an attractor A, whose dynamics is the reverberation reliably detected in single-neuron recordings. A is Amit's candidate for a representation of the stimulus. The neurons participating in the attractor dynamics, detected by their unusually high firing rates, form a Hebbian cell assembly H.

There is an important conceptual distinction between:

(1) the cell assembly CA, which is a set of neurons;

(2) the attractor A, a subset of the state space $\mathbf{R}^N$ composed of trajectories;

(3) the attractor dynamical system DS, obtained by considering only trajectories in A.

DS contains the most detailed information. A without the dynamics is an elusive geometrical object, from which important numerical dynamical invariants such as Liapunov exponents and entropy can be estimated. CA contains little dynamical information, but is easier to detect.

As an illustration, consider an assembly of $N = 3$ cells participating in an attractor, each reverberating periodically. If their periods have rational ratios, then AT is a topological circle. If two of the three periods have irrational ratio, the attractor AT is a torus (surface of a doughnut) in which the trajectory dense. And more complicated things can happen – the attractor can be a fractal, or a knot of arbitrary type. Extremely complex dynamics can occur robustly; yet in all these cases CA consists of exactly the same three cells. The complications for large N are almost inexpressible, largely unexplored, and poorly understood.

Amit leaves a key point unresolved: Is the stimulus represented by CA, by A, or by DS? The answer has crucial implications, for it constrains how the rest of the nervous system makes use of the stimulus token. If the full attractor dynamics DS are needed to respond to the stimulus, then there must be sophisticated process-

ing which distinguishes between different dynamics in the same set of neurons. If only *CA* is used, then the nervous system performs the much simpler task of distinguishing different reverberating group of neurons – at the price of ignoring the rich cognitive possibilities inherent in the dynamics.

Single cell probes may be useful in this problem, as they permit time-lag reconstruction of the dynamics (Packard et al. 1980; Takens 1981). Amit (sect. 2, para. 5) seems to identify an attractor with the cell assembly, ignoring the dynamics. But he also speaks of "activity distributions in the reverberations" as representations of stimuli; this seems close to the W. Freeman's emphasis on spatial patterns of neural activity (Freeman 1987; Freeman & Baird 1987; Freeman & Viana di Prisco 1986). [See also Skarda & Freeman "How Brains Make Chaos in Order to Make Sense of the World," *BBS* 10(2) 1987.]

A related set of issues concern the "Hebbian paradigm": How does it work? What kind of biologically possible Hebbian training methods produce attractor representations of stimulus classes? Does training go on during testing? If so, what prevents previously formed attractors from being destroyed; if not, what shuts off training? What kind of Hebbian algorithms produce the temporal correlations between attractors representing different stimuli? How does the order of presentation of stimuli affect the formation of attractors? Do simultaneously active attractors influence each other?

Hebb postulated assemblies of assemblies, corresponding to higher order concepts, in which a whole cell assembly is considered a single unit. These send signals to each other, perhaps merely by overlapping. Do similar reverberations take place in these more abstract units? High-order Hebbian assemblies have been invoked in a recent philosophy book, in order to provide a physical substrate for mathematical abstractions (Maddy 1993; reviewed in Hirsch 1995).

Neural network researchers pays lip service to Hebb's notion of a Hebbian cell assembly. But they usually ignore the rich cellular structure, preferring instead to consider an assembly as a single unit with a single input-output map. It is refreshing to see cell assemblies taken seriously by Amit and his school.

Additional tests of Amit's attractor neural networks

Ralph E. Hoffman

Department of Psychiatry, Yale University School of Medicine, Box 208038, New Haven, CT 06520-8038. **hoffman@biomed.med.yale.edu**

Abstract: Further tests of Amit's model are indicated. One strategy is to use the apparent coding sparseness of the model to make predictions about coding sparseness in Miyashita's network. A second approach is to use memory overload to induce false positive responses in modules and biological systems. In closing, the importance of temporal coding and timing requirements in developing biologically plausible attractor networks is mentioned.

Amit describes a modification of attractor neural network (ANN) simulations which has interesting properties, namely, that neuronal responses are correlated if otherwise unrelated stimuli are learned in fixed order. Response correlations of simulated neurons seem to fall off at roughly the same rate relative to their sequential association when compared to neurons recorded by Miyashita. The critical parallel findings are that for both the ANN simulation and monkey neurons correlations fall to chance levels at about the same serial position number. This is a "surprise" finding that makes the model worthy of careful attention. I am less clear, however, about whether the fall-off correlation rate was "strongly predicted" by the model or whether these results were obtained by *post hoc* tweaking of the various parameters that determine the functioning of the model. Could different but equally plausible parameters (governing, for example, neuronal response functions) yield correlation fall-offs to chance levels at a serial position number of two rather than 5–6 (as indicated in Fig. 4)?

It appears from Figure 5 that memories stored in Amit's ANN were sparsely coded. Brisk responses were obtained for only 2/100 stored memories for the particular neuron displayed. In the now classic Hopfield (1982) model, all neurons in the network participated in the representation of particular gestalts (with neuronal deactivation coding for as much information as neuronal activation). Perhaps sparse coding is what allowed neurons in Amit's network to code information at less than saturation firing levels. Can the relative sparseness of coding structure for the latter be used to get some handle on relative sparseness of codings in Miyashita's biological networks? For example, in his 1988 study, only 57 of 206 neurons demonstrated shape-selective delay discharge (Miyashita 1988). This 25% "hit rate" becomes more understandable if only a relatively small fraction of neurons in anterior ventral temporal cortex is activated by a particular shape.

On a related note, what was the memory capacity of Amit's network? The standard memory capacity for Hopfield's original model was roughly 10% of the total number of neurons. On this basis it would be predicted that the memory capacity of a module of the sort described in section 2 would be 10% of 10^5 or 10^4. That, alas, is not a great deal of memory if the module we are talking about has to store and recall all the key visual patterns in the animal's phenomenal world. My speculation is that sparse coding, if implemented in an enlightened fashion, could push up the memory threshold of artificial systems, thereby increasing their biological plausibility. Nonetheless, at some point memory overload could occur. In this case, Gestalts might become blended together and, on a behavioral level, errors of commission could arise. It would be extremely interesting to see whether memory overload imposed on Amit's ANNs would produce certain types of response errors which could then be replicated in Miyashita-type neurophysiological experiments. For example, it might be the case that (over)training on excessively long sequences of patterns yields an increased frequency of commission errors by both artificial network and monkey. If my speculations are correct, then a second "choice" stimulus prone to producing false positive responses should be coded by a particularly large population of coactivated neurons derived from a blend of patterns induced by memory overload. This would constitute another "surprise behavior" of artificial ANNs that might be sought in Miyashita-type neural modules to test for attractor-like properties.

My major reservation regarding ANNs is that real-world information processing does not involve a series of frozen images but rather a continuous, nonstop flow of sensation. The timescales involved may be insufficient to allow neurobiological networks to settle into well-delineated attractors (Bialek & Rieke 1992). For example, some preprocessing decisions in the mammalian visual cortex seem necessary after only 2–3 spikes have been induced by the stimulus (ibid). This seems hardly sufficient to allow attractor dynamics to emerge. Observations such as these have caused some researchers to investigate whether patterns of spikes (or interspike intervals) over time can serve as a functional code (ibid). Such coding systems could be much more time-efficient compared to the sustained activation levels required by binary or gray scale coding systems (see, for instance, Hopfield 1984) of "traditional" ANNs.

To retain an attractor orientation for temporally continuous perceptual processes, the time required by a network to dislodge itself from an attractor (and the mechanisms for doing so) needs to be carefully considered. This issue has been usefully reviewed by Freeman and Skarda (1990), who suggest that perception is broken up into momentary "frames" which are biologically organized so that "the down-time between perceptual frames . . . is minimized . . . and the perceptual process may move with the seeming smoothness of a video" (p. 168). ANNs need to achieve this flexibility in order to be useful models of real-world perceptual processes. A good first step is Amit's examination of how temporal sequences of gestalts might be learned by networks;

moreover, ANNs, after training, need to be able to flexibly flow from one attractor to the next in response to a continuous flow of input information. Whether the output code of the network is expressed as an anatomic distribution of firing rates or as an anatomic distribution of evolving spike patterns (which does not depend primarily on averaged rates) remains an important unanswered question. If the latter is the case, then the Hebb rule, as a neural learning algorithm, will need an update.

Hebb's accomplishments misunderstood

Michael Hucka,[a] Mark Weaver,[b] and Stephen Kaplan[c]

[a]Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109; [b]Corvus Development, Inc., 2088 Georgetown Blvd., Ann Arbor, MI 48105; [c]Department of Psychology, and Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109. **michael.hucka@umich.edu; mark.weaver@um.cc.umich.edu; stephen.kaplan@um.cc.umich.edu**

Abstract: Amit's efforts to provide stronger theoretical and empirical support for Hebb's cell-assembly concept is admirable, but we have serious reservations about the perspective presented in the target article. For Hebb, the cell assembly was a building block; by contrast, the framework proposed here eschews the need to fit the assembly into a broader picture of its function.

Amit's work represents a significant achievement in understanding and modeling the dynamics within a Hebbian cell assembly. Unfortunately, the target article also reveals a flawed perspective that we believe leaves Amit poorly positioned to exploit and build upon his results. This flawed perspective manifests itself in two ways. First, he fails to make contact with both empirical evidence and related models which could serve to support and inform his own efforts. (Indeed, there is a danger that the reader will be left with the impression that the cell assembly is only now being rescued from obscurity some 45 years after Hebb proposed it.) Of greater concern, however, is the fact that these failures to make contact are not due to oversight; rather, they follow from a principled approach to cognitive modeling that involves deliberate avoidance of a broad cognitive theory. We argue that this is both strategically flawed and, ironically, in direct conflict with Hebb's perspective – Amit's position is conceptually much closer to the behaviorist position.

1. A broader sampling of empirical evidence is called for. The research by Miyashita et al. (Miyashita 1988; Miyashita & Chang 1988; Sakai & Miyashita 1991) provides exciting evidence for cell-assembly structures in the brain. However, in reading Amit's target article, one may get the impression that it is practically the only evidence. On the contrary, the research adds to a substantial existing body of empirical results on the topic. Direct evidence for maintained neural activity goes back to at least the 1950s (Burns 1951). Freeman has long studied EEG activity in the olfactory systems of rabbits and has developed an attractor model based on the results (Freeman 1975). Evidence for reverberatory activity similar to Miyashita's can be found in several lines of work (e.g., Goldman-Rakic 1990; Miller et al. 1993). For example, Goldman-Rakic has studied working memory in the prefrontal cortex of monkeys and found neurons that maintain their activity for several seconds during the delay period (but not the stimulus presentation) of a delayed-response visual task. More recently, Laurent and Davidowitz (1994) reported detailed evidence for assembly-based representations of odor information in the olfactory system of locusts. They proposed that "odor quality is encoded not only by an assembly of synchronously oscillating neurons but by a particular succession of different, but overlapping, oscillating assemblies" (p. 1874).

2. A broader sampling of theoretical issues is called for. Amit is surely right that Hebb's cell-assembly construct has received less attention than his learning postulate. As Amit notes, the most widely studied models of the past decade have been feed-forward systems – networks which are incompatible with cell assembly theory because they are fundamentally incapable of supporting reverberation. He is not the first to note their limitations; indeed, feed-forward connectionist models have been accused of constituting a behaviorist revival (Lachter & Bever 1988; Pinker & Prince 1988; also see Kaplan et al. 1990). Nevertheless, the importance of the cell-assembly concept and of Hebb's "processing cut," as Amit calls it, has not been lost on researchers. There are in fact two classes of connectionist models compatible with the cell assembly concept, and a large number of examples in each class. The first class includes those such as Hopfield's and Amit's that cast internal representations as attractors; representative examples include work by Anderson et al. (1977), Freeman (1975), Kanerva (1988), and Hinton & Sejnowski (1986). The second class is closer to Hebb's original conception: cell assemblies as distinct subpopulations of neurons. These cell-assembly analogues have been variously termed "classification couples" (Edelman 1987), "recognition codes" (Grossberg 1987), "object representations" (Kaplan & Kaplan 1982), and "cell assemblies" (Braitenberg 1984; Palm 1982).

Amit's attractor neural network, of which we see only a glimpse in the target article, is an important advance over many existing models in its coverage of biologically relevant details and the depth of its theoretical analysis. But in the context of a proposal for understanding cognitive phenomena, it must be kept in mind that the model's scope is quite limited. If one is going to propose that such network "modules" are a fundamental component of cognitive function, one must be prepared to explain something about how they work together to give rise to more complex phenomena – an issue that *has* been addressed in other models (e.g., Edelman 1987; Grossberg 1987; Kaplan et al. 1990) and indeed in Hebb's own theory.

3. Cell-assemblies and cognitive theory. We believe that the preceding issues are symptoms of a deeper problem. In his conclusion, Amit suggests that although cell-assemblies may "even suggest a substrate for psychology itself" (sect. 9, para. 5) constructing such a theoretical framework is a temptation we should resist, lest we get too far ahead of ourselves. He concludes:

> The lessons learned from these experiments include the one which advises restraint. . . . Our imagination concerning brain computation is still too much constrained by formal mathematics, by computer languages, and by artificial intelligence. . . . It is most likely that attending a while longer to the details of the contact between modeling and experiment will keep open options which a premature harvest of speculation would foreclose. [sect. 9, last para.)

But Hebb's (1949) *Organization of behavior* must rank as one of the great "premature harvests of speculation" of our time. Nearly fifty years later, Hebb's speculations continue to influence the course of experimental and theoretical brain research, as the present enterprise, for example, plainly demonstrates. In contrast, the circumscribed, bottom-up approach and avoidance of cognitive theory that Amit argues for is actually very similar to the position taken by the behaviorists. It is interesting to note just how many of the constraints of the collective imagination Hebb was able to transcend; in the late 1940s, psychology was dominated by behaviorism and neuroscientists could offer direct evidence neither for synaptic learning nor reverberation. Perhaps even more important than his particular theoretical contributions is the example Hebb set for how to conduct a dialogue between cognitive theory and experimental neuroscience, and this Amit seems to have missed entirely.

This principled avoidance of a larger theoretical framework is not a purely philosophical problem – it has practical implications for precisely the kind of research program that Amit favors. In particular, if reverberation is to form the basis of a cognitive theory, the cell assembly must serve as a building block in a larger, more complex system – a system that will have its own emergent dynamics that will influence the individual assemblies. If one lacks a notion of the kind of environment in which cell assemblies

function, one is unlikely to be able to explain fully their operation or explore their potential. For this reason, we fear that Amit's results – technically impressive as they are – will not easily prove to be extendable.

The functional meaning of reverberations for sensoric and contextual encoding

Wolfgang Klimesch

Department of Physiological Psychology, University of Salzburg, 5020 Salzburg, Austria. **klimesch@edvz.sbg.ac.at**

Abstract: Amit argues that the local neuronal spike rate that persists (reverberating) in the absence of the eliciting stimulus represents the code of the eliciting stimulus. Based on the general argument that the inferred functional meaning of reverberation depends in part on the type of representational assumptions, reverberations may only be important for the encoding of contextual information.

The crucial starting point for the following arguments is a seemingly plausible assumption of Amit's. He proceeds from the idea that a stimulus activates a particular subset of neurons and that (after the stimulus is turned off) the activated cell assembly may either decay rapidly or reverberate for some time. Only if a stimulus leads to a state of reverberation can or will the stimulus be perceived. In other words, the cell assembly (or assemblies) representing the stimulus information must be put in a state of reverberation that outlasts the exposure of the stimulus. Furthermore, Amit assumes that the activation of the cell assembly serves to "tag" this activated (passive) memory.

Even in considering Amit's reference to working memory (in para. 5 of sect. 2), I see two possible interpretations of what reverberation may mean: (1) Reverberation may just mean that a particular cell assembly is selected (tagged) from passive memory or (b) reverberation may reflect the activation of additional or other neuronal circuits that represent specific contextual information about the encoded stimulus. In terms of cognitive psychology, "passive memory" is closely related to semantic (long-term) memory, whereas "tag" captures the essential meaning of episodic memory (cf. Tulving 1984).

To clarify this point it may be helpful to distinguish between the following two meanings of encoding. The encoding of sensory information (as a process of recognizing a presented stimulus) aims at the semantic understanding of perceived information. Long-term memory (LTM) holds the information which is essential for this encoding process. Within the framework of working memory, encoding means the creation of a new code that primarily comprises episodic information. Thus, reverberation may in principal refer either to semantic LTM or to episodic memory.

With respect to the encoding of sensory (semantic) information, interesting findings were reported by Gray and Singer (1987) together with other researchers at the Frankfurt MPI (see e.g., Engel et al. 1992). They have provided convincing evidence that a visual code, established through a perceptual process, can be described as a cell assembly which responds with a synchronous oscillatory discharge pattern within a broad frequency range of about 30 to 70 Hz which is termed the gamma band. They assume that the synchronous oscillatory firing pattern of distributed cortical cells reflects a stage of cortical integration in the sense that the information provided by different feature detectors is integrated into a single visual code. This assumption is substantiated by the important finding that even widely distributed but synchronously oscillating cell assemblies fire with zero phase lag. Feedback loops, connecting different cell groups of the cortex are obviously the means which enable this surprising ability.

In contrast to this approach, Hebb's conception has the disadvantage that in a particular cortical region and within a given time span, only a single code or feature can be activated, because the enhanced firing rate is the only cue which allows one to distinguish the relevant code from irrelevant information. However, during a search process in LTM, a huge variety of codes will be activated at the same time and possibly in the same brain region. Thus, different and topographically overlapping cell assemblies will be activated at the same time. Consequently, it will be impossible to distinguish between different codes. In trying to avoid this problem, one may instead assume that assemblies can be functionally defined by a state of synchronous firing of cortical neurons, rather than by an enhanced average firing rate (for a more detailed discussion see, e.g., Klimesch 1995).

In summarizing my arguments, a simple conclusion can be drawn. Reverberations (in the sense of a persisting neuronal discharge pattern) do not seem necessary for sensory (semantic) encoding but may be of substantial importance for the encoding of contextual information. The encoding of contextual – or any other type of new – information may very well depend on reverberations which in this case may reflect a process of consolidation. Because Amit's arguments are based on the delay period separating the presentation of two already learned stimuli, the reverberating activity observed during this period may very well be related to the encoding of contextual information.

An evolutionary perspective on Hebb's reverberatory representations

David C. Krakauer[a] and Alasdair I. Houston[b]

[a]*BBSRC NERC Ecology & Behaviour Group, Department of Zoology, University of Oxford, Oxford OX1 3PS and* [b]*School of Biological Sciences, University of Bristol, Bristol BS8 1UG, United Kingdom.*
krakauer@vax.ox.ac.uk

Abstract: Hebbian mechanisms are justified according to their functional utility in an evolutionary sense. The selective advantage of correlating content-contingent stimuli reflects the putative common cause of temporally or spatially contiguous inputs. The selective consequences of such correlations are discussed by using examples from the evolution of signal form in sexual selection and model-mimic coevolution. We suggest that evolutionary justification might be considered in addition to neurophysiology plausibility when constructing representational models.

