

Part I. Physiological Basis of Learning and Memory Storage

Some Properties of a Neural Model for Memory

(with J. A. Anderson, M. M. Nass, W. Freiberger, and U. Grenander)

AAAS Symposium, December 30, 1972

I had always been interested in biological problems. My project for the Westinghouse Science Talent Search, of which I was a finalist in 1947, concerned the development of strains of the bacterium *bacillus subtilis* more resistant to penicillin than the wild type. This choice of project was certainly influenced by the availability of a wonderful biology laboratory at the Bronx High School of Science, where I spent every possible afternoon hour when classes had finished.

My decision to choose physics was influenced by several factors. At the end of World War II, physics was very much in the public eye. The explosion of the first nuclear bomb had created a sensation. And for years I had had a burning desire to understand the theory of relativity and quantum theory, to understand the deep mysteries of the universe. And, as for many aspiring young scientists, my great idol was Albert Einstein. Einstein with his blown hair and mysterious smile; how much one wanted to understand the deep thoughts of that saint-like figure.

In my early years at Columbia College, when the decision had to be made, I chose physics. Only this way, I thought to myself, would I ever be able to participate in the deepest of all mysteries: the fundamental laws of nature, the meaning of space, time, relativity, and quantum particles. If I didn't learn early, I would never really understand. The rest, I hoped, I could somehow do later. I recall John Ward's shock at the Institute for Advanced Study in Princeton when I confided to him that after I had finished all of the problems in physics I would return to biology (having written a few operas on the way).

Thus, in the late sixties when papers on superconductivity were becoming longer, more technical, and somehow less exciting, in my peripatetic search for new objects of interest, I returned to my old love. In particular, I was fascinated by statements in various articles and books, paraphrased roughly: "Although much is known about the structure and function of individual neurons, almost nothing is known about how memory is stored and retrieved". This seemed to me particularly anomalous since, for example, so much is known about how memory is stored in computers. When I began to think about the problem, it occurred to me that memory might somehow be a collective property of very large ensembles of neurons. I had already had some success teasing subtle collective properties from large ensembles of interacting electrons and had been working on aspects of many-body theory for almost a decade, and it seemed possible that perhaps there might be some technical connection. This proved illusory, but it got me going.

So, when a young graduate student, Menasche Nass, appeared in my office and said that he was interested in working on a problem with me for his Ph. D., somehow (I don't remember exactly how) the possibility of working on a biological problem arose. I warned Menasche

that among the normal risks one takes, this was really a high-risk thing to do. However, after some thought, Menasche decided he wanted to do it. It turned out he was an excellent student who accomplished a great deal in his Ph. D. thesis, unfortunately, some half a generation ahead of his time. The career options when he finished were not brilliant in theory, and Menasche decided he didn't want to be an experimentalist. He has since pursued a successful career as a tax lawyer in Los Angeles.

As I recall, I had come upon a paper by Longuet-Higgins proposing a model for memory based on a hologram analogy. The essential notion was that rather than storing an entire item of memory in a single spot, the memory would be distributed over a region. And, as we and everyone else was saying even at that time, this would be contained somehow in the synaptic junctions between neurons.¹

So, my first assignment to Menasche (a summer project) was to try to devise a more physiological rendering of the type of holographic memory that Longuet-Higgins had suggested. This was toward the beginning of the summer of '70. In grand graduate student style, Menasche told me, when I encountered him again in September, that he hadn't solved the problem but he had found in the literature, an extremely interesting proposal. He then showed me a paper of James Anderson, at that time a Research Associate at the Rockefeller University in New York. In fact Anderson's proposal seemed highly attractive. It was a physiologically possible distributed memory that did seem as though it could be acquired in an actual nervous system.

I contacted Jim and there began a collaboration that lasted for many years. Jim, by the way, is now Chairman of the Department of Cognitive and Linguistic Sciences at Brown University.

The following is an excerpt from the first article we wrote together. It provides a summary of some of our very early thinking.

¹Synaptic junctions arose in the course of evolution as part of the solution to the problem of establishing communication between one portion of an animal and another. Once excitable membranes became available, their use in various cells such as muscles and neurons (especially in neurons) provided a means of electrical communication. When the animal becomes large one can, of course, string a single neuron from one end to the other; but this is impractical in most situations and probably risky since if the neuron is severed it is likely more difficult to replace a single long cell than a smaller, shorter one. Thus the problem of communicating between neurons arises. A straightforward means would presumably be to have a direct electrical contact and this does, in fact, occur. A seemingly less straightforward means is to transmit the information from one neuron to another by the very complex mechanism of chemical transmission: a chemical transmitter released unto the synaptic cleft, diffuses and attaches itself to special receptors on post-synaptic membrane, opens channels and produces currents in the post-synaptic dendrite. These propagate passively to the cell body of the post-synaptic neuron and, if these are sufficient, produce action potentials which travel along the axon of the post-synaptic neuron. Thus the information flow continues.