This seems too apposite an opportunity for us to fail to quote from William James 1902 (cited in James 1987) where he writes, "There are innumerable kinds of connexion that special things have with other special things; and the ensemble of any one of these connexions forms one sort of system by which things are conjoined." Amit has eloquently demonstrated and justified the recent interest biologically minded physicists have shown in the problem of forming connections between correlated but contingent stimuli. Although we feel unqualified to comment on the physics or neurophysiology, we would like to present a series of points which we hope will demonstrate the evolutionary character of the problem. Our approach will be, first, to suggest that Hebbian learning, with its ability to construct temporal correlations of the Miyashita type (Miyashita & Chang 1988), is adaptively very sensible. We will then present a few consequences of such correlations which naturally follow from such a mechanism; these will be of some evolutionary interest.

The vast literature on connectionist modelling that has emerged in recent years presents a bewildering suite of alternative learning rules for training the weights of a model network. The usual justifications for these training rules are either engineering expediency, or neurophysiological plausibility. The theoretical demonstration by Amit and his colleagues that Hebbian learning rules (correlations between pairs of neurons during the presentation of a stimulus) can induce correlations between internal representations, deserve a further justification drawn from a consideration of evolution or functional utility. The argument is essentially one that maintains that correlating small *sensu* Miyashita, or as Amit puts it

"correlations form between attractors representing semantically meaningless, uncorrelated stimuli" is a selectively advantageous process. Stimuli uncorrelated in content but correlated in time or space are often derived from the same cause. The philosopher Reichenbach (1956) formulated this observation as the principle of the common cause. A mechanism that allows such correlations to emerge provides a nondeductive means of forming categorical representations of the external world. Consider the case of a predator who provides both an immediate and unique visual impression and also a prior auditory indication of its presence. The ability of the sound of the predator to prime visual recognition of the predator is a selectively advantageous trait: the category of "predator" may be wired so as to spontaneously elicit an escape response. The Hebbian learning rule is therefore thought of as a mechanism which acts to construct unitary representations of adaptively important features in the external world. This of course underpins the psychological principle of association holding between constituent events.

Given such a correlating mechanism, how might it be further exploited during the course of evolution? We shall dwell briefly on multi-component displays used during sexual selection, and the evolution of mimetic morphologies.

Males from many species in a large number of animal taxa are adorned with brightly coloured plumage and engaged in a range of behaviours intended to attract the attention and subsequent sexual favours of females (Anderson 1994; Darwin 1871). For example, the male anuran, *Bufo calamita* attracts females through a combination of a large body size and a high amplitude, high pulse rate, low pitch call (Arak 1988). The passerine *Geospiza conirostis* attracts females using a combination of song type and body colour (Grant 1985). Why should males invest in numerous displays when each reveals ostensibly the same piece of information? The formation of correlated representations of a single object through uncorrelated stimuli could provide critical information when one modality is partially occluded. For example, poor weather conditions may diminish visibility but leave sound unaffected. In such a context it is still important for females to choose the best males. Priming effects would allow the female to reach a more informed decision through supplemental auditory information as opposed single visual signalling channel. A further observation on signal form has recently been made by Enquist and Arak (1993) concerning spurious states in artificial neural networks (analogous to undesired basins of attraction in attractor neural networks). Enquist and Arak argue that spurious states provide avenues through which low quality males might exploit the preference behaviours of females. In other words, there are signals to which networks are prone to respond that lie outside the training set experienced during the course of evolution, consequently low quality males using such signals by chance experience disproportionately high fitness. An exploration of the dynamics of such exploitation has allowed us to investigate the diversity of signal forms (Krakauer & Johnstone 1995).

A number of poisonous species of animal have evolved characteristic patterns of colouration (aposematic colouration) which act to warn predators of their unpalatability. Other species lacking poisons (Batesian mimics) have evolved similar patterns and consequently benefit by proxy from this means of deterrence. This example is intended to provide a possible behavioural analogue to the phenomena Amit refers to as false alarms. The "test stimulus," here the mimic, is correlated with items in the "subset," here the aposematic species. Consequently mimics elicit representations reserved for aposematics. As the frequency of mimics increases, the predators learn to ignore the signal through a new association (the warning pattern becomes paired with palatability) and the deterrence effect declines. Hence the positive adaptive value of generalisation by a receiver for a given class of signals leads inexorably to the possibility for exploitation. Thus for an evolutionist the Hebbian assembly is a double edged sword. To paraphrase the James quote (1902) of our opening gambit, *special things* are not always connected to the *other things* that an animal would choose.

Distributed cell assemblies and detailed cell models

Anders Lansner and Erik Fransén

SANS (Studies of Artificial Neural Systems), Department of Numerical Analysis and Computing Science, Kungl. Tekn. Högskolan, S-100 44 Stockholm, Sweden. **ala@sans.kth.se**

Abstract: Hebbian cell-assembly theory and attractor networks are good starting points for modeling cortical processing. Detailed cell models can be useful in understanding the dynamics of attractor networks. Cell assemblies are likely to be distributed, with the cortical column as the local processing unit. Synaptic memory may be dominant in all but the first couple of seconds.

Cell assemblies, attractor networks, and cortical function. Amit's target article highlights the striking analogies between Hebb's cell-assembly theory and modern attractor network models. It relates them to cortical function as revealed by experiments on awake behaving monkeys. We fully support this general scheme. In fact, the case can be made even stronger. The pattern reconstruction and prototypal operation of attractor networks have been suggested to underlie Gestalt perception, language production, and memory phenomena (See, e.g., Quinlan 1991). Closely related contemporary work includes theoretical and anatomical investigations by Braitenberg and Palm, EEG-models as reviewed by Wright and Kydd, and studies of dynamic associative memory in the olfactory cortex pioneered by Freeman (see Lansner & Liljenström 1994).

Significance of neuronal properties. Hebb's ideas were in fact subject to early computer simulations (MacGregor & McMullen 1978; Rochester et al. 1956). However, these came out essentially negative. More recent simulations suggest that motor neurons (as were used in the previous models) do not readily support reverberatory activity whereas cortical pyramidal cells do (Lansner 1982; Lansner & Fransén 1992). This is mostly due to differences in afterhyperpolarization properties.

In addition, these simulations using compartmentalized Hodgkin-Huxley type model neurons showed that (1) Hebbian synapses can be used to store memories in these networks; (2) the time to reach an attractor is around 100 milliseconds, that is, close to experimental perceptual reaction times; (3) fast spiking local inhibitory interneurons are well suited for supplying lateral inhibition.

Amit proposes that specific noise and unstructured inhibition is behind low rate activity. However, our simulations show that low firing rates can also be obtained in an assembly of pyramidal cells without any inhibition or specific noise, provided we include saturating excitatory synapses (kainate/AMPA and NMDA, Fransén & Lansner 1994). Thus, the role of inhibition here is still an open question.

Local vs. distributed cell assemblies. The most important point where our view differs from that expressed by Amit is in how theory is mapped onto living cortex. Amit puts several assemblies together in a spatially localized module. Our hypothesis is that cell assemblies are large aggregates of interconnected columns that are spread over the cortex, perhaps over multiple areas. We have done simulations of distributed cell assemblies with average conduction delays of up to at least 10 milliseconds without deterioration of function (Fransén et al. 1992). In addition, we regard the cortical (mini)column as a processing unit, and not the individual cortical neuron as Amit does. This allows for sparse long-range connectivity and a functionally symmetric connectivity without cell-to-cell symmetry (Lansner & Fransén 1994). This makes the connectivity of the model more like that of real cortex. Of course, neither model reflects the true complexity of cortical functional architecture with layers, blobs, hypercolumns, and so on.

In addition, we see no reason to assume that cell assemblies can only exist in association cortex. They are likely to be found also in the sensory and motor cortex, which also has the necessary horizontal connectivity and synaptic plasticity. There might, how-

ever, be differences from region to region with respect to stimulus dependence, temporal properties etc. For example, cell assemblies in sensory areas might support pattern completion without reverberations during the stimulus interval (Fransén & Lansner 1994). A possible way to distinguish the two hypotheses would be to lesion a local module (e.g., one of the type studied in AVT), and assess how this affects performance. If marked effects are seen, locality is more likely, whereas little or no effect would be suggestive of a distributed representation.

Role of reverberations. We feel that Amit overemphasizes the role of independent, long-lasting reverberations for memory. As an alternative, we propose that "passive" synaptic memory is dominant in all but the first hundred milliseconds, up to perhaps a minute. Since an active assembly may extend over a large part of the brain and suppress other activity, synaptic memory fits our hypothesis of distributedness better than assuming many simultaneous reverberations. Synaptic connections that are enhanced and weakened on a fast time scale may well support different forms of short and intermediate term memory. The sign of a "tagged" assembly could then be a "trace" of potentiated synapses. Diffuse activation would trigger such an assembly relatively easier than others. Other synapses could maintain the temporal connections that produce the temporal correlations as discussed by Amit.

Conclusions. It is extremely important to formulate a theory of cortical associative memory and we can see no better starting point than the one taken by Amit. But when mapping the cell assembly theory to real neuronal networks we should be careful to leave open alternative interpretations as long as they are compatible with experimental data.

Attractors – don't get sucked in

Peter M. Milner

Department of Psychology, McGill University, Montreal, Que., Canada, H3A 1B1. **ps64@musica.mcgill.ca**

Abstract: Every immediate memory is unique; it is therefore unlikely to consist of an attractor or even a combination of attractors. In the present state of knowledge about the chemistry of synaptic transmission, there is no reason to look beyond neurons that directly receive sensory afferents for the afterdischarges that correspond to active memories.

> In speculating, one can at least be specific enough so that, when further anatomical and physiological information is made available to the psychologist, the errors of earlier speculation such as this will be apparent at once, and the necessary changes clearly indicated. D. O. Hebb (1949, p. 80)

Mid-century was a period of upheaval for both psychology and neurophysiology. Until then it was almost universally believed that synaptic transmission was effected by the passage of an electric current from an axon terminal to a dendrite. As a consequence it was considered most likely that learning required the growth of synaptic knobs, and that, because of the brevity of the electric pulses, they would summate only if they arrived almost simultaneously. Furthermore, it was difficult to conceive of an electrical synapse that could inhibit. These misconceptions led Hebb astray in his speculation about the cell assembly.

On behavioral and introspective grounds it is clear that stimuli establish persistent traces almost instantly, too quickly to be explained by synaptic growth. A popular candidate for maintaining memory during the interval was reverberation, the circulation of impulses round closed loops (Hilgard & Marquis 1940; Müller & Pilzecker 1900). It was not an elegant solution. Hebb (1949, p. 62) believed, for example, that a reverberatory trace might account for the ability to repeat a series of digits immediately after hearing them, but has anyone seriously attempted to show how a purely dynamic trace can store (and play back on demand) a number like 363363363, which the brain does with relative ease?

Even in 1949, when I first read *The Organization of Behavior*, I found it difficult to believe that activity started in so complex and busy a structure as the cerebral cortex would complete even one cycle of reverberation without some chemical marker to keep it on track. Now that we know synaptic transmission involves a series of extremely rapid (and in some cases long-lasting) conformational changes at the molecular level (Kennedy 1989), the sort of dynamic storage that Hebb considered necessary fifty years ago seems less plausible than ever. It is surely time to heed the words quoted at the head of this commentary and consider what the new information about synaptic transmission might mean for the cell assembly. Amit's remark (sect. 5, para. 3) that "the selective activity distribution can persist for as long as 16 seconds, in a rather noisy environment" strengthens my conviction that the persistence must be attributed to an effect local to the observed neurons, or to neurons directly coupled to them; definitely not to feed-back around a cortical loop.

As Amit makes clear, arguments supporting an active trace are irrefutable, but reverberation round closed loops is not the only way, and in my opinion certainly not the best way, of attaining it. An alternative that I tentatively suggested many years ago (Milner 1957) is that cells primed by facilitatory input are repeatedly fired by nonspecific afferents. For example, neurons sensitized by sensory input or by the activity of other cortical neurons may be fired for a time by input from the widespread network that according to Llinás and Ribary (1993) continuously sweeps the awake cortex at a frequency of about 40 Hz.

A large portion of Amit's target article is concerned with attractor networks, which he sees as relating Hebb's reverberations to computer models of the Hopfield net type and to the experiments of Miyashita et al. I once thought that the attractor model provided a good explanation for some aspects of stimulus equivalence (triggers from similar stimuli set off identical reverberations) and in fact the first computer simulations of the cell assembly (Rochester et al. 1956), which were based on my MkII version (Milner 1957), were of the attractor type. But since then I have had second thoughts.

In a room full of people I know immediately which one is my wife. Leaving a restaurant I know whether I have the right coat. If someone reads me a word to translate, it is not only the word that persists in my head, I also remember the voice, which is not usually replaced by the attractor for the familiar voice it most resembles. It seems unlikely that we have separate attractors for every object and characteristic that we can discriminate. If the attractor model promotes categorization then we need to know why it does not seriously interfere with discrimination.

This dilemma is pertinent to Amit's claim, made in the penultimate paragraph of section 3, that attractor representation is the only dynamic that can distinguish naturally between familiar and unfamiliar stimuli. Familiarity should not be confused with classifiability. That I recognize an object as a dog does not mean that I am acquainted with it. The problem of recognition is more complicated than Amit makes it seem. I can indicate which of two objects I have seen most recently whether the objects were initially familiar, unfamiliar but classifiable, or, within limits, utterly baffling to me. Moreover, it is not enough that familiar objects produce a neural activity different from other objects; no matter what objects are being compared, the relevant difference must activate a common output, a neural representation of the concept "familiar," "recently seen," or whatever (Milner 1989).

I do not have space here to discuss the experiments of Miyashita and his colleagues (1988a; 1988b; 1991; 1993) that Amit regards as conclusive evidence for the existence of self-maintaining reverberations in attractor neural nets. It is sufficient to say that, taking into consideration Hebb's postulate that associations occur between neurons that are simultaneously active, the data are even more consistent with the postulate advanced here that input

produces an immediate and persistent threshold change at learning synapses.

Another ANN model for the Miyashita experiments

Masahiko Morita

Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba, Ibaraki 305, Japan. **mor@is.tsukuba.ac.jp**

Abstract: The Miyashita experiments are very interesting and the results should be examined from a viewpoint of attractor dynamics. Amit's target article shows a path toward realistic modeling by artificial neural networks (ANN), but it is not necessarily the only one. I introduce another model that can explain a substantial part of the empirical observations and makes an interesting prediction. This model consists of such units that have nonmonotonic input-output characteristics with local inhibition neurons.

I have been studying associative memory and examined the Miyashita experiments from the viewpoint of the dynamics of associative neural networks. Discussions in the target article are important and quite reasonable; I agree, except for the following point.

It is true that modification of conventional ANN models is required to explain the results of the Miyashita experiments and the article shows a way to do realistic modeling. However, it seems questionable that the elaborate modification of the single neuron dynamics is really necessary. In the following, I will describe a model which might give a simpler explanation of some of the empirical observations.

Amit points out that all previous ANN models produced a bimodal distribution of rates in an attractor contrary to the empirical observation. I independently noticed and examined this point and found that this problem can be solved by modifying the analog Hopfield model only a little, that is, by adopting a neuronal model whose output is a nonmonotonic function of the input (Fig. 1). This small modification causes a critical change in the network dynamics: the distribution of output values in an attractor becomes broad because the dispersion of inputs is reflected in the output distribution without saturation. The nonmonotonic model also enlarges the memory capacity of autoassociative neural networks and attractivity of the attractors (Morita 1993).

Another important property is that this model can have attractors of a "line type" whereas the original Hopfield model has only point attractors isolated mutually (Morita 1994). This enables us to store sequential patterns in the network in a simple and natural manner.

Of course the nonmonotonic neuron itself is not realistic. However, a local circuit consisting of a few neurons can realize

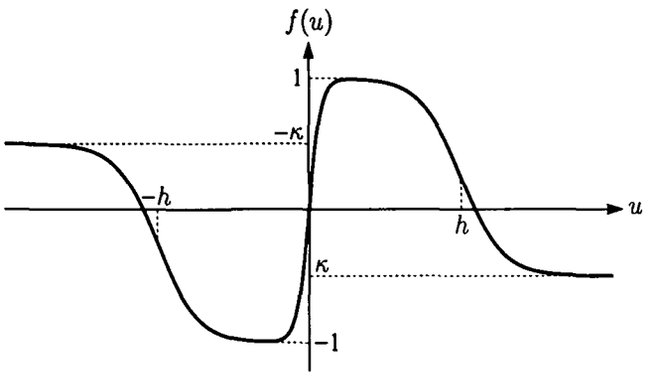

Figure 1 (Morita). Nonmonotonic transfer function. Conventionally, monotonically increasing sigmoid function has been used for the transfer function.

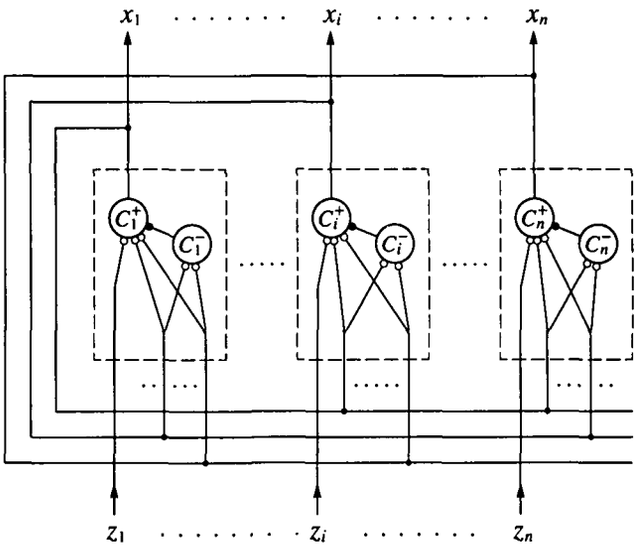

Figure 2 (Morita). Structure of the model. The part surrounded by broken lines represents a unit.

nonmonotonic input-output characteristics. The simplest example is shown in Figure 2 (Morita 1992).

In this model, an output neuron C_i^+ and an inhibitory neuron C_i^- compose a unit: C_i^- receives signals from other units in common with C_i^+ and send a strong inhibitory signal to C_i^+ when the total in put is large. Thus the output of the unit is a nonmonotonic function of the total input as shown in Figure 3.

The behavior of this model is consistent with the observations in the Miyashita experiments. Moreover, an interesting prediction is derived: the neuron which responds to two different figures will exhibit only a very weak response if both figures are presented at the same time. This may seem strange, but it necessarily occurs in this model, since such a neuron receives such large input signals from the neurons responding to either of the figures that the output becomes small because of the nonmonotonic property. After making this prediction, I heard from Professor Miyashita that such a phenomenon was actually observed in his experiments.

Although this model involves some conflicts with the models in the target article (for example, with the first prediction in sect. 8), it is useless to discuss now which model is correct. In a sense, every model that can make a substantial contribution to brain research is correct and should be developed. I expect that Professor Amit and

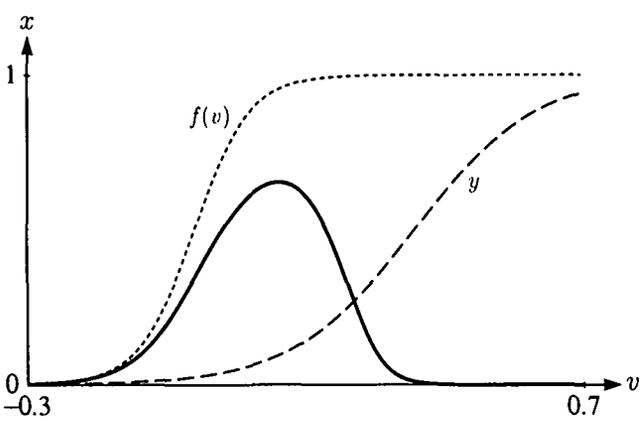

Figure 3 (Morita). Input-output characteristics of a unit (solid line). The abscissa is the total input v to the unit, and the ordinate is the output x. Without the inhibition neuron, x is a sigmoid function of v (dotted line), but it decreases with an increase of the output y (broken line) of the inhibition neuron.

his models will continue to make a contribution. I too will continue to further improve my model, being stimulated by the target article.

The problems of cognitive dynamical models

Jean Petitot

EHESS, Mathematical Center, 54 bd. Raspail, 75 006, Paris, France and CREA, Ecole Polytechnique, 1 rue Descartes, 75 005, Paris, France. **petitot@poly.polytechnique.fr**

Abstract: Amit's "Attractor Neural Network" perspective on cognition raises difficult technical problems already met by prior dynamical models. This commentary sketches briefly some of them concerning the internal topological structure of attractors, the constituency problem, the possibility of activating simultaneously several attractors, and the different kinds of dynamical structures one can use to model brain activity: point attractors, strange attractors, synchronized arrays of oscillators, synfire chains, and so forth.

The main idea of Amit's article is to identify the psychological concept of Hebbian reverberation with the dynamical concept of attractor and to ground a computational theory of mental representations on this basis. This ANN (Attractor Neural Network) perspective on *cognition* raises difficult technical problems.

1. A historical remark. In the context of a physicalist approach of connectionist networks Amit has been one of the first to emphasize the interest of a true dynamical perspective, but prior mathematical work already exists. As far as I know, it was Zeeman (1965; 1976) who in the late sixties introduced the cogent and seminal idea that one could use dynamics to bridge the gap between the small scale neural level and the large scale psychological one. According to him, mental contents could be modeled by attractors of neurally implemented dynamical systems and the temporal flux of mental representations by sequences of bifurcations of attractors.

The main limitation of these early dynamical models compared to current ones was that the effective neural dynamics were unknown. Deep theorems showing that there exist *universal* prototypes ("normal forms") for the relevant dynamical structures (e.g., universal unfolding of singularities of energy functions) nevertheless made it possible to work out dynamical cognitive models (e.g., for categorical perception or image processing; see Petitot 1989; 1995). These early models already showed the kinds of technical problems facing a dynamical cognitive theory. I wish to stress some of them here.

2. The internal structure of attractors. In an ANN perspective, what can be the internal structure of attractors? In the case of symmetric weights (Hopfield model) there exists an energy function and attractors are therefore point attractors. But in the case of asymmetric weights one can get topologically complex attractors; and routes towards chaos (such as the doubling period subharmonic cascade) are observable (see Sompolinsky et al. 1988, Dayon et al. 1993, or Renal & Rohwer 1990 work).

The fact that the topology of attractors is in general highly nontrivial could be essential for understanding the semantics of mental contents.

3. Attractor syntax and the constituency problem. If one identifies a mental representation with an attractor, then one must take up the challenge of modeling dynamically the syntactic constituent structures. In their replies to Smolensky's fundamental 1988 *BBS* paper "On the Proper Treatment of Connectionism," Fodor & Pylyshin (1988) and McLaughlin (1990) have shown dramatically that connectionism lacks any correct account of constituency and compositionality.

They were essentially right concerning a very weak (PDP) form of connectionism (see Petitot 1991), but this is no longer the case if one adopts a stronger form. Indeed, according to Thom (1980), it is possible to work out an "attractor syntax" using bifurcations (more precisely universal unfoldings of singularities) in an original way (see Petitot (1995). What is Amit's response to the constituency problem?

4. The simultaneous activation of several attractors. Another difficult problem concerns the simultaneous activation of several attractors. Indeed, a dynamical system can only be in a single asymptotic state at any time. If one uses attractors for explaining how several representations can be tagged by a stimulus and can be self-maintained in memory until further processing, then one faces a problem. One possible solution could be to activate attractors of a slow/fast dynamical system sequentially, but such an attractor-chaining is not satisfactory because in many cognitive tasks the co-activation must be done in parallel.

To tackle this difficulty, it seems that some sort of symbolic computing is unavoidable.

5. What kind of dynamics? My last point concerns the different kinds of dynamical structures one can use to model brain activity. Some solutions to the constituency problem (the so-called "binding problem") use results of experiments on cortical oscillations. Since the pioneering findings of Gray and Singer (see e.g., Engel et al. 1992) much work has shown that neural modules (e.g., orientation columns in the primary visual cortex) behave as oscillators and that their synchronization is stimulus-dependent and codes for the coherence of the stimuli. According to the "labeling hypothesis," the constituent structures of mental representations can be retrieved using as labels for the constituents the common phase of the synchronized oscillators they are implemented in. The problem of synchronizing of weakly coupled oscillators is a very difficult one which can be tackled only with sophisticated tools of statistical physics (Kuramoto and Nishikawa 1987: phase transitions, Daido 1990: renormalization group) or of qualitative dynamics (Kopell and Ermentrout 1990).

Even though these results are controversial (they can be significantly improved using pulse-coupled oscillators), they show that many kinds of dynamical structures can be relevant: point attractors, strange attractors, synchronized arrays of oscillators, and so forth.

It would also be interesting to see in what *exact* sense Miyashita's results confirm the ANN hypothesis. Indeed, similar experimental results can be interpreted in a different way. For example, Bienenstock (1994) uses the concept of *synfire chains*, that is, neural modules supporting wave-like patterns of activity. According to Abeles (1991), synfire chains can reverberate in different modes, depending on the context of their activation. They can also learn to recognize sequences of synchronized volleys and can dynamically bind with each other via synchronization. They might accordingly represent another major mechanism for local information processing in the cortex. What does Amit think of their links with ANN models?

Local or transcortical assemblies? Some evidence from cognitive neuroscience

Friedemann Pulvermüller and Hubert Preissl

Institut für Medizinische Psychologie und Verhaltensneurobiologie, Universität Tübingen, 72074 Tübingen, Germany. **pumue@tuebingen.de**

Abstract: Amit defines cell assemblies as *local cortical neuron populations* with strong internal connections. However, Hebb himself proposed that cell assemblies are distributed over different cortical areas (nonlocal or *transcortical assemblies*). We review evidence from cognitive neuroscience and neuropsychology supporting the assumption that cell assemblies are transcortical.

Translating cognitive terms into the language of neurobiology is certainly a worthwhile enterprise. Hebb's concept of cell assemblies has frequently proven useful in this endeavor. However, the term "cell assembly" has been used in various different ways

(Gerstein et al. 1989). Most authors use it to characterize a neuronal population (1) with strong reciprocal internal connections which can therefore be considered (2) a functional unit. Amit uses the term to refer to local neuron clusters (so-called modules) restricted to gray matter beneath 1 mm^2 of cortical surface. In contrast, Hebb himself has used the term to refer to neuronal populations widely distributed even over large cortical areas, such as various visual cortices (Hebb 1949, see for example, pp. 73 ff). Others have developed the Hebbian approach further, assuming that the whole cortex is a huge associative memory where neurons of diverse areas may strengthen their connections when frequently active at the same time (Braitenberg 1978; Braitenberg & Schüz 1991; Palm 1982; Singer 1994). At this point, it cannot be decided with certainty whether local assemblies or widely distributed transcortical assemblies exist in the cortex. However, according to our view, evidence available at present suggests that cortical assemblies are not restricted to a tiny voxel of cortical space.

Some evidence for transcortical assemblies comes from recent experiments in cognitive neuroscience. As Amit correctly points out, frequent perception of a stimulus can be assumed to lead to the formation of a cortical cell assembly, while new and uncommon stimuli are unlikely to have such brain-internal representations. Different neurophysiological responses to common and uncommon stimuli have not only been obtained from individual neurons and local neuron clusters, but also by using large-scale neurophysiological methods such as the electroencephalogram (EEG) and the magnetoencephalogram (MEG). Perhaps the most common stimuli perceived by human subjects are frequently-used words, while random letter combinations and pseudowords are very uncommon. It has been shown repeatedly that EEG and MEG responses to words and pseudowords reliably differ (Holcomb & Neville 1990; Pulvermüller et al. 1995). Differences can be found, for example, in high-frequency brain responses of the gamma-band (> 20 Hz), where pseudowords elicit less spectral power compared to words (Lutzenberger et al. 1994; Pulvermüller et al. 1994). Stronger gamma-band responses to words can be explained by assuming cell assemblies in which fast circulation of neural activity takes place. Word presentation activates such an assembly, while pseudoword presentation fails to "ignite" an assembly. It appears very unlikely that large-scale brain responses are generated by a small local neuron population (Birbaumer et al. 1990). Differential gamma-band responses in the EEG or MEG recorded from the human brain strongly suggest that cell assemblies underlying these responses are transcortical (Pulvermüller et al. 1994; Lutzenberger et al. 1995).

Another clue comes from behavioral experiments. If two copies of a meaningful word are simultaneously presented in the left and right visual half-fields, processing is speeded compared to unilateral presentation of only one copy of the word. This processing advantage after bilateral presentation (bilateral gain) cannot be observed for pseudowords (Mohr et al. 1994b). One possible conclusion from this result is that processing units representing words are distributed over both hemispheres. Summation processes in these assemblies allow for faster responses when the network is stimulated twice. These transcortical assemblies must be held together through fibers of the corpus callosum. Consistent with this assumption, the word-specific bilateral gain is absent in split-brain patients (Mohr et al. 1994a). A similar argument can be made for the representation of motor movements. Bilateral elbow movements can be carried out as fast as unilateral elbow movements, while unilateral finger movements are usually faster compared to bilateral finger movements (Anson & Bird 1993). Bilateral elbow movements may be fast because summation processes take place in transcortical assemblies held together via the callosum. Interestingly enough, there are no transcallosal connections between the left and right motor areas representing finger movements; hence no transcallosal summation processes can take place when bilateral finger movements are performed (Wickens et al. 1994). This can explain why bilateral finger movements are relatively slow. These behavioral experiments (bilateral gain for words but not for pseudowords, and not in split-brain patients; fast bilateral elbow movement, but slow bilateral finger movements) suggest that cortical processing units can even be distributed over both hemispheres.

In conclusion, there is evidence for transcortical assemblies coming from various empirical fields. One may ask whether there is also evidence supporting Amit's statement that assemblies should be local. To our knowledge, there is not. The only piece of evidence for local assemblies Amit refers to is the finding that, in a particular delayed response experiment, neurons in "a small part (about 1 mm^2) of anterior ventral temporal cortex" showed "stimulus selective persistent activity . . . during the delay period." Amit takes this as "convincing evidence for the local maintenance of a reverberation by the feed-back in the synaptic structure" (target article, sect. 5, para. 3). However, if stimulus-selective persistent activity is present in one cortical locus, this does not imply that such activity is absent elsewhere. Fuster (1994) who has investigated this issue for many years emphasizes that "while a monkey waits to perform a motor act in accord with a recent sensory stimulus, as in a delay task, *innumerable neurons in widespread areas of its neocortex* undergo sustained elevation of firing" (p. 243). This widespread activity in different cortical areas is likely to be caused by processing units held together through long-distance connections between pyramidal cells.

Amit is aware of the possible relevance of long-distance connections for local cortical processing. For example, he acknowledges that continuous feedback and feedforward projections between primary cortices and a local "module" may be relevant for processing in the latter (sect. 4 of target article, second last paragraph). Nevertheless, he fails to draw the inevitable conclusion that, in this case, processing units would be nonlocal, and that the maintenance of activity in such units would be due to local as well as long-distance connections. Taking into account recent data from cognitive neuroscience, a neuronal assembly should be defined as neuronal network consisting of several local neuron clusters held together by long-distance connections. This concept corresponds to Hebb's intuitions.

ACKNOWLEDGMENTS

For comments on an earlier version of this manuscript, we wish to thank Bettina Mohr. Work is by grants Pu 97/2-2 and Pu 97/5 from the Deutsche Forschungsgemeinschaft (DFG).

How to decide whether a neural representation is a cognitive concept?

Maartje E. J. Raijmakers and Peter C. M. Molenaar

Department of Psychology, University of Amsterdam, 1018 WB Amsterdam, The Netherlands. **op_raijmakers@macmail.psy.uva.nl**

Abstract: A distinction should be made between the formation of stimulus-driven associations and cognitive concepts. To test the learning mode of a neural network, we propose a simple and classic input-output test: the discrimination shift task. Feed-forward PDP models appear to form stimulus-driven associations. A Hopfield network should be extended to apply the test.

A major subject of Amit's target article concerns self-maintaining reverberations as neural representations or cognitive concepts. He argues that a neural representation of a stimulus cannot be active only during the presentation of the stimulus but also afterwards. Experiments of Miyashita et al. (Miyashita & Chang 1988; Miyashita 1988; Sakai & Miyashita 1991) show empirical evidence for reverberations as the internal code. In a Hopfield network with Hebbian learning dynamics, in contrast to feed-forward networks, the formation of attractors of spike rates that represent presented stimuli can be simulated. This kind of model, attractor networks, is proposed as a good and simple model of the formation and activation of neural representations or, as Amit says, cognitive

concepts. We agree with Amit that attractors are better candidates for neural representations of stimuli than the neural representations formed in feed-forward networks. Amit gives an excellent enumeration of arguments for this.

The point we wish to make in this commentary, however, is that neural representations are not cognitive concepts per se. There exists an extended philosophical discussion about the distinction between cognitive concepts and stimulus-driven associations (including stimulus-response associations, Reese 1989). We will describe a simple empirical test, which only concerns the input-output relation (of animals, humans, or neural networks) that discriminates between the formation of stimulus-driven associations and the formation of cognitive concepts. This test, the discrimination-shift task, was extensively applied from the fifties till the early eighties to animals, children, and adults. One point that Kendler and Kendler (1975) wanted to make, for example, is that although a simple discrimination task can be learned equally well by animals and humans; the former learn by forming simple stimulus-driven associations and humans (older than 6 years) learn by forming mediated concepts.

In the standard discrimination-shift task, subjects learn to discriminate on the basis of reinforcement contingencies between four stimuli which are presented in two distinct pairs. The stimuli are distinguishable on two dimensions: for example, shape (round/triangle) and color (white/black). Each stimulus pair appears in two configurations of which only the positions of the stimuli differ. The task comprises three phases: the pre-shift phase, the reversal-shift (RS) phase, and the extradimensional-shift (EDS) phase. The pre-shift phase continues until the number of correct responses in a sequence of adjacent trials meets a given criterion. After the pre-shift phase, learning continues, but the reinforcement is changed by either an RS or an EDS. An RS implies that all stimuli that received positive reinforcement get negative reinforcement, and vice versa. An EDS means that the dimension upon which the reinforcement is based, shape or colour, is shifted. The main difference between the two shifts is the number of stimulus-response relations that change: After an RS all relations change, whereas an EDS changes only half of the relations. On the basis of this distinction, Kendler and Kendler (1962) conclude that behaviorist models (e.g., Spence 1936) predict that EDSs are learned faster than RSs (i.e., the EDS needs fewer trials before criterion). In contrast, a model that presumes the use of a mediating concept (selective encoding) is expected to learn the RS faster because only the link between the mediating concept and the response should be changed.

It has been repeatedly found that college students execute an RS more rapidly than an EDS (Buss 1956; Harrow & Friedman 1958; Kendler & D'Amato 1955). In contrast, rats, pigeons, fish, and monkeys are found to execute an EDS faster than an RS (Kelleher 1956; Schade & Bitterman 1966; Tighe 1964). Kendler and Kendler (1962), report a study with kindergarten children in which the half who performed the pre-shift phase above median executed RS faster than EDS and for the other half of the children the reverse was true. All kinds of variations on this task (e.g., trial-by-trial analysis and the optional-shift task) give converging evidence for the distinction between animals and adults.

We performed an extended simulation study with three-layer feed-forward PDP-networks performing the discrimination-shift and related tasks (Raijmakers et al., submitted). It turned out that the learning behavior of all tested network configurations is equivalent to forming stimulus-driven associations, which agrees with behavioristic models on all performed tasks. The same was shown for ALCOVE with respect to the standard task (Kruschke 1992).

As Amit argues, the way neural representations are formed and activated in feed-forward networks differs qualitatively from the formation of neural representations in attractor networks. The question is now whether the formation of neural representations in Hopfield-like models simulates the learning of cognitive concepts. We argued that this can be tested by means of the discrimination-shift paradigm.

By definition of the task, some of the stimuli in the discrimination-shift task are highly correlated since they share one feature. A preliminary simulation study made clear that in order to simulate discrimination-shift learning by a Hopfield-like model it should be extended so that highly correlated stimuli can be learned on the basis of reinforcement and can be relearned with contradicting reinforcement. In line with Amit's target article, this should be a modular system which learns both inter- and intra-modular connections, since the stimuli have more than one distinguishable feature. According to Amit, the linking of separate neural representations is still an open problem. Hence, in our view it is too early to speak of cognitive concepts when we are dealing with neural representations that are formed in attractor networks by means of Hebbian learning.

Reverberations of Hebbian thinking

Josef P. Rauschecker

Section on Cognitive Neuroscience, Laboratory of Neuropsychology, National Institute of Mental Health, Bethesda MD 20892-4415.
josef@helix.nih.gov

Abstract: Cortical reverberations may induce synaptic changes that underlie developmental plasticity as well as long-term memory. They may be especially important for the consolidation of synaptic changes. Reverberations in cortical networks should have particular significance during development, when large numbers of new representations are formed. This includes the formation of representations across different sensory modalities.

What Amit refers to in his target article as "Hebb's paradigm" or, under its more common name, "Hebb's cell-assembly theory," incorporates at least two distinctly different roles for reverberations in cortical networks: (1) Reverberations in the "closed loop" of a cell assembly may *denote* short-term memory (Hebb & Donderi 1987); or (2) reverberatory activity may help to *induce* synaptic changes that are the substrate of long-term memory (Hebb 1949). In addition, Amit assumes as a third role for local cortical reverberations that they *are* in fact the internal representations themselves, which cognitive science has postulated (Fodor 1975; Pylyshyn 1980) and modern neurobiology seems to support (Singer & Gray 1995). It is important that these three potential roles for cortical reverberations are clearly distinguished, realizing that they may have nothing to do with one another except for being caused by the same underlying network structure. From a plain engineering point of view, every network with feedback connections (or "attractor network" in the terminology of Hopfield [1982] and Amit) has the capacity to reverberate. Hence reverberations may be considered just a byproduct of the network design, not necessarily bearing any functional significance, until it can be demonstrated experimentally.

Evidence for reverberations as a substrate for short-term memory is sparse, even though the idea, initially put forward by Lorente de Nó (1949), has always remained popular. It is especially worth noting that the concept of working memory has become increasingly influential among neuroscientists today (Goldman-Rakic 1992; Miller et al. 1991). Most studies point to the prefrontal cortex as an essential brain site involved in working memory. Other areas, including inferior temporal and parietal cortex, are thought of as the highest sensory regions from which prefrontal cortex draws when activating memories. We can imagine easily that information being shuttled back and forth between sensory and prefrontal regions would result in reverberations. However, as Amit is careful in pointing out, in comparison short term and working memory one must distinguish between "active" and "passive" memories, working memory falling under the former category, whereas short-term memory has conventionally been subsumed under the latter category.

The idea that reverberations *are* the internal representations has originally been pursued by theoreticians, because it relieves us from postulating the "little green man" looking at these representations. The concept includes the assumption that the information about representations is stored in the structure of the underlying network, that is, the weight of its synaptic connections, but that the information does not come "alive" until activated by an external stimulus or through another cell assembly. Again this is made clear by discriminating between active and passive memories. Passive memories are those that are structurally written down as "traces" (in the Hebbian sense); active memories are those evoked by external events and are thus intimately related to the internal representations or, in Amit's framework, cortical reverberations.

All such memories are content-addressable, and their most prominent feature is associativity (Kohonen 1984; Palm 1982). Associations are formed when two stimulus configurations co-activate two different cell assemblies, which can be situated in quite disparate cortex regions. When one of them is activated again later by one of the stimuli, the other becomes co-activated, and the memory of the second stimulus springs to mind as well. Thus incomplete information may be sufficient for retrieving the whole.

Perhaps the best evidence for a role of cortical reverberations is available in long-term plasticity. Most of the evidence comes from studies of visual cortex in young kittens that are exposed to different kinds of visual patterns (see Rauschecker 1991 for review). Strictly speaking, however, even in this well-explored system, synaptic changes are induced by correlation of pre- and postsynaptic activity (Sejnowski 1977; von der Malsburg 1973), as long as this correlation exceeds the necessary threshold (Singer 1990). While the presence of reverberations is not an absolute necessity, the availability of some sort of feedback is an important requirement because it signals the postsynaptic activation back to the synapses, in order to initiate their strengthening (or decay). The feedback could be provided by recurrent collaterals or other axonal connections, but it could equally well be produced within the membrane of the postsynaptic cell itself, making use of depolarizations finding their way back into the dendrites. The ensuing biochemical cascade most likely includes the action of calcium-dependent protein kinases (Stevens et al. 1994). The time it requires to bring about permanent synaptic change has often been termed a "consolidation period" and has been found to exist both in long-term memory (McGaugh & Herz 1972; Squire 1987) and in developmental plasticity (Rauschecker & Hahn 1987). It is attractive to think that reverberations may help to accelerate this consolidation.

Tying two of the potential roles for cortical reverberations together, it is intuitively clear that a young organism, which is still in the process of building an enormous number of new representations every day, would need a large amount of plasticity in its connections. It also makes sense to reduce this plasticity later in life, in order to stabilize at least some of the representations. While it is commonly assumed that synaptic plasticity is controlled largely by chemical factors, it is certainly possible that closure of the sensitive period for cortical development is also brought about by the fact that increasing memory space is being taken up by an increasing number of stored representations.

The formation of representations that link different sensory modalities is of particular importance during development. As an example, one may just think of the development of speech and language (Kuhl et al. 1991) or of spatial perception (Rauschecker & Sejnowski 1994). In crossmodal plasticity, one representation can be partially replaced by another, but only at a level where the two representations share the same neural code (Rauschecker 1995; Rauschecker & Korte 1993; Rauschecker et al. 1992).

At this level, the distinction between development and learning, one being controlled largely by intrinsic events, the other by extrinsic ones, starts to vanish. Both processes are controlled by neural activity in the cortical network, or possibly (in Amit's framework) by its reverberations. Cortical activity or reverberations form the bridge between maturation, experience-dependent plasticity, and learning.

Association and computation with cell assemblies

Frank van der Velde

Unit of Experimental and Theoretical Psychology, Leiden University, 2333 AK Leiden, The Netherlands. **vdvelde@hlerul55**

Abstract: The cell assembly is an important concept for cognitive psychology. Cognitive processing will to a large extent depend on the relations that can exist between different assemblies. A potential relation between assemblies can already be seen in the occurrence of (classical) conditioning. However, the resulting associations between assemblies only produce behavioristic processing or so-called regular computation. Higher-level cognitive abilities most likely result from nonregular computation. I discuss the possibility of this form of computation in terms of cell assemblies.

The discussion of the reverberations in cell assemblies initiated by Amit is of direct relevance to the question of how cognition is generated by neural processes. Hebb's influential idea of the (reverberating) cell assembly as the basis of internal representation has now gained both experimental and theoretical support, as clearly described in the target article. The picture that emerges is that of the cortex consisting of a large set of local modules or assemblies. Tanaka (1993) indeed describes such a set of assemblies in the form of columns in the inferotemporal cortex, which is thought to be responsible for visual object recognition. Each column is selective for a set of related (and abstract) visual features, yet adjacent columns are typically responsive to different visual forms.

From the perspective of cognitive psychology, the relevance of assemblies will then be determined by the internal structure of an assembly (the features or temporal correlations that it represents) and by the relations that exist *between* different assemblies. It can be argued that to a large extent (higher) cognitive computation will depend on the relations between existing assemblies, because theoretical considerations show that a module of about 10^5 neurons can only store and retrieve a limited number of representations or attractors (e.g., see Amit 1989), which is indeed what is described by Tanaka (1993).

Furthermore, the experiments of Miyashita et al. discussed in the target article show that it takes quite a long period of training to form a particular assembly. Hence, learning could in many cases consist of the formation of new relations between existing assemblies. A case in point is given by (classical or operational) conditioning. In conditioning a (cognitive) association is formed between two hitherto unrelated but familiar stimuli. The reverberations described by Amit might explain how this could take place. The first (familiar) stimulus will result in the activation of its assembly (or assemblies), because the stimulus is represented by an attractor in that assembly. Because this stimulus is represented by a reverberating attractor, activation in the assembly will persist after the stimulus has disappeared. The attractor could still be active when the second stimulus activates its particular assembly. Hence, during a period of time neurons in both assemblies will be active. The simultaneous activity of these neurons could result in an association between the assemblies by means of the Hebb learning rule, as found in the process of long-term potentiation or LTP, which is a (long lasting) strengthening of the synaptic efficacy between a pre- and postsynaptic cell (e.g., see Baudry & Davis 1991).

The reverberating activity in an assembly plays a crucial role in this description. First, it results in the simultaneous and persistent activation of the pre- and postsynaptic cells, which is necessary for LTP to occur. Simultaneous and persistent activation may be difficult to achieve with neurons whose activity was hitherto

uncorrelated. But if each neuron belongs to a (different) assembly, persistent activity is much more likely (because it results from the internal structure of each assembly), and the simultaneity of activity no longer depends completely on the simultaneous presence of the stimuli. Second, the association between reverberating assemblies could proceed relatively quickly. An association between assemblies could already be formed even if only a small fraction of the neurons are actually associated, because the activation of a fraction will be enough to reactivate the whole attractor in the assembly (thereby producing the association). Hence, the possibility of relations (associations) between assemblies is in part determined by the internal structure of a cell assembly, in particular, by the possibility of reverberations as described in the target article.

Yet associations between assemblies would only produce cognition as described by behaviorism (e.g., see Amsel & Rashotte 1984). Since the "cognitive revolution" of the fifties and sixties it has been the accepted view that (higher) cognitive behavior (for example language processing) cannot be modeled or explained on the basis of mere associations (e.g., see Simon & Kaplan 1989). Instead, computations are needed to model cognition according to this view. Only in recent years has this (classical) view been challenged as a result of the rise of connectionism (e.g., see Hintzman 1993). It can be argued, however, that connectionism has not yet generated the productivity and representational adequacy necessary to model language processing, precisely because until now connectionism has reproduced the associate processes described by behaviorism (e.g., see Pinker & Prince 1988; Reilly & Sharkey 1992).

It is incorrect to describe the distinction between behaviorism and classical cognitive psychology in terms of computation per se. Behavioristic processes are also computable processes, but of a restricted sort. They consist of so-called regular languages or computation, produced by finite-state automata. In contrast, the processes described by classical cognitive psychology consist of so-called nonregular languages or computation. The distinction between regular and nonregular computation can be described on the basis of the Turing machine. The Turing machine is a finite-state automaton, which contains its program connected to a working memory in the form of a tape. It is important to note that all programs on the Turing machine (and thus on the computer) are finite-state automata. In other words, the rules used in any computation are associations as described by behaviorism. Without the tape, the Turing machine would only produce regular languages, thus behavioristic processes. With the tape, however, it can produce nonregular languages because it can store "intermediary" results on its tape, to be used later in the production. This possibility of storing intermediary results is the basis of the greater productivity of all nonregular languages. Hence, it is the crucial difference between behaviorism and classical cognitive psychology (see Van der Velde 1994).

Intermediary results can be stored because the tape (working memory) is separated from the program. This is true for all automata that generate nonregular languages, although the nature of the working memory can differ (such as a stack in case of the pushdown automaton). When such a working memory is used in a computation, representations (symbols) are copied and stored in the memory. As a result, precisely ordered strings of individual symbol tokens of arbitrary composition (or constituent structure, see Fodor & Pylyshyn 1988) can be stored and manipulated. However, as Amit has pointed out, representations with cell assemblies are different from those in the computer (or Turing machine) because they are not copied and stored in or removed from a separated memory. Instead, representations given by assemblies are unique and their place remains invariant. This characteristic seems to preclude nonregular computation, and thus cognitive processing as described by classical cognitive psychology, with cell assemblies.

Yet the difficulty can be resolved by looking at it from a different angle. The crucial characteristic of nonregular computation is not the possibility of making individual copies of representations per se, but the possibility of associating representations with different (and unique) position tags. For example, the string *ABBA* on the tape positions h, i, j and k is in fact similar to the set of associations $A - h$, $B - i$, $B - j$ and $A - k$. These associations provide the possibility of producing nonregular languages with cell assemblies (Van der Velde 1995). Assemblies that represent symbols are associated with assemblies that represent position tags. Both types of assemblies are unique and their place is invariant. It is only through their associations that strings of individual symbol tokens can be formed. A dynamical system with assemblies can produce these associations at the appropriate moment using the process of LTP described above (for details, see Van der Velde 1994). It is important to note that the reverberating nature of activation in an assembly, as described in the target article is a prerequisite for this system to operate in an appropriate manner.

How do local reverberations achieve global integration?

J. J. Wright

Mental Health Research Institute, and Swinburne Center for Applied Neuroscience, Melbourne, Victoria 3052, Australia. **jjwacortex.mhri.edu.cu**

Abstract: Amit's Hebbian model risks being overexplanatory, since it does not depend on specific physiological modelling of cortical ANNs, but concentrates on those phenomena which are modelled by a large class of ANNs. While offering a strong demonstration of the presence of Hebb's "cell assemblies," it does not offer an equal account of Hebb's "phase sequence" concept.

Amit's target article presents us with a powerful demonstration of the relevance of ANN modeling to central cognitive processing. In view of this achievement, excessive criticism would be carping but two general types of reservation seem in order.

The first of these reservations concerns the extent to which ANN models are overexplanatory when applied to real neural behaviours, if we are willing to excuse those aspects of experiments which are not adequately modelled. Amit's discussion in section 7 shows that ANN models, when interpreted carefully, are fairly robust in their capacity to match experiment, even when they involve great simplifications of physiological reality. He appeals to the persistence of simplifications in his "toy model" to account for "some systematic differences in the details of the attractor correlation coefficients and rate distributions."

In this light it is reassuring that he advances strong and clear predictions for future psychological experiments (sect. 6). These predictions appear to be robust over most classes of ANN, but they do not correspondingly help refine the physiological model. As an example of this concern, a colleague and I have calculated that the asymmetry of real intracortical connections is very high (Liley & Wright 1994) – much higher than in the models used by Amit. While we suspect (but have not yet demonstrated) that highly asymmetric ANNs, under Hebbian learning, could also be used to model the Miyashita findings, it would be disturbing if this proved not to be the case.

It is similarly uncertain whether the low firing rate model applied by Amit (Amit & Tsodyks 1991a; 1991b) is always applicable in the cortex. In a recent publication (Wright & Liley 1995) we have developed a low firing rate lumped model, based upon leaky-integrate-and-fire neural dynamics similar to those applied by Amit. When couplings based upon estimates of real cortical synaptic densities are introduced, this model simulates many properties of the EEG. However, to meet other considerations, relatively high synaptic gain has to be introduced. The model is then highly unstable, in the sense that it readily exhibits transition to high cell firing rates. Since real brains are readily provoked into grand mal epilepsy, this property appears plausible.

A further, and more general reason for considering that realistic ANN models of cortex will exhibit transitions between high and low firing rates, is that they may then belong to those classes of cellular automata operating on the "edge of chaos" (Langton 1990). In such automata particularly rich dynamic interactions with computational properties are favoured.

Perhaps an even simpler reason to consider models exhibiting transitions to high firing rates, is that local reverberations must develop a sufficiently high output signal/noise ratio, if they are to have significant influence on the subsequent state trajectory of the entire brain.

The second reservation concerns the danger of making too strong a set of predictions about cognitive phenomena (sect. 6) from a neural network model which is an incomplete model of cognition. That the present model is incomplete is surely the case. Amit has warned us not to indulge in excessive speculation on the basis of the success of his model, but these findings demand further speculation, not least because they represent a powerful vindication of Hebb's own ideas and encourage more of the same.

In his opening statements (sect. 1) Amit asserts that most of the results in Miyashita's seminal experiments "could be predicted on the basis of simple observations on common cognitive phenomena, without recourse to any specific model." Here he has us, in imagination, struggling to find an appropriate translation for a given word. We try and fail, and maybe "the entire episode, word task, withers away from our consciousness, only to surface resolved hours or days later."

Does Amit really want us to believe that some small ensembles of cells have continued to reverberate, storing this problem, throughout the entire waiting period, through sleep and waking, and all manner of distractions? Wouldn't these intervening events have changed the attractor state of the cells a number of times? What did the resolution itself consist of? What was the underpinning of the subject's belief that the problem was soluble? What sustained the system in attempting to solve this problem and not others?

I presume that the answers to these queries might be concerned with Hebb's "phase-sequence" concept, rather than that of the "cell assembly" (Hebb 1949). Hebb was at pains to consider the details of transition between sequences of active cell assemblies, but was perhaps less persuasive and specific in his account of how this comes about than elsewhere in his work. Would Amit please give us his further insights into this global aspect of the organisation of brain function, since he has shown us that interacting local ANNs with Hebbian learning can take us some of the way?

Author's Response

Empirical and theoretical active memory: The proper context

Daniel J. Amit

Racah Institute of Physics, Hebrew University, Jerusalem; Istituto di Fisica, Universita di Roma, Rome, Italy. **damita@il.ac.huji.fiz.ilios**

Abstract: The context of the target article is delimited again, underlining the intended location of the argument in the bottom-up hierarchy of brain study. The central message is that collective delay activity distributions (reverberations) in cortical modules extend the role of a spike (a potential information carrier across long distances) to an active memory of structured, learned information that can be carried across long time intervals. Moreover, the population code of the reverberations makes them readable down the cortical processing stream. Most of the critical comments are then interpreted and addressed in relation to misreading of the proper context. The price for the limitation of the context (in cognitive, behavioral, and computational terms) is compared with the advantages of a clear, direct contact with experiment on the one hand and with a well controlled body of modeling and analysis on the other.

R1. A metaphor to define the context

The wide spectrum of reactions to the target article indicates to me that perhaps the dimensions and the orientation of the context addressed has not been made sufficiently clear. I would therefore try to clarify it by a metaphor: the Hodgkin-Huxley model for the mechanism of spike emission of a neuron. Clearly, neither the phenomenon of neural spike emission nor the model accounting for it have resolved any cognitive problem, behaviorist or otherwise. Yet the idea that spikes may be a basic element in the construction of an account for brain function was immediately appealing. There was no direct connection of spikes to behavior, except perhaps for the fundamental observation that whatever the brain concludes must travel over long distances and spikes can provide a natural carrier. Nor was the model compelling in its details. In fact, by now many elements have been added to the model: additional ionic channels, time constants, and so on. And yet the mere beginning of an analytic control over a neurophysiological phenomenon of computational appeal was itself very appealing.

Spikes have led to an amazing range of speculation. Most impressive is perhaps the beautiful, if extremely naive, program of McCulloch and Pitts, who proposed to connect the apparent ability of brains to carry out logical calculus to the fact that spikes, synapses, and thresholds can provide AND and OR gates. Hence, a long chain made of links of very few neurons each could generate the most complex logical predicates. Spikes and synapses have also allowed the invaluable speculations of Hebb discussed in the target article.

What I am trying to call attention to is the possibility that another step in the spike tradition, rather than in the Hebbian one, can be achieved. At least in one simple situation we are again at a point where a very clear, independent brain phenomenon can be observed and is accompanied by a plausible model, hence a semblance of analytic control. The additional rung in the ladder builds on spikes, taking into account that synapses are weak and therefore if spikes are to play a significant role, neurons must act in concert. Yet, the concert cannot be too harmonic, since the neural context is very noisy. Just as the spike is an intrinsic neural property, so the attractor, the reverberation, is an intrinsic property of a realistic assembly of *neurons and synapses.* The potential for long distance transmission of information by the spike is supplemented by an assembly code of potentially long, autonomous duration. Moreover, the resulting codes are related both experimentally and theoretically to learning. The outcome is a collective code which can propagate long distances, since it is composed of spikes. What it propagates can, in principle, be read (at the "receiving end") because a sufficient number of neurons participate and can overcome the intrinsic weakness of synapses.

The phenomenon is observed *in vivo,* in performing

animals, as spikes are. Its structure is analyzable, since it is essentially independent of the details of the provoking stimulus. It can be discussed and tested, independently of its cognitive content or role. Not that the cognitive role is less important, but it is more difficult and may require many of the modifications suggested by the commentaries. Just imagine trying to give a cognitive role to spikes in the analogy.

R2. Simulations and mathematics

The line of research summarized telegraphically in the target article employs mathematical analysis as well as numerical simulations, not exclusively simulations (**Hirsch**). To get a glimpse of the mathematics (see Amit 1989; Amit & Tsodyks 1991; Amit et al. 1985; Griniasty et al. 1993). And yet in every case we resort to extensive numerical simulations. There are basically two motivations for the use of simulations:

1. Our mathematics may be considered "disgraceful," since it is neither rigorous nor exhaustive. The unfortunate fact is that complex systems (and one naturally considers one's own brain that way) have no rigorous mathematical treatments. We have been analyzing them by various techniques of probability theory, of statistical mechanics, of disordered systems, and so forth. As beautiful and powerful as those tools are, the solutions they provide are at best approximate; moreover, it is not always clear that the set of quantities exposed by the theory provides a complete description of the system's dynamics. Hence one motivation is to ensure that the analytic solutions are close enough to the dynamics of the model that had been set up, and that no special features are hidden, uncaptured by the approximations involved in the theory.

2. But no less important is the communicative role of simulations. There is a large linguistic gap between theorists (modelers) and neurobiologists. What most often preoccupies the biologist is not whether we have used rigorous mathematics to solve for the essential properties of the model, but the model itself. One way to produce a common language is to set up a credible simulation of the model that is accessible in a form familiar to someone who might be carrying out an experiment on the system presumed to be modeled.

R3. Assemblies and attractors

It should be emphasized that the assembly referred to in the target article is not the set of neurons that maintain an elevated spike rate following the presentation of a familiar stimulus (**Hirsch**). The assembly is that body of properly interconnected neurons, able to support *all* the different attractors (100 in Miyashita's, 1988, case) but many more in principle. This may be the column, with its 100,000 cells. Each reverberation is expressed by about 1,000 neurons with elevated rates. These cells may be shared by different attractors, as is observed in experiment and is accounted for by theory. One of the main achievements of the nonrigorous theory developed in the last 14 years has been to demonstrate in a quantitative way that it is possible to have a large number of attractors (with learnable structure) in a single assembly. Thus the reverberation is not a selection of an assembly (**Klimesch**), it is much more: in the selected assembly, which stores (by learning) many potential reverberations in its synaptic structure, the stimulus selects one which can then propagate across the delay. It may also be part of a wider web of interacting modules (see sect. R8, local and global assemblies, below).

Finally, the types of attractors, critical points, and bifurcations, from Zeeman to Thom, resulting from differential equations (**Hirsch**) are really very simple compared to what a system of neurons, inundated by noise, with conflicting interactions (excitatory and inhibitory synapses) and with very ambiguous boundaries can produce. The attractors discussed in the target article are neither fixed points nor limit cycles. They are closer to the dynamical structures suggested by **Freeman.** The structure of attractors reviewed in the target article, and based on selective persistent rate distributions, is very naturally accessible to the next neural module down the computational line, if not to the mathematician. I am less sure that this is true of Lyapunov exponents or power spectra of noise.

As for reading the attractor, I share **Fuster**'s view that it must involve the output of a relatively large group of neurons which on average have spike rates above some threshold. Otherwise, the animal seems to make errors (**Fuster,** and Funashi et al. 1989). Where we differ – in perspective rather than on technical grounds – is that Fuster prefers to consider the sampled group of neurons to be widely distributed while we have shown that there exists an option in which a single local module can maintain active an internal representation of a type that can be read by neurons down stream. Moreover, the local picture has the additional attraction that already at this level one finds, for the first time, structural features of the presumed internal representations (the Miyashita correlations). I tend to believe that Fuster's option is the way the different local attractors "bind."

But given that there is an agreement on the question of a population code, I am somewhat puzzled as to the relevance of **Fuster**'s Figure 2. Unless one can show that many of the neurons that participate in the same delay activity have the particular time course of spiking, the fact that a single cell has it will be washed out when the population average is performed. Such population averaging is unavoidable, when the "content" of the delay activity is to be communicated for downstream processing. It is dictated by the weakness of single synaptic contacts (see also discussion of local and global assemblies, sect. R8). I would say in passing that the appearance of a neuron with similar features in the model of Zipser et al. (1993) is to be taken with caution – not because of the way learning is affected in that model, but rather because of the very particular neurons used.

Finally, it may be worth emphasizing that there are no "Amit-style" attractors (**Bienenstock & Geman**). There are attractors observed in single electrode recordings. There are corresponding attractors in structureless neural networks, with quasi-realistic neurons, connectivity probabilities, noise levels, and synaptic values that also sustain selective delay activities.

R4. Defense of specific features

1. Is there evidence for columnar organization (**Dalenoort & de Vries**)? At the time of writing the target article I was informed of such organization on the scale of 1mm^2 in

interior areas of cortex as inferotemporal only by informal communication. Since then it was reported several times by Miyashita (1988), Goldman-Rakic (1990), Tanaka (1992), and others, in electrophysiological as well as histologic studies of both IT and prefrontal cortex. The first clearly depends on sampling, but essentially all cells exhibiting the persistent delay activity were bunched together, while outside the column and in the same region they were not observed. An estimate of how many such modules can be formed in the entire zone would constitute a guess, but in IT of a monkey it may be a couple of hundred.

2. It is not the case in our models that a critical level of neural excitation drives the network rapidly to maximum rates (**Dalenoort & de Vries**). This would have been the case in the absence of inhibition. Excessive excitation in presence of feedback inhibition results, typically, in activity reduction or even extinction. This effect may provide a mechanism by which one network may inhibit another, despite the fact that long range axons are usually excitatory.

3. The relation of persistent activities to modulators (**Ahissar**) can be intended in two ways: (a) That neuromodulators are essential for synaptically controlled attractor dynamics to manifest itself, or (b) that modulators replace synaptic potentiation. The data produced by Ahissar support the first reading, I believe, though it seems that the commentator would rather have the data imply the second.

Note, first, that no selectivity is present in the data. There is one cell that is driven by a single stimulus. It maintains a delay activity when a neuro-modulator is around and does not do so in its absence. I would suggest that in the given module, in the anesthetized brain, the level of spontaneous activity is too low and as a consequence neurons' depolarizations lie too low and selective activity does not manifest itself. The injection of the modulator seems to restore higher spontaneous activity levels, which in turn make attractor dynamics possible. It would be interesting to test this hypothesis. But if this were the case, we would still have to decide who determines the selectivity – and back to Hebb, synapses and reverberating assemblies.

The second reading has it that all and only stimulus driven neurons would be active in the delay. (a) This is contrary to the findings by Sakai and Miyashita (1991), Fuster (1973), and others. (b) It does not account for persistent activity in neurons which are not selective for a particular stimulus, but only for its neighbors in the training sequence, and carry persistent activities due to the correlations found by Miyashita. (c) It does not distinguish between stimuli that have been learned and those that have not, a distinction clearly observed in the delay activity recordings. (d) It is not accompanied by a computational proposal.

4. I have no empirical evidence that delay activities persist for more than a minute (**Lansner & Fransén**). But to conclude from that working memory is mediated by synaptic changes (**Milner** and **Chown**) goes against the empirical evidence (see point 3, above).

5. Is there an upper limit to the amount of time an assembly can remain in an attractor (**Chown**)? We do not have any empirical evidence for the existence of an upper bound on the duration of a reverberation. Chown mentions 5 seconds as an upper bound, measured in experiments related to memory consolidation time. Such times are in contrast with the 16 seconds observed by Miyashita (20 seconds in **Fuster**'s data). Moreover, consolidation time can be related to the duration of reverberations only if one assumes that significant learning is taking place while the network is in an attractor. As I argue in the discussion of learning (sect. R6) this is not an advisable hypothesis to make.

6. Do the present models provide mechanisms for inactivating an attractor (**Chown**)? There are several mechanisms (in the models) that will interrupt a reverberation: (a) If an unfamiliar (unlearned) stimulus arrives while the module is in a selective attractor, all neurons in the network will relax to spontaneous rates (Amit & Brunel 1994). (b) If a different learned stimulus arrives, the network will switch from one attractor to a new one. (c) If the background noise afferent from the spontaneous activity outside the module is significantly decreased, the attractor activity will cease. This last feature may provide a mechanism for implementing attention.

7. It is gratifying that the theory (the model) of the assembly dynamics does not require structured "closed loops" to reverberate "around" (**Milner**). The reverberating attractor behavior is produced in networks connected at random, provided the probability for a connection is not too low, and the connectivity suggested by anatomy is plenty.

8. **Wright** raises a valid doubt about the stability of networks with neurons tightly engineered to produce low rates in the attractors (as in Amit & Tsodyks 1991). Since the publication of that paper, we have become aware of the fact that inhibition takes much of the burden of maintaining low rates away from the neural gain function, as has been also confirmed in simulations of large networks of integrate and fire excitatory and inhibitory spiking neurons (Amit & Brunel 1995). This is also the answer to the concern about the level of the rates in the delay activities (**Ahissar**). We can obtain elevated selective delay rates which are essentially limited from below only by the existence of spontaneous activity. The data presented in the target article, reproduced from a detailed study (Amit et al. 1994), have neurons at 20–25 spikes/second.

9. This issue connects also to the mechanism that controls the rates (**Lansner & Fransén**). There may be more than one mechanism that can do the job. The comment suggests the saturation of synaptic efficacies by kainate/AMPA and NMDA. We have opted for excitatory noise and unstructured inhibition for several reasons. First, it is much simpler. Second, the simpler mechanism also suffices for the reproduction and the stabilization of spontaneous activity (see Amit & Brunel 1995). For such low rates, a couple of spikes per second, it would seem difficult to invoke synaptic saturation.

10. Attractors are not characterized by discrete frequencies (**Fuster**) in two senses: (a) In an attractor an entire subpopulation of the local module has elevated spike rates. There is a rather wide distribution of rates among the neurons active in each attractor (Amit et al. 1994). (b) Activity in a given attractor may have different rates depending on the unselective afferents received from outside the module. This may be an alternative way of understanding the remarkable cooling experiments reproduced in Fuster's commentary. I cannot exclude the possibility that Fuster's explanation is the one that stands. Both mechanisms are available and the data advanced are not sufficient to discriminate which one is operative in the particular experiment.

11. Variability from trial to trial and multi-electrode

recordings (**Ahissar**): (a) There is in fact a fair amount of variability from trial to trial, but what impresses me much more is the full half of the cup: the reproducibility of average spike rates between trials at the level of single cell recordings. (b) Our proposal is to see how far one can go with average rates. The observed rates are consistent with both anatomy and elementary physiology. Should we celebrate this fact or emphasize the noise instead? Note that the difference in rates between selective and nonselective cells is far beyond the intertrial variability. (c) When summed over some 1,000 participants in a selective attractor, as would be required for reading an assembly code (see discussion in sect. R3, assemblies and attractors), much of the variability will be eliminated. (d) A similar type of variability appears in the models and does not create any problem of classification. (e) The alternative proposal is multi-electrode recordings. What is the associated learning and computational proposal? How does it deal with intertrial variability in rates? The technology has been around for a number of years, and it has not proved a panacea.

12. We do not put "several assemblies together" (**Lansner & Fransén**). In fact, we take a single local assembly, in which we model neurons by neurons! Modeling entire assemblies as individual processing units requires an additional level of modeling to show that the new units operate in the assumed fashion while respecting, internally, biological plausibility. We find this additional level unnecessary. Moreover, we believe this is a more faithful description of the anatomy than that expressed by widely spread assemblies (see discussion of local and global assemblies, sect. R8). Our network performs the task with just a few percent local connectivity, which is in accordance with anatomy, if the network does not extend beyond a scale of 1 mm.

13. I agree that when excitatory and inhibitory neurons are intermixed the response of the new composite unit to a common input may be nonmonotonic (**Morita**). In our networks, with excitatory and inhibitory neurons mixed in an *unstructured* way, if nonselective input is afferent to the network, its output rate first increases and then decreases. I fail to see the need, or the utility, for structured composite elements of the memory network because: (a) there is no anatomical evidence for it; (b) with the typical strength of synapses in cortex, a composite unit of few neurons will not work; (c) the unstructured network with simple elements works quite well and we are not promised any bonus for the additional complication.

14. We assume no symmetry of the connectivity (**Wright**). Whenever symmetry is invoked it is done in order to simplify analysis. But every consequence has been tested in simulations with symmetry removed. Perhaps an extra clarification concerning symmetry is required. In models reproducing the Miyashita correlations, symmetry may appear at two different points: (a) In the general connectivity, that is, where any two neurons have the same connection strength when the afferent and efferent neurons are interchanged. Persistent, selective rate distributions are very robust to the removal of this type of symmetry (Amit et al. 1994). (b) In a given synaptic efficacy, with respect to the interchange of the terms which connect the present stimulus with the preceding and the successive ones in the training sequence. The results are somewhat more sensitive to deviations from this symmetry.

15. **Hoffman** raises questions concerning coding levels and storage capacity. Experiments give an estimate of about 1% coding level. This means that about one neuron in a hundred in the assembly is driven by a stimulus. This is the type of coding level we have used when we tested whether the phenomenon of the conversion of temporal correlations persisted in networks with realistic neurons (Amit et al. 1994). When the attractors become correlated, the fraction of neurons participating in the delay activity becomes somewhat higher. The storage capacity of networks increases very significantly as the coding level of the stored patterns decreases (Amit 1989; Gardner 1987; Willshaw et al. 1969), as suggested in the commentary.

16. The curves shown in Figures 5 and 6 (**Hoffman**) are all taken from the model with realistic neurons in which coding was sparse and rates were low. There is no direct relation between the two, however. The rates depend mostly on the nature of the inhibition and on the neural response function. Sparsely coded networks can operate at saturation levels and in models often do.

17. The magic 5 (**Hoffman**): Strangely, and this is the magic, in the least realistic model the range of 5 in the correlations requires no parameter adjustment and is invariant under the variation of the only parameter in the model (Griniasty et al. 1993). This is not the case for realistic networks (Amit et al. 1993). In such networks the range of correlations do vary, as suggested in Hoffman's commentary. It depends on several parameters such as the strength of inhibition, the strength of the interstimulus coupling and so on. In this sense, the model of the realistic network does not produce oracular magic. Instead, it provides a predictive, testable framework for monitoring network properties and learning protocols. What remains impressive, I believe, is that a relatively simple network, with a simple realization of a learning scenario, can reproduce the data in its full detail *and* be strongly predictive.

18. There are indeed many potential internal structures for attractors (**Petitot, Hirsch**). Some important ones relate to chaos, synchronization, and synfire. They may all be pertinent to the understanding of a system as complex as a brain. My point of view is that the simplest are attractors which preserve elevated spike rates without invoking correlations between spike emission times. Simplicity in a context of such complexity is not to be cast away simply for the sake of mathematical richness. Simplicity in this context implies: (a) very easy experimental access and classification; (b) readability by another neural system; (c) clear correspondence with a class of models; and (d) models accessible to analysis leading to predictions.

If simple attractors are to be abandoned or supplemented by more complex ones, some weighty arguments should be brought to bear. The price is very high, though perhaps unavoidable. I have not yet seen such compelling reasons.

R5. What is not there

In the course of studying and developing the theory of attractor neural networks, a very rich world of phenomena was exposed. Some of them have proven robust; others have turned out to be ephemeral features of the toy model.

1. Unfortunately, the beautiful cosmos of spurious states, which has turned out to be the most popular aspect of our first study (Amit 1989; Amit et al. 1985), is totally absent in models with more realistic neural elements.

Spurious states are a plethora of attractors which are not reflections of the stimuli presented for training. They are additional attractors, symmetric mixtures of all stimuli, generated in an uncontrollable way by the highly nonlinear neural dynamics. The loss (from the standpoint of **Krakauer & Houston** or **Hoffman**) is compensated for by the fact that the uncontrollable and symmetric way in which these spurious attractors infest the dynamics has proved an embarrassment rather early in the game (Amit 1989). One would hope that the mechanism of Enquist and Arak (1993) which gives inferior males an opportunity in their affective life, can be salvaged even in the absence of spurious states.

2. In networks in which attractors engage neurons at spike rates so much below saturation, there seem to be no simultaneously coactive attractors. Inhibition is too strong.

3. In this connection it may be useful to clarify a three-way distinction between spurious attractors, correlated attractors, and Miyashita-type correlated attractors. The first, as mentioned above, are attractors which are not learned, but parasitically accompany a set of learned attractors whenever the neural elements have a high gain. The second may be a set of attractors learned from correlated stimuli. The third, the attractor correlations that had motivated the target article, are generated by the process of learning even when there are no correlations in the set of stimuli used for training or if correlations are eliminated by earlier stages in the processing (see Atick 1992). Then, in testing, one presents the uncorrelated stimuli and the persistent activity distributions are correlated.

R6. On the matter of learning

1. Does overlearning take place in reverberating systems (**Chown, Hirsch**)? The question has been raised in the context in which several modules may be reverberating simultaneously. In this case, Hebbian synaptic dynamics may potentiate all synapses between any two such assemblies. Even though I have been trying to stay clear of multiassembly systems, the single assembly does give some partial answers to this question. Note, first, that the same issue presents itself even more menacingly on the level of the single assembly. If that assembly reverberated it would go on learning and submerge all synapses to end up with a single attractor (Dong & Hopfield 1992). Yet the experiments discussed in the target article show quite clearly that this does not happen. Up to 100 attractors survive many repetitions of long reverberations in each. This is not because they are instantly learned each time, as I have argued above. It is because attractor dynamics do not imply overlearning, either experimentally or theoretically.

Present evidence therefore suggests that some synaptic modification may take place during testing, or during the persistence of an attractor (**Hirsch**). But not very much.

2. How does learning avoid it? A suggestion that works both in extensive simulations (Amit & Brunel 1994) and in electronic implementations of learning attractor networks (Badoni et al. 1994) is the following: realistic neurons can sustain three different levels of spike rates – spontaneous (very low); attractor (low); and stimulus driven (relatively high). This is an empirical fact. If one assumes (and this is a productive speculation) that potentiation takes place only between neurons of which at least one has a high (driven) rate and none has spontaneous rate, then attractor dynamics do not produce overlearning.

3. This point relates to a misconception, namely, that reverberations are a mechanism for synaptic potentiation (**Rauschecker**). I do not believe that this is Hebb's position, nor is it a tenable one. Significant structured synaptic plasticity must occur before reverberations can persist in a given neural module. Consequently, reverberations cannot underlie learning. Learning in attractors is neither necessary nor desirable. It is unnecessary because it can be done during stimulus presentation. It is undesirable because it leads to destructive overlearning, as mentioned above.

4. Learning algorithms that respect the constraints mentioned in point 2 above, and hence not producing overlearning, allow for learning of the Miyashita correlations (Brunel 1995). What is essential is that the information of a previous stimulus be carried across the delay. This is guaranteed within the attractor paradigm by the attractor associated with the preceding stimulus. Moreover, if one attractor is active and the next stimulus arrives, there are pairs of neurons of which one has a driven rate and the other attractor rate, and correlations can be coded (Brunel 1995).

Hence, it is not merely the Hebbian mechanism which creates the correlated attractors (**Krakauer & Houston**), but the combination of this type of learning with the fact that when stimuli are presented in fixed order, there is an agent (the uncorrelated attractor) that can carry the information from one stimulus to another.

5. In the process of stimulus presentation, attractors can be destroyed if there is a significant change in the statistics of stimuli in the input stream (Amit & Brunel 1994). Upon testing, one presents either stimuli that have been learned or random new stimuli, we would therefore expect very small changes in the structure of the attractors. It is only when a new image begins to be presented frequently in the process of testing that a significant deformation in the landscape takes place. This should be tested experimentally, though it is by no means easy.

6. I do not share the suggestion by **Rauschecker** that storing in memory may block learning. As mentioned above, not many hard facts are known about learning. To me it seems that learning and its manifestations are dynamical properties of neurons and of synapses. What is or is not in memory is not an additional variable, and unless introduced as such, cannot play a legitimate role in evolutionary description. In other words, a synapse knows its efficacy and the activity of its two neighboring neurons. It does not know what the network knows. Hence the change of its state must depend on these variables and on the biochemistry of the synaptic dynamics. The latter may vary with evolution, but no viable proposal on how this can be connected to memory storage levels has been put forward.

7. The question of storage capacity (**Hoffman**) takes on a very different meaning, in situation of realistic learning (Amit & Brunel 1994; Amit & Fusi 1994). Memory overload and associated mixing and confusions are a rather special feature of networks in which all memories are put in symmetrically, by hand, into the synaptic structure, as in Hopfield networks. When stimuli are gradually learned, there is no overloading breakdown. Instead, the network becomes a palimpsest in which new memories replace old ones. Storage capacity then measures the distance into the past for which memories can be recalled. The capacity of networks with realistic neurons and sequential gradual learning is not well known.

R7. Inferotemporal vs. olfactory

I see basically no disagreement with **Freeman,** whose investigations of the dynamical and computational functions of the olfactory system are in a similar vein to ours. I have preferred to concentrate on the Miyashita experiments and, had I been better informed, I would also have mentioned those of the Goldman-Rakic (1990) group and of Miller and Desimone (1994; see, e.g., **Hucka et al.**). One can clearly identify an autonomous mode of the local module, because the absence of the stimulus is twice confirmed: once on the screen and once neurophysiologically. Such experiments provide access to the sensitive modes of the assembly, which will eventually be driven by external afferents. It allows us, in my view, to distinguish features connected to learning from those imposed by stimuli.

The phenomena mentioned by **Freeman** represent a finer structure than the attractors in the space of mean spike rates. Yet our pet attractors are neither point attractors nor limit cycles (**Hirsch**). They are driven by noise, as Freeman suggests, and are expressions of stochastic dynamics, with very strange attractors, though mean rates are stable. Perhaps this minor controversy is due to my misinterpretation of the Hebbian token "reverberation." I have used it because I did not read any notion of periodicity in Hebb's usage (**Milner**). A reverberation can be considered an analog of a resonance in an autonomous assembly. I believe it is the complexity of assemblies that distanced Hebb's notion from that of Laureate de Nò (1949).

Perhaps a comment is appropriate about issues of connectivity. I agree that "the packing density [of neurons] is very high" (**Freeman**). Yet, estimates of the probability for a direct contact between two neurons within a range of 1mm vary from 1–10%. My rule of thumb is as follows: In a column of $1mm^2$ parallel to the cortical surface there are about 100,000 cell bodies. On the dendrites of each neuron there are about 20,000 synapses. Adding the estimate that about one half of these contacts come from the near zone and the other half from beyond (Braitenberg & Schutz, 1991, p. 141), one has 10,000 synapses as collaterals in the local module, which leads to a 10% probability for direct collateral connection. In a recent study (Amit & Brunel 1995) we have reconciled the stability of spontaneous activity with the coexistence of structured attractors and have found that this level of connectivity can do the job.

Freeman raises an objection to the idea of the local module, based on his experience of the lack of barriers in the olfactory system. It may be that the olfactory system is different from IT and one is probing two complementary parts of the same elephant. Another possibility is that in the presence of the stimulus activity is observed over wider areas of the same cortical region than during delays in the absence of a stimulus. Both possibilities are accessible to experiment.

On the other hand, I should correct a misconception that my text has created. In the 1mm module there are, during the delay, both active and "silent" neurons for each stimulus. This is the only way in which as many as 100 stimuli could have been coded into the module of 10,000 neurons. Thus, silent neurons are as much part of the code as active ones, as **Freeman** suggests. Yet, the action for all 100 stimuli appears in the $1mm^2$ column, as far as IT cortex is concerned. To say that the relevant "functional unit" is $1cm^2$ rather than $1mm^2$ is to implicate the entire area of the monkey's IT, where recordings have been made. Selective delay activities are not found outside the small module, for each class of stimuli.

There will, of course, be many other modules and areas of cortex engaged in processing the computational consequences of any particular stimulus, but in the absence of the driving stimulus, IT and prefrontal cortex, at least, seem to have chosen a more localized representation for active memory (see discussion of local and global assemblies, which follows).

R8. Local or global assemblies

The contrast is not as acute as implied in the comments (**Pulvermüller & Preissl, Fuster, Lansner & Fransén**). Given the anatomy of cortex, I would argue that if there are persistent reverberations in extended assemblies they are a result of collaboration between reverberations that exist within localized assemblies. In other words, if the synaptic collaterals in a local module of about 1mm in diameter cannot autonomously sustain a reverberation, the afferents from remote areas are unlikely to do the job. The probability of a connection between neurons in such modules is too low to have the required effect. Basically, it all comes back to the fact that synapses are weak, in the sense that a large number of excitatory synaptic inputs are required in order to provoke a postsynaptic spike and thus propagate information.

On the other hand, given a set of local modules that have self-sustaining reverberations, those can easily be excited by afferents from remote areas. This is analogous to the way a resonant state in a system is excited by a very weak driving force. Though I have concentrated on an (I hope unambiguous) $atom^1$ – the local module, I do believe that interactions between remote modules exciting and composing local reverberations are essential for constructing any computational structure of interest at the next level. I have not attempted it, partly for a lack of intellectual courage (**Hucka et al.**), but mostly because I find it exciting that a new element in the bottom-up hierarchy (see sect. R1, metaphor) may be brought under control both experimentally and theoretically. This new element gives an insight not only into collective neural dynamics, but also into the interface between such dynamics and learning.

I interpret the beautiful investigations using imaging to observe large scale phenomena (**Pulvermüller & Preissl**) as complementing the picture of reverberations in local modules. At this stage, neither single cell recordings nor imaging can resolve the bridge between the local and the global. I would suggest that one of the most urgent projects in neurophysiology today is to establish the correlation between the two. What we know about local and global connectivities, about plausible values for synaptic efficacies, and about levels of spontaneous activity in the wide cortex convinces me that local reverberations are here to stay as a dynamic building block. They, in themselves, probably do not carry much cognitive import. The next step may be correctly perceived in the concluding sentence of Pulvermüller & Preissl's commentary.

But whenever there is the possibility that disjoint assem-

blies participate in a related task, which is unavoidable, the problem of "binding" rears its head (**Petitot, Bienenstock & Geman**). Gray and Singer (1987) and others have put forth the very attractive idea that it could be solved by synchronization, as observed in the primary visual cortex of anesthetized cats. To me this issue seems still largely open. First, one would expect such a fundamental mechanism to be more ubiquitous than observed. More important, no viable cortical reading mechanism sensitive to synchrony has been put forth. Finally, I would expect that the proposal of the mechanism of binding to be accompanied by a mechanism of unbinding, which I suspect is an even harder nut to crack. It may appear simple if the two different cortical areas have separate receptive fields. But similar mechanisms must also deal with the binding and unbinding of color from shape, for example.

R9. Computational time and attractor dynamics

Another question that projects beyond the context of the target article concerns the amount of time required for a network to reach an attractor compared to the observed psychophysical reaction or perception times (**Hoffman**). I cannot offer a definitive answer. Some experiments do give the impression that things may be taking place on time scales of a few tens of milliseconds. Given that rates in attractors may be as low as 20 spikes/sec, a neuron will on average emit about 2 spikes in the time in which an entire process is supposed to take place. Can this be sufficient time to reach attractors?

Several things can be said in this regard:

1. The target article did not intend to argue that every neural computation has to pass via attractors. Fast reflexes and processes in early attention probably have little to do with attractors. They may be feedforward computations of sorts. In this perspective, attractors would reenter the picture if some re-elaboration of the stimulus is to take place long after it has disappeared.

2. There is a common misconception, generated by an over-orthodox interpretation of the Hopfield (1982) toy model, namely, that interspike intervals are identified with neural potential updates at discrete time intervals (**Hoffman**). This was not the intention of the gospel. It was merely a schematization for the sake of creating a parallel with certain well known models of physics. Networks of realistic neurons do not operate like that. The neurons in the assembly have, at any given moment, a wide distribution of depolarizations and when the stimulus arrives, some start emitting spikes almost instantly, beginning the process of relaxation into an attractor, spontaneous or structured, in a relatively short time. In the models we have considered (Amit & Tsodyks 1991) such times are 50–80 msec. To repeat, the above does not imply that attractors necessarily underlie all brain processing, but only that arguments based on time constraints are not sufficient to disqualify them.

3. The alternatives are not less problematic. To suggest that the internal temporal structure of a very limited number of spikes is the key to fast computation, is to ignore (a) the level of noise and fluctuation in the system; (b) the fact that to date we have no viable proposal as to how such sequences are learned; and (c) the absence of a viable candidate that can read such fine code in equally short time.

R10. Attractors and cognition

It appears to me that the comment of **van der Velde** indicates a promising extension of the attractor picture in the direction of cognitive psychology. I say this with some trepidation, because I feel out of my element with "regular" and "irregular" computations. It is surely true that attractors can be used to create associations between assemblies via Hebbian learning and open the way for modeling the conditioned reflex as well as allowing some type of binding.

I believe, however, that the picture is richer. First, note that the creation of the Miyashita correlated attractors is already a "computation" based on intermediate results. In order to generate these correlations the assembly first had to generate a synaptic structure that could subsequently support a reverberation. These uncorrelated reverberations, in turn, are the carriers of information about one stimulus until the arrival of the next. As a consequence, the representations of the stimuli, which are at first uncorrelated, change significantly and end up carrying the context.

Is this a Turing machine with or without a tape? Is it a regular or irregular computation? I do not know the answer. The application of the Turing scheme to describe neural computation in a real brain is not completely obvious. Where is the program in the brain? And what is a memory? If a program exists, its mere definition will, in my view, be a revolutionary step toward the understanding of brain function. The mere demonstration of the existence of a program is beyond what seems imaginable. And all we have are noisy neurons and unreliable synapses.[2]

It should be emphasized that none of the internal representations carried by the attractors is completely stable. This can be deduced indirectly from experiment: The fact that reverberations corresponding to stimuli remote in the training sequence are uncorrelated indicates that at some stage in the training process all reverberations were uncorrelated and that those uncorrelated attractors have been modified during training. In fact, the transformation of uncorrelated attractors into correlated ones, caused by contextual proximity, may be one "dynamic" way in which elements are composed "into a relationally specific composite" (**Bienenstock & Geman**). Also in material devices which embed what has been learned from the reverberating modules (Badoni et al. 1994), one observes the slow drift of attractors under the impact of variations in the input flux of stimuli. Should we consider this collective plasticity of the synaptic matrix of the assembly as a memory or as the modification of a program?

This connects directly to the possibility that attractors are "cognitive concepts" (**Raijmakers & Molenaar**). I have not proposed that they are and do not think they are. The opening metaphor (sect. R1) is of attractors as spikes, rather than attractors as brains. This should underline the fact that the attitude informing the attractor approach is diametrically opposed to that informing thinking, reading, and talking feed-forward networks. The distance between attractors and cognition is to be filled by bold speculations. Some imaginative speculations have been put forward (see **Hoffman,** and the review by Ruppin 1995). In my view, these are still too naive and put too much emphasis on generating cognitive interpretations based on single networks and on properties that they possess only at a very specific stage of modeling. More promising perhaps is the

approach to semantic and cognitive categorization considered as an outcome of the interaction of several attractor networks (Lauro Grotto et al. 1994).

This all boils down to the fact that I have no answer to the important question of whether an attractor network can perform properly in a discrimination-shift test. In fact, as with many aspects of this modeling scheme, the question concerns not so much whether a certain cognitive-behavioral expectation can be confirmed or falsified but rather how the elements provided should be used to express the input-output relation. I would venture to say more. Feed-forward thinking has reduced much of mental function to mechanical input-output relations, which are usually confirmed by 2–3 layer networks. My hope is that the new bottom-up element of assemblies learning, modifying and carrying attractors may allow a fresh look at the cognitive behavioral questions themselves. The discrimination-shift test is a timely, exciting phenomenon to try to model with attractor networks and then put to the test.

Much criticism (**Dalenoort & de Vries, Bienenstock & Geman**) seems related to the fact that I have restricted myself to the minimal reverberating cortical module and avoided discussing assemblies of assemblies at the risk and the cost of foregoing complex cognitive and behavioral phenomena. Or, as they put it, I have made a physical picture of attractors that does not bring out the intricacies of human cognition. Or, still better I have not chosen to play where the "action" is. I agree.

R11. Attractors and monkey cognition

The observation by Miyashita (1988) (highlighted by **Bienenstock & Geman**) that the monkey performs as well for "new" stimuli as for known ones, underlines how little can be said at this stage about the cognitive significance of the selective reverberations. A minimalist response would refer back to the metaphor in section R1 and to the fact that selectively reverberating assemblies are a fact of life, just as spikes are. Moreover, they are related to learning, and via learning create context-sensitive representations. Strangely enough, the same criticism is not raised against spikes themselves. For some reason, spikes are accepted as legitimate building blocks in every one of the sophisticated mechanisms proposed in the commentary. What attractor representations are used for by the brain is at present a subject for speculation, or for some additional experiments.

But one can say a little more without departing too much from solid evidence:

How does the monkey manage to recognize "new" stimuli? I can only speculate. There may indeed be a place, outside of area AVT, the anterior ventral temporal cortex, where stimuli create attractors rapidly. See Bayliss & Rolls 1987. The cost, I believe, is in not being able to handle too many of them concurrently. That is, if a large number of "new" stimuli is shown before a match, the fast learning mechanism will fail. Slowly learned attractors, learned by many presentations, can be accumulated.

It is possible that "new" stimuli are recognized by something more than Hebbian attractors (**Bienenstock & Geman**), but it is not necessarily the case, especially if one keeps in mind that information has to be carried in a robust way for a long time after the stimulus has been turned off.

I fail to see what underlies the statement that binding cannot be done by coactivating populations of neurons for the respective pieces (**Bienenstock & Geman**). I rather think it can, though I have only circumstantial evidence that it does (see discussion in sect. R11). At the risk of being repetitious, I would paraphrase the above objection using spikes instead of spiking assemblies. This does not seem to be strongly objectionable. Is it just because spikes have been used longer in the process of training *we* have been subjected to? Each of us can have his pet speculations, as long as contact with experiment remains tenuous. I tried to remain on more solid ground.

R12. Disclaimers

My warning against premature speculation was not intended to deter anyone, least of all Hebb (**Hucka et al.**), from speculating. In a sense, it was an attempt to draw a productive lesson from self-criticism. Everyone of us dabbling in models is tempted to extrapolate and speculate from what we know, at a rather low level, to surface cognitive phenomena. I have been there too. And then when the Miyashita correlations were discovered, they came as a shock. They indicate that the border between what is computation and what is being represented may be very far from anything we have been imagining. To me, these experiments are not only a demonstration of the poverty of our imagination. They convey a very strong message, namely, that there are experimental ways of shedding light on these questions. My warning was intended to highlight the above observation.

I admit that Hebb made an ingenious speculation about the connection of learning to the formation of reverberating assemblies, and in the absence of essentially any experimental evidence. At this level, however, the speculation is not of the computational kind. It is again a construction of second level elements from spikes, as suggested in the metaphor (sect. R1). In that sense I do not feel guilty of usurping historical credit. I believe that essentially he had it all. That is, for example, why I would not dare to answer **Petitot**'s question about how I think attractor syntax takes place. I made an attempt in this direction in my book (Amit 1989), and I regret that this particular proposal is in print. Perhaps there is no choice but to make such leaps. But right now, experiment offers such rich data that we may gain more by testing our imagination against it.

I also espouse Hebb's point of view that the introspection leading to assemblies and reverberations, as an expression of the process of learning and/or innate synaptic organization, should have strong implications for psychology. The construction of a rich metaphor for psychology, based on attractors, their creation, their mutation, their interactions, their possible structure and semantics, and so on, is still awaiting the courageous adventurer who will ignore my warnings. I would also consider the evolutionary attractor scenario proposed by **Krakauer & Houston** to fall in this category. But building metaphors for psychology or social behavior differs, in my view, from proposing explicit computational schemes. The Miyashita picture, if captured correctly in the target article, should lead to a moment of reflection.

Milner objects to my assertion that (1) "attractor representation is the only dynamic that can distinguish naturally between familiar and unfamiliar stimuli" and that (2) "rec-

ognition" is simple. If the word "only" is removed from the first assertion, it corresponds to what I said and I suppose answers the objection. Concerning the second, I do not know how recognition is affected, and I hope I have not implied I did. The sense in which "familiar" was used was very narrow and technical, and has nothing to do with cognition. A stimulus is familiar if it was embedded in the synaptic structure during learning.

The little motivating story about the word that hovers for hours and days should be taken with a grain of salt (**Wright**). Do I want it to be believed? It is definitely not part of the restricted context for which there is empirical evidence. Yet, mechanically, it is possible. The question then is, as Wright points out, who protects it against ongoing life? I have no answer, but that does not make the poetic hypothesis wrong. There are many other pertinent questions for which I (we) do not have answers. Some are raised by Wright, such as the description of the dynamics of the magic moment in which a cognitive task has been accomplished. There is no reason why, among the missing pieces of the puzzle, some screening mechanism could not exist for protecting the hovering word. Perhaps the experiments of Miller et al. (1993), in which a monkey is able to perform DMS (delay match to sample) trials despite intervening stimuli, are closing in on such a mechanism. Perhaps also, when we have a better idea of the relation between EEG, MRI, and spike recording, we will be able to test whether such long hovering times are plausible. To conclude this interlude I would like to put the shoe on the other foot: what are the plausible alternatives?

R13. Philosophy and methodology

I would dispose of the *philosophical* issues in short order, not for lack of respect, but because they really belong in a different context.

1. Philosophical issues should not intervene in the construction of a scientific theory. They never have. In the construction of a scientific theory one works by ostension. No philosophical issue can be resolved in the process. To quote de Santillana & von Dechend (1969): "The problem of number remains to perplex us, and from it all of metaphysics was born." And yet imagine if the use of numbers had to wait for the resolution of the perplexity.

2. I do not claim to have solved either the "problem of representation" or the "psychophysical" problem (**Edelman**). I had no such objective in mind. I have stumbled upon an experimental fact with a theoretical "correlate" which simply *is* a representation. Even philosophy must appreciate the fact that for us to be puzzled about representations, representations must exist in some naive sense to be so named. So here are some. And the very limited and intuitive sense in which the term is used is carefully delineated in the target article.

3. To Plato's facetious comment on the "flock of birds" (**Edelman**) I would add, as a footnote, that had one discovered a selective flock of birds in peoples' skulls for different stimuli, the fact would have been considered of great scientific interest, even if Plato's perplexities would not have been put to rest.

4. I promise not to solve any problem of cognitive philosophy. I would hope that some of the phenomena, interpreted in the particular way suggested in the target article, may serve as metaphors for the production of some new metaphysics by cognitive philosophers. But this is a side concern.

5. Finally, Popper, critical tests, and refutations (**Ahissar**). When in the history of science has this been a productive methodology? What has ever been honestly refuted? I find it quite remarkable that the only community among the natural sciences in which Popperian banners have followers is biology. A community with achievements as wonderful as neurobiology should free itself from such slogans. (a) Naive observation of scientific practice suggests that a theory and/or a model should be judged by the conceptual framework it offers; the range of seemingly unrelated data it connects; its economic use of *ad hoc* assumptions; its beauty; the ease with which it can be extended to new problems; the range of predictions it makes; and most important, by the quality of the competition. (b) What if some predictions are not confirmed experimentally? It may teach us something about the quality of the experiment. Or, it may be corrected by a modification of the model or the theory, which should then be resubjected to the algorithm described under (a) above. Such modifications of the theory are not a sign of weakness or lack of integrity. They are part of a valid process of learning about the mysterious system we are trying to study.

I regret to belabor these issues, which have been quite clear to Popper's followers (e.g., Lakatos and Kuhn).

And yet, if a neurophysiological experiment should demonstrate that variations in the stimulus (leading to small variations in the afferents at the reverberating assembly) provoke a significantly different delay activity distribution, serious rethinking of the picture will have to take place.

NOTE

1. I used this expression in this response before I became aware of its pejorative sense (**Bienenstock & Geman**).

2. Recent work of van der Velde, including an article submitted to *BBS*, proposes very stimulating answers to these questions.

References

Letters *a* and *r* appearing before authors' initials refer to target article and response respectively.

Abeles, M. (1991) *Corticonics: Neuronal circuits of the cerebral cortex.* Cambridge University Press. [JP]

Amit, D. J. (1989) *Modeling brain function.* Cambridge University Press. [arDJA, JP, FVDV]

(1993) In defense of single electrode recordings. *Network* 3:385. [aDJA]

(1994) Persistent delay activity in cortex: A Galilean phase in neurophysiology? *Network: Computation in Neural Systems* 5:429–36. [MWH]

Amit, D. J. & Brunel, N. (1994) Learning internal representations in an attractor neural network with analogue neurons. *Network* 6(3):359 [rDJA]

(1995) Global spontaneous activity and local structured (learned) activity in cortex. Submitted. [rDJA]

Amit, D. J., Brunel, N. & Tsodyks, M. V. (1994) Correlations of Hebbian reverberations. *Journal of Neuroscience.* 14:6435. [arDJA]

Amit, D. J. & Fusi, S. (1994) Learning in neural networks with material synapses. *Neural Computation* 6:957. [rDJA]

Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985) Spin-glass models of neural networks. *Physiological Reviews* A32:1007. [rDJA]

Amit, D. J. & Tsodyks, M. V. (1991a) Quantitative study of attractor neural network retrieving at low spike rates: 1. Substrate—spikes, rates and neuronal gain. *Network* 2:259. [arDJA, JJW]

(1991b) Quantative study of attractor neural network retrieving at low spike rates: 2. Low-rate retrieval in symmetric networks. *Network* 2:275. [arDJA, JJW]

(1992) Effective neurons and attractor neural networks in cortical environment. *Network* 3:121–37. [EA]

Amsel, A. & Rashotte, M. E. (1984) Mechanisms of adaptive behavior: *Clark*

Hull's theoretical papers, with commentary. Columbia University Press. [FVDV]

Anderson, J. A., Silverstein, J. W., Ritz, S. A. & Jones, R. S. (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review* 84:413–51. [MH]

Anderson, M. (1994) Sexual selection. *Monographs in behavior and ecology,* ed. J. R. Krebs & T. Clutton-Brock. Princeton University Press. [DCK]

Anisfeld, M. & Knapp, M. (1968) Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology* 77:171. [aDJA]

Anson, J. G. & Bird, Y. N. (1993) Neuromotor programming: Bilateral and unilateral effects on simple reaction time. *Human Movement Science* 12:37–50. [FP]

Arak, A. (1988) Callers and satellites in the natterjack toad: Evolutionary stable decision rules. *Animal Behavior* 36:416–32. [DCK]

Artola, A. & Singer, W. (1933) Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends in Neurosciences* 16(11):480–87. [EC]

Atick, J. J. (1992) Could information theory provide an ecological theory of sensory processing? *Network* 3:213. · [rDJA]

Atwood, H. L. & Nguyen, P. V. (1990) Physiological properties of crustacean motor neurons and the alteration of these properties. In: *Frontiers in crustacean neurobiology.* Birkhauser Verlag. [EC]

Badoni, D., Bertazzoni, S., Buglioni, S., Salina, G., Amit, D. J. & Fusi, S. (1995) Electronic implementation of an analog attractor neural network with stochastic learning. *Network* 6:125. [rDJA]

Baudry, M. & Davis, J. L., eds. (1991) *Long-term potentiation: A debate of current issues.* MIT Press. [FVDV]

Baylis, G. C. & Rolls, E. T. (1987) Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Experimental Brain Research* 65:614. [rDJA]

Bialek, W. & Rieke, F. (1992) Reliability and information transmission in spiking neurons. *Trends in Neuroscience* 15:428–33. [REH]

Bienenstock, E. (1994) A model of neocortex. Technical report, Division of Applied Mathematics, Brown University [JP]

Birbaumer, N., Elbert, T., Canavan, A. G. M. & Rockstroh, B. (1990) Slow potentials of the cerebral cortex and behavior. Physiological Reviews 70:1–41. [FP]

Braitenberg, V. (1978) Cell assemblies in the cerebral cortex. In: *Theoretical approaches to complex systems* [Lecture notes in biomathematics, vol. 21], ed. R. Heim & G. Palm. Springer. [FP]

(1984) *Vehicles: Experiments in Synthetic Psychology.* MIT Press. [MH]

Braitenberg, V. & Schutz, A. (1991) *Anatomy of the cortex: Statistics and geometry.* Springer-Verlag. [arDJA, WJF, FP]

Bridgeman, B., Van der Heijden, A. H. C & Velichkovsky (1994) A theory of visual stability across saccadic eye movements. *Behavioral and Brain Sciences* 17:247–92. [DCB]

Brunel, N. (in press) Stochastic learning of temporal correlations between stimuli in attractor neural networks. *Neural Computation.* [arDJA]

Buhmann, J., Divko, R. & Schulten, K. (1989) Associative memory with high information content. *Physical Review* A39:2689. [aDJA]

Burns, B. D. (1951) Some properties of isolated cerebral cortex in the unanesthesized cat. *Journal of Physiology* 112:156–75. [MH]

Burr, D. C., Holt, J., Johnstone, J. R. & Ross, J. (1982) Selective depression of motion selectivity during saccades. *Journal of Physiology (London)* 333:1–15. [DCB]

Burr, D. C. & Ross, J. (1982) Contrast sensitivity at high velocities *Vision Research* 23:3567–69. [DCB]

Burr, D. C., Morrone, M. C. & Ross, J. (1994) Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature* 371:511–13. [DCB]

Buss, A. H. (1956) Reversal and nonreversal shifts in concept formation with partial reinforcement eliminated. *Journal of Experimental Psychology* 52:162–66. [MEJR]

Campbell, F. W. & Wurtz, R. H. (1978) Saccadic omission: Why we do not see a greyout during a saccadic eye movement. *Vision Research* 18:1297–1303. [DCB]

Chekaluk, E. (1994) Is there a role for extraretinal factors in the maintenance of stability in a structured environment? *Behavior and Brain Sciences* 17:92. [DCB]

Chown, E. (1994) Consolidation and learning: A connectionist model of human credit assignment. Doctoral dissertation, University of Michigan. [EC]

Cugliandolo, L. (1994) Correlated attractors from uncorrelated stimuli. *Neural Computation* 6:220. [aDJA]

Cummins, R. (1989). Meaning and mental representation. MIT Press. [SE]

Daido, H. (1990) Intrinsic fluctuations and a phase transition in a class of large populations of interacting oscillators. *Journal of Statistical Physics* 60:753–800 [JP]

Dalenoort, G. J. (1982) In search of the conditions for the genesis of cell asemblies: A study—in self-organization. *Social Biol. Struct.* 5:161–87. [GJD]

(1990) Towards a general theory of representation *Psychological Research* 52:229–37. [GJD]

Damasio, A. R. & Damasio, H. (1991) Cortical systems underlying knowledge retrieval: Evidence from human lesion studies (background manuscript for the Dahlem Conference on Exploring Brain Function: Models in Neuroscience, Berlin). [aDJA]

Darwin, C. (1871) *The descent of man, and selection in relation to sex.* Murray. [DCK]

De Santillana, G. & von Dechend, H. (1969) *Hamlet's mill.* Gambit. [rDJA]

Dong, D. W. & Hopfield, J. J. (1992) Dynamic properties of neural networks with adapting synapses. *Network* 3:267. [rDJA]

Doyon, B., Cessac, B., Quoy, M., Samuelides, M. (1993) Chaos in neural networks with random connectivity. *International Journal of Bifurcation and Chaos* [JP]

Edelman, G. M. (1987) *Neural Darwinism: The theory of neuronal group selection.* Basic Books. [MH]

Edelman, S. (1995) Similarity and the chorus of prototypes. *Minds and machines* 5:45–68. [SE]

Emery, J. D. & Freeman, W. J. (1969) Pattern analysis of cortical evoked potential parameters during attention changes. *Physiology & Behavior* 4:67–77. [WJF]

Engel, A. K., König, P., Kreiter, A., Schillen, T. & Singer, W. (1992) Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neuroscience* 15(6):218–26. [WK, JP]

Enquist, M. & Arak, A. (1993) Selection of exaggerated male traits by female aesthetic senses. *Nature* 361:446–48. [DCK]

Field, D. J. (1987) Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A* 4:2379–94. [DCB]

Fodor, J. A. (1975) *The language of thought.* Crowell. [JPR]

Fodor, J. A. & McLaughlin, B.P. (1990) Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35:183–204.

Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. In: *Connections and symbols,* ed. S. Pinker & J. Mehler. MIT Press. [JP, FVDV]

Fransén, E. & Lansner, A. (1994) Low spiking rates in a network with overlapping assemblies. In: *The neurobiology of computation: Proceedings of the annual computational neuroscience meeting,* ed. J. M. Bower. Kluwer. [AL]

Fransén, E., Lansner, A. & Liljenström, H. (1992) A model of cortical associative memory based on Hebbian cell assemblies. In: *Computation and neural systems,* ed. F. Eeckman & J. M. Bower. Kluwer. [AL]

Freeman, W. J. (1967) Analysis of function of cerebral cortex by use of control systems theory. *Logistics Review* 3:5–40. [WJF]

(1968) Analog simulation of prepyriform cortex in the cat. *Mathematical BioScience* 2:181–90. [WJF]

(1975) *Mass action in the nervous system.* Academic Press. [WJF, MH]

(1979) Nonlinear gain mediating cortical stimulus-response relations. *Biological Cybernetics* 33:237–47. [WJF]

(1987) Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biological Cybernetics* 56:139–43. [WJF, MWH]

(1992) Tutorial in neurobiology. *International Journal of Bifurcation & Chaos* 2:451–82. [WJF]

(1993) Valium, histamine, and neural networks. *Biological Psychiatry* 34:1–2. [WJF]

(1995) *Societies of brains: A study in the neuroscience of love and hate.* Erlbaum. [WJF]

Freeman, W. J. & Baird, B. (1987) Relation of olfactory EEG to behavior: Spatial analysis. *Behavioral Neuroscience* 101:393–408. [MWH]

Freeman, W. J. & Barrie, J. M. (1994) Chaotic oscillations and the genesis of meaning in cerebral cortex. In: *Temporal Coding in the Brain,* ed. G. Buzsaki, R. Llinàs, W. Singer, A. Berthoz & Y. Christen. Springer-Verlag. [WJF]

Freeman, W. J. & Skarda, C. A. (1990) Chaotic dynamics versus representation. *Behavioral and Brain Sciences* 13:167–68. [REH]

Freeman, W. J. & Viana di Prisco, G. (1986) EEG spatial pattern differences with discriminated odors manifest chaotic and limit cycle attractors in olfactory bulb of rabbits. In: *Proceedings of the First Trieste Meeting on Brain Theory,* ed. G. Palm & A. Aertsen. Springer-Verlag. [MWH]

Fujita, I., Tanaka, K., Ito, M. & Cheng, K. (1992) Columns for visual features of objects in monkey inferotemporal cortex. *Nature* 360:343–46. [SE]

Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61:331. [rDJA]

Fuster, J. M. (1973) Unit activity in prefrontal cortex during delayed-response

performance: Neuronal correlates of transient memory. *Journal of Neurophysiology* 36:61. [aDJA]
(1989) The prefrontal cortex: *Anatomy, physiology, and neuropsychology of the frontal lobe,* 2d ed. Raven. [JMF]
(1990) Inferotemporal units in selective visual attention and short-term memory. *Journal of Neurophysiology* 64:681–97. [JMF]
(1994) *Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate.* MIT Press. [FP]
(1995) Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate. MIT Press. [JMF]
Fuster, J. M., Bauer, R. H. & Jervey, J. P. (1982) Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. *Experimental Neurology* 77:679–94. [JMF]
(1985) Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research* 330:299–307. [JMF]
Fuster, J. M. & Jervey, J. (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science* 212:952–55. [JMF]
(1982) Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *Journal of Neuroscience* 2:361–75. [JMF]
Gardner, E. (1987) Maximum storage capacity in neural networks. *Europhysics Letters* 4:481. [rDJA]
Gerstein, G. L., Bedenbaugh, P. & Aertsen, A. M. H. J. (1989) Neuronal assemblies. *IEEE Transactions on Biomedical Engineering* 36:4–14. [FP]
Goldman-Rakic, P. S. (1990) Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. In: *Progress in Brain Research,* vol. 85, ed. H. B. M. Uylings, C. G. Van Eden, J. P. C. De Bruin, M. A. Corner & M. G. P. Feenstra. [MH]
(1992) Working memory and the mind. *Scientific American* 267:110–17. [JPR]
Grant, B. R. (1985) Selection on bill characters in a population of Darwin's finches, *Geospiza fortis,* on Isla Genovesa, Galápagos. *Evolution* 39:523–32. [DCK]
Gray, C. & Singer, W. (1987) Stimulus-dependent neuronal oscillations in the cat visual cortex area 17. *Neuroscience* (Suppl.) 22:434. [WK]
Griniasty, M., Tsodyks, M. V. & Amit, D. J. (1993) Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation* 5:1. [arDJA]
Grossberg, S. (1987) Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11:23–63. [MH]
Haidarliu, S., Shulz, D. & Ahissar, E. (1995) A multielectrode array for combined microiontophoresis and multiple single-unit recordings. *Journal of Neuroscience Methods* 56:125–31. [EA]
Harnad, S. (1990) The symbol grounding problem. *Physica D* 42:335–46. [SE]
Harrow, M. & Friedman, G. B. (1958) Comparing reversal and nonreversal shifts in concept formation with partial reinforcement control. *Journal of Experimental Psychology* 55:592–98. [MEJR]
Hebb, D. O. (1949) *The organisation of behaviour.* Wiley. [aDJA, GJD, PMM, FP, JPR, JJW]
Hebb, D. O. & Donderi, D. C. (1987) *Textbook of psychology,* 4th ed. Erlbaum. [aDJA, JPR]
Hilgard, E. R. & Marquis, D. G. (1940) Conditioning and learning. Appleton-Century. [PMM]
Hinton, G. & Sejnowski, T. (1986) Learning and relearning in Boltzmann machines. In: *Parallel Distributed Processing,* vol. 1, ed. D. E. Rumelhart & J. L. McClelland. MIT Press. [MH]
Hintzman, D. L. (1993) Twenty-five years of learning and memory: Was the cognitive revolution a mistake? In: *Attention and performance 14,* ed. D. E. Meyer & S. Kornblum. MIT Press. [FVDV]
Hirsch, M. W. (1995) Realism in mathematics. *Bulletin of the American Mathematical Society* 32:137–47. [MWH]
Hoffman, R. E. (1987) Computer simulations of neural information processing and the schizophrenia-mania dichotomy. *Archives of General Psychiatry* 44:178. [aDJA]
Holcomb, P. J. & Neville, H. J. (1990) Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes* 5:281–312. [FP]
Hopfield, J. J. (1982) Neural networks and physical systems with emergent selective computational abilities. *Proceedings of the National Academy of Sciences USA* 79:2554–58. [aDJA, GJD, REH, JPR]
(1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences USA* 81:3088–92. [REH]
Ilg, U. J. & Hoffmann, K.-P. (1993) Motion perception during saccades. *Vision Research* 33:211–20. [DCB]
Ito, M. (1992) Posttetanic depression. In: *Enclyclopedia of learning and memory,* ed. L. R. Squire. Macmillan. [EC]
James (1902/1987) *Pragmatism: A new name for an old way of thinking.* The Library of America. [DCK]
Kanerva, P. (1988) *Sparse distributed memory.* MIT Press. [MH]
Kaplan, S. & Kaplan, R. (1982) *Cognition and Environment.* Praeger. Republished by Ulrich's, Ann Arbor, Michigan, 1989. [MH]
Kaplan, S., Sonntag, M. & Chown, E. (1991) Tracing recurrent activity in cognitive elements (TRACE): A model of temporal dynamics in a cell assembly. *Connection Science* 3:179–206. [EC]
Kaplan, S., Weaver, M. & French, R. (1990) Active symbols and internal models: Towards a cognitive connectionism. *AI & Society* 4:51–71. [MH]
Kelleher, R. T. (1956) Discrimination learning as function of reversal and nonreversal shifts. *Journal of Experimental Psychology* 51(6):379–84. [MEJR]
Kendler, H. H. & Kendler, T. S. (1962) Vertical and horizontal processes in problem solving. *Psychological Review* 69(1):1–6. [MEJR]
(1975) From discrimination learning to cognitive development: A neobehavioristic odyssey. In: *Handbook of learning and cognitive processes,* ed. W. K. Estes. Erlbaum. [MEJR]
Kendler, T. S. & D'Amato, M. F. (1955) A comparison of reversal shifts and nonreversal shifts in human concept formation behavior. *Journal of Experimental Psychology* 49:165–74. [MEJR]
Kennedy, M. B. (1989) Regulation of neuronal function by calcium. *Trends in Neurosciences* 12:417–20. [PMM]
Klimesch, W. (in press) Memory processes described as brain oscillations in the theta and alpha band. *Psycoloquy* 95.6.55.memory-brain.1.klimesch. [WK]
Kohonen, T. (1984) *Self-organization and associative memory.* Springer. [JPR]
Kopell, N. & Ermentrout, G. B. (1990) Phase transitions and other phenomena in chains of coupled oscillators. *SIAM Journal of Applied Mathematics* 50:1014–52. [JP]
Krakauer, D. C. & Johnstone, R. U. (in press) The evolution and honesty in animal communication: A model using artificial neural networks. *Phil. Trans. Roy. Soc. B.* [DCK]
Kruger, J., ed. (1991) *Neuronal cooperativity.* Springer-Verlag. [EA]
Kruschke, J. K. (1992) ALCOVE: An examplar-based connectionist model of category learning. *Psychological Review* 99(1):22–44. [MEJR]
Kuhl, P. K., Williams, K. A. & Meltzoff, A. N. (1991) Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance* 17:829–40. [JPR]
Kuramoto, Y. & Nishikawa, I. (1987) Statistical macrodynamics of large dynamical systems: Case of a phase transition in oscillator communities. *Journal of Statistical Physics* 49: 569–605 [JP]
Lachter, J. & Bever, T. G. (1988) The relationship between linguistic structure and associative theories of language learning: A constructive critique of some connectionist learning models. *Cognition* 28(1–2):195–247. [MH]
Langton, C. G. (1990) Computation to the edge of chaos: Phase transitions and emergent computation. *Physica* 42D:12.
Lansner, A. (1982) *Information processing in a network of model neurons: A computer simulation study* (Technical Report No. TRITA-NA-8211). Stockholm, Sweden: NADA, Royal Institute of Technology. [AL]
Lansner, A. & Fransén, E. (1992) Modeling Hebbian cell assemblies comprised of cortical neurons. *Network* 3:105–119. [AL]
(1994) Improving the realism of attractor models by using cortical columns as functional units. In: *The neurobiology of computation: Proceedings of the annual computational neuroscience meeting,* ed. J. M. Bower. Kluwer. [AL]
Lansner, A. & Liljenström, H. (1994) Computer models of the brain—How far can they take us. *Journal of Theoretical Biology* 171:61–73. [AL]
Lashley, K. S. (1951) In search of the engram. *Symposia of the Society of Experimental Biology* 4:454–82. [GJD]
Laurent, G. & Davidowitz, H. (1994) Encoding of olfactory information with oscillating neural assemblies. *Science* 265:1872–75. [MH]
Lauro Grotto, R., Reich, S. & Virasoro, A. M. (1994) The computational role of conscious processing in a model of semantic memory, ed. M. Ito. *Proceedings of the IIAS Symposium on Cognition, Computation and Consciousness,* Kyoto, August 31, 1994. [rDJA]
Lem, S. (1985) *Star diaries.* Harcourt Brace Jovanovich. [SE]
Liley, D. T. J. & Wright, J. J. (1994) Intracortical connectivity of pyramidal and stellate cells: Estimates of synaptic densities and coupling symmetry. *Network* 5:175–79. [JJW]
Llinás, R. & Ribary, U. (1993) Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences USA* 90:2078–81. [PMM]
Locke, J. (1690) *An essay concerning human understanding.* Available electronically on the Internet at URL gopher://gopher.vt.edu:10010/02/116/3. [SE]
Lorente de Nó (1949) In: *Physiology of the nervous system,* ed. J.F. Fulton. Oxford University Press. [aDJA, JPR]
Lutzenberger, W., Pulvermüller, F. & Birbaumer, N. (1994) Words and pseudowords elicit distinct patterns of 30-Hz activity in humans. *Neuroscience Letters* 176:115–118. [FP]

Lutzenberger, W., Pulvermüller, F., Elbert, T. & Birbaumer, N. (1995) Local 40-Hz activity in human cortex induced by visual stimulation. *Neuroscience Letters* 183:139–42. [FP]

MacGregor, R. J. & McMullen, T. (1978) Computer simulation of diffusely connected neuronal populations. *Biological Cybernetics* 28:12–127. [AL]

Mackay, D. M. (1970) Elevation of visual threshold by displacement of visual images. *Nature* 225:90–92. [DCB]

Macknik, S. L., Bridgeman, B. & Switkes, E. (1991) Saccadic suppression of displacement at isoluminance. *Investigative Ophthalmology and Visual Science* [Suppl.] 32:899. [DCB]

Maddy, P. (1993) *Realism in mathematics.* Oxford University Press. [MWH]

Magleby, K. L. (1987) Short-term changes in synaptic efficiency. In: *Synaptic function,* ed. G. M. Edelman, W. E. Gall & W. M. Cowan. Wiley. [EC]

Mangler, G. (1975) Consciousness: Respectable, useful, and probably necessary. In: *Information processing and cognitive psychology,* ed. R. L. Solso. Erlbaum. [EC]

Mason, A., Nicoll, A. & Stratford, K. (1991) Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro. *Journal of Neuroscience* 11:72. [aDJA]

Matin, E. (1974) Saccadic suppression: A review and an analysis. *Psychological Bulletin* 81:899–917. [DCB]

McGaugh, J. L. & Herz, M. J. (1972) *Memory consolidation.* Albion. [JPR]

McKenna, T. M., Ashe, J. H. & Weinberger, N. M. (1989) Cholinergic modulation of frequency receptive fields in auditory cortex: I. Frequency-specific effects of muscarinic agonists. *Synapse* 4:30–43. [EA]

McNaughton, B. L., Barnes, C. A. & Andersen, P. (1981) Synaptic efficacy and EPSP summation in granule cells of rat fascia dentata in vitro. *Journal of Neurophysiology* 46:952. [aDJA]

Metherate, R. & Weinberger, N. M. (1989) Acetylcholine produces stimulus-specific receptive field alterations in cat auditory cortex. *Brain Research* 480:372–77. [EA]

Miller, E. K. & Desimone, R. (1994) *Dual mechanism for short-term memory in inferior temporal cortex.* NIMH reprint. [rDJA]

Miller, E. K., Li, L. & Desimone, R. (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* 254:1377–79. [JPR]

(1993) Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience* 13(4):1460–78. [MH, rDJA]

Miller, R. R. & Marlin, N. A. (1984) The physiology and semantics of consolidation. In: Memory consolidation: *Psychobiology of cognition,* ed. H. Weingartner & E. S. Parker. Erlbaum. [EC]

Milner, P. M. (1957) The cell assembly: Mk. II. *Psychological Review* 64:242–52. [PMM]

(1989) A cell assembly theory of hippocampal amnesia. *Neuropsychologia* 27:23–30. [PMM]

Miyashita, Y. (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335:817–20. [aDJA, EA, GJD, MH, REH, PMM, MEJR]

Miyashita, (1988) *Nature* 335:819.

Miyashita, Y. & Chang, H. S. (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331:68–70. [aDJA, EA, GJD, MH, DCK, PMM, MEJR]

Mohr, B., Pulvermüller, F., Rayman, J. & Zaidel, E. (1994) Interhemispheric cooperation during lexical processing is mediated by the corpus callosum: Evidence from the split-brain. *Neuroscience Letters* 181:17–21. [FP]

Mohr, B., Pulvermüller, F. & Zaidel, E. (1994) Lexical decision after left, right and bilateral presentation of content words, function words and non-words: Evidence for interhemispheric interaction. *Neuropsychologia* 32:105–124. [FP]

Morita, M. (1992) A neural network model of the dynamics of a short-term memory system in the temporal cortex. *Systems and Computers in Japan* 23(4):14–24. [MM]

(1993) Associative memory with nonmonotone dynamics. *Neural Networks* 6:115–26. [MM]

(1994) Smooth recollection of a pattern sequence by nonmonotone analog neural networks. *Proceedings of the 1994 IEEE International Conference on Neural Networks* 2:1032–37. [MM]

Muller, G. E. & Pilzecker, A. (1900) Experimentelle Beiträge zur Lehre vom Gedächtniss. *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane, Erganzungsband* 1:1–288. [PMM]

Niki, H. (1974) Prefrontal unit activity during delay alternation in the monkey. *Brain Research* 68:185. [aDJA]

O'Keefe, J. & Speakman, A. (1987) Single unit activity in the rat hippocampus during a spatial memory task. *Experimental Brain Research* 68:1. [aDJA]

Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. (1980) Geometry from a time series. *Physical Review Letters* 45:712. [MWH]

Palm, G. (1982) *Neural Assemblies.* Springer-Verlag. [MH, FP, JPR]

Petitot, J. (1989) Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics* 15(1/2):25–71. [JP]

(1991) Why connectionism is such a good thing: A criticism of Fodor's and Pylyshyn's criticism of Smolensky. *Philosophica* 47:1:49–79. [JP]

(1995) Morphodynamics and attractor syntax: Dynamical and morphological models for constituency in visual perception and cognitive grammar. In: *Mind as Motion,* ed. T. van Gelder & R. Port. MIT Press. [JP]

Pinker, S. & Prince, A. (1988a) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1–2):73–193. [MH]

(1988b) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. In: *Connections and symbols,* ed. S. Pinker & J. Mehler. MIT Press. [FVDV]

Plato (360 BC) *Theaetetus* [transl. B. Jowett]. Available electronically on the Internet at URL gopher://gopher.vt.edu:10010/02/131/23. [SE]

Pulvermüller, F., Lutzenberger, W. & Birbaumer, N. (in press) Electrocortical distinction of vocabulary types. *Electroencephalography and Clinical Neurophysiology.* 94:357–70. [FP]

Pulvermüller, F., Preissl, H., Eulitz, C., Pantev, C., Lutzenberger, W., Elbert, T. & Birbaumer, N. (1994) Brain rhythms, cell assemblies, and cognition: Evidence from the processing of words and pseudowords. *Psycoloquy* 5(48). [FP]

Putnam, H. (1988) *Representation and reality.* MIT Press. [SE]

Pylyshyn, Z. W. (1980) Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3:111. [JPR]

Quine, W. V. O. (1960) *Word and object.* MIT Press. [SE]

Quinlan, P. (1991) *Connectionism and psychology. A psychological perspective on new connectionist research.* Wheatsheaf, New York: Harvester. [AL]

Quintana, J., Fuster, J. M. & Yajeya, J. (1989) Effects of cooling parietal cortex on prefrontal units in delay tasks. *Brain Research* 503:100–10. [JMF]

Quintana, J., Yajeya, J. & Fuster, J. M. (1988) Prefrontal representation of stimulus attributes during delay tasks: I. Unit activity in cross-temporal integration of sensory and sensory-motor information. *Brain Research* 474:211–21. [JMF]

Raijmakers, M. E. J., Koten, S. & Molenaar, P. C. M. (in press) On the validity of simulating stagewise development by means of PDP-networks: Application of catastrophe analysis and an experimental test of rule-like network performance. *Cognitive Science.* [MEJR]

Rauschecker, J. P. (1991) Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. *Physiological Reviews* 71:587–615. [JPR]

(1995) Compensatory plasticity and sensory substitution in the cerebral cortex. *Trends in Neurosciences* 18:36–43. [JPR]

Rauschecker, J. P. & Hahn, S. (1987) Ketamine-xylazine anaesthesia blocks consolidation of ocular dominance columns in kitten visual cortex. *Nature* 326:183–85. [JPR]

Rauschecker, J. P. & Korte, M. (1993) Auditory compensation for early blindness in cat cerebral cortex. *Journal of Neuroscience* 13:4538–48. [JPR]

Rauschecker, J. P. & Sejnowski, T. (1994) Processing of visual and auditory space and its modification by experience. In: *Advances in Neural Information Processing Systems,* vol. 6, ed. J. D. Cowan, G. Tesauro & J. Alspector. [JPR]

Rauschecker, J. P., Tian, B., Korte, M. & Egert, U. (1992) Crossmodal changes in the somatosensory vibrissa/barrel system of visually deprived animals, *Proceedings of the National Academy of Sciences USA* 89:5063–67. [JPR]

Reese, H. W. (1989) Rules and rule-governance: Cognitive and behavioristic views. In: *Rule-governed behavior: Cognition, contingencies, and instructional control,* ed. S. C. Hayes. Plenum. [MEJR]

Reichenbach, H. (1956) *The direction of time.* University of California Press. [DCK]

Reilly, R. G. & Sharkey, N. E. (1992) Representational adequacy. In: *Connectionist approaches to natural language processing,* ed. G. Reilly & N. E. Sharkey. Erlbaum. [FVDV]

Renals, S. & Rohwer, R. (1990) A study of network dynamics. *Journal of Statistical Physics* 58:825–48,

Rochester, N., Holland, J. H., Haibt, L. H. & Duda, W. L. (1956) Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on Information Theory* IT-2:80–93. [AL, PMM]

Ruppin, E. (1995) Neural modeling for psychiatric disorders. *Network* 6(3). [rDJA]

Sakai, K. & Miyashita, Y. (1991) Neural organization for long-term memory of paired associates. *Nature* 354:152–55. [aDJA, EA, GJD, MH, PMM, MEJR]

Sakai, K., Naya, Y. & Miyashita, Y. (1994) Neuronal tuning and associative mechanisms in form representation. *Learning and Memory* 1:83–105. [SE]

Sayer, R. J., Redman, S. J. & Andersen, P. (1989) Amplitude fluctuations in small EPSPs recorded from CA1 pyramidal cells in the guinea pig hippocampal slice. *Journal of Neuroscience* 9:840. [aDJA]

Schade, A. F. & Bitterman, M. (1966) Improvement in habit reversal as related to the dimensional set. *Journal of Comparative and Physiological Psychology* 62:43–48. [MEJR]

Sejnowski, T. J. (1977) Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology* 4:303–21. [JPR]

Shiori, S. & Cavanagh, P. (1989) Saccadic suppression of low-level motion. *Vision Research* 29:915–28. [DCB]

Sholl, D. A. (1956) *The organization of the cerebral cortex.* Wiley. [WJF]

Simon, H. A. & Kaplan, C. A. (1989) Foundations of cognitive science. In: *Foundations of cognitive science,* ed. M. I., Posner. MIT Press. [FVDV]

Singer, W. (1990) The formation of cooperative cell assemblies in the visual cortex. *Journal of Experimental Biology* 153:177–97. [JPR]

(1994) Putative functions of temporal correlations in neocortical processing. In: *Large scale neuronal theories of the brain,* ed. C. Koch & J. Davis. MIT Press. [FP]

Singer, W. & Gray, C. M. (1995) Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* 18:555–86.

Skarda, C. A. & Freeman, W. J. (1987) How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* 10:161–95. [WJF, MWH]

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [JP]

Sompolinsky, H. (1986a) The theory of neural networks: The Hebb rule and beyond. In: *Heidelberg Colloquium on glassy dynamics,* ed. L. van Hemmen & I. Morgenstern. Springer-Verlag. [aDJA]

(1986b) Neural networks with nonlinear synapses and static noise. *Physics Review* A34:2571. [aDJA]

Sompolinsky, H., Crisanti, A., & Sommers, H.-J. (1988) Chaos in random neural networks. *Physics Review Lett.* 61:259–62 [JP]

Spence, K. W. (1936) The nature of discrimination learning in animals. *Psychological Review* 43:427–49. [MEJR]

Squire, L. (1987) *Memory and brain.* Oxford University Press. [JPR]

Stevens, C. F., Tonegawa, S. & Wang, Y. (1994) The role of calcium-calmodulin kinase II in three forms of synaptic plasticity. *Current Biology* 4:687–93. [JPR]

Takens, F. (1981) Detecting strange attractors in turbulence. In: *Lecture notes in mathematics 898,* ed. D. Rand & L. S. Young. Springer-Verlag. [MWH]

Tanaka, K. (1992) Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology* 2:502–5. [arDJA, SE]

(1993) Neuronal mechanisms of object recognition. *Science* 262:685–88. [FVDV]

Tanzi, E. (1893) I fatti e le induzioni nell'odierna istologia del sistema nervosa. *Rivista Sperimentale di Freniatria* 19:419–72. [GJD]

Thom, R. (1980) *Modèle mathématiques de la Morphogenèse.* Paris, Christian Bourgois. [JP]

Tighe, T. J. (1964) Reversal and nonreversal shifts in monkeys. *Journal of Comparative and Physiological Psychology* 58(2):324–26. [MEJR]

Tsodyks, M. V. & Feigel'man, M. V. (1988) The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters* 46:101. [aDJA]

Tulving, E. (1984) Précis of *Elements of episodic memory. Behavioral and Brain Sciences* 7:223–68. [WK]

Uchikawa, K. & Sato, M. (in press) Saccadic suppression to achromatic and chromatic responses measured by increment-threshold spectral sensitivity. *Journal of the Optical Society of America A.* [DCB]

Van der Velde, F. (1994) *Integrating connectionism and symbol manipulation: The importance of implementation in psychology.* Technical report, Leiden University. [FVDV]

(in press) Symbol manipulation with neural networks: Production of a context-free language using a modifiable working memory. *Connection Science.* [FVDV]

Virasoro, A. M. (1988) Categorization in neural networks and prosopagnosia. *Physics Reports* 184:99. [aDJA]

Von der Malsburg, C. (1973) Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 14:85–100. [JPR]

Von Neumann, J. (1954) *The computer and the brain.* Yale University Press. [aDJA]

Wickens, J., Hyland, B. & Anson, G. (1994) Cortical cell assemblies: A possible mechanism for motor programs. *Journal of Motor Behavior* 26:66–82. [FP]

Willshaw, D., Buneman, O. P. & Longuet-Higgins, H. (1969) Nonholographic associative memory. *Nature* 222:960. [rDJA]

Wilson, M. A. & McNaughton, B. L. (1993) Dynamics of the hippocampal ensemble code for space [see comments]. *Science* 261:1055–58. [EA]

Wright, J. J. & Liley, D. T. J. (1995) Simulation of electrocortical waves. *Biological Cybernetics* 72(4):347–56. [JJW]

Zeeman, E. C. (1962) The topology of the brain and visual perception. In: *Topology of 3-Manifolds and Related Topics,* ed. M. K. Fort Jr. Prentice Hall. [MWH]

Zeeman, E. C. (1965) Topology of the brain. *Mathematics and computer science in biology and medicine.* Medical Research Council. [JP]

Zeeman, E. C. (1976) Brain modelling. In *Structural stability, the theory of catastrophes and applications in the sciences,* Lecture notes in mathematics 525:367–372. Berlin, Springer. [JP]

Zipser, D., Kehoe, B., Littlewort, G. & Fuster, J. (1993) A spiking network model of short-term active memory. *Journal of Neuroscience* 13:3406–3420. [JMF, rDJA]