One might speculate that a great advantage of chemical transmission between neurons is the relative ease for modification of the transmission efficacy. Thus the same action potential in the pre-synaptic axon can produce a different current or response in the post-synaptic dendrite. And it is this possibility that produces the dramatic new capability of information storage.

SOME PROPERTIES OF A NEURAL MODEL FOR MEMORY

James A. Anderson

The Rockefeller University, New York, NY 10021, USA

Leon Cooper and Menasche M. Nass

*Department of Physics, Brown University,
Providence, RI 02912, USA*

Walter Freiberger and Ulf Grenander

*Division of Applied Mathematics, Brown University,
Providence, RI 02912, USA*

2:00 pm December 30, 1972

Washington Hilton, Hemisphere Room

AAAS Symposium, Theoretical Biology and Biomathematics

ABSTRACT

A model of long-term memory, motivated by the anatomy and physiology of the mammalian central nervous system is proposed. We suggest that what is of importance to the nervous system is the collective individual activities of large numbers of individual neurons and that just such a collection of activities constitutes the memory trace. We assume that the long-term memory is the sum of such individual traces. With a minimum of some of the functions of a biological memory and that many of its properties are reminiscent of the brain. Throughout the emphasis is on consistency with the known physiology.

Enough is known about the mammalian central nervous system to allow us to suggest some general principles that a neural model for memory must satisfy.

First, the biological memory system is inherently noisy in that what is recalled is rarely (if ever) identical to what is stored and the input to the system is almost never identical to what has been stored. Also, we know that nervous activity often displays the character of noise-like processes: electrical activity, particularly in higher centers such as cortex, shows Gaussian or Poisson distributions of amplitudes or spike activity.

Second, the physical representation of memory appears to be distributed over the brain, parts of a single 'memory' presumably occurring in many different spatial locations. This property has recently been called the 'holographic' property because, by analogy to a hologram, information is not stored locally (point-to-point) as in an ordinary photograph, but globally. This distributed nature of the memory has been generally recognized since Lashley conducted his famous ablation experiments on rat cortex.

Third, there is no evidence that at any time is part of the brain left 'vacant' for future storage. Rather, most parts of the brain show continuous activity. Nor does the amount of information a human brain can store show saturation.

Fourth, there appears to be no strong evidence in the higher centers of the mammalian central nervous system for the existence of 'pontifical' or 'decision making' neurons, although such neurons may be present in some invertebrates. Highly parallel, numerous sensory afferents excite simultaneously large populations of neurons in cortex and thalamus. Action in the mammalian central nervous system appears to involve processing by large numbers of basically rather similar neurons and the critical steps in neural processing do not appear to depend on one or only a few neurons out of this population.

Fifth, areas capable of memory storage are written over, again and again. This conclusion is strongly suggested by our previous points and is supported by the neurophysiology of cortex. To quote Sir John Eccles (1971), "Any cortical neuron does not exclusively belong to one engram ('memory') but, on the contrary, each neuron and even each synaptic junction would be built into many engrams."

Sixth, there are in the human central nervous system approximately 10^{10} neurons with at least 10^3 connections apiece. Can the details of these connections be important? Specifying, for example, 10^{13} connections at birth would involve the storage of a vast amount of genetic information. It is not likely that the amount of DNA present in the chromosomes could specify so many connections. This suggests that the central nervous system, although globally organized, probably is locally random. Studies by Creutzfeldt and Ito (1968) suggest that neurons in visual cortex receive most of their input from a small number of locally randomly assorted fibers from the lateral geniculate body. Other data suggests the same kind of conclusion.

Seventh, processing in the central nervous system is highly parallel. The visual system, for example, has over a million parallel input channels feeding into cerebral cortex. Other areas of cortex show similar organization and cortical interconnections are dense and highly parallel.

All of the above considerations have been used in the past, in whole or in part, in the construction of various models of nervous system function. Some models have proceeded by analogy with the digital computer. Both the brain and the computer carry on 'processing' in some sense. We believe, however, that the analogy ends there. The modern computer is a very fast digital machine capable of performing

serially comparatively simple operations. Faced with a task of biological significance, analyzing a complex or noisy pattern, for example, the computer is a very poor second. Further, considerable evidence now indicates that most neurons in mammals do not behave in binary fashion. Apparently what is important to the nervous system is not the simple presence or absence of a spike, but the average firing frequency of the cell over a brief interval. Perkel and Bullock (1968) list various codes employed in nervous systems. Most of these, and the ones most often found, depend on temporal patterns of spike long in comparison to the duration of a single spike.

Where the computer analogy breaks down most severely is in the storage and retrieval of information. We do not wish to discuss these aspects of computers, other than to say that information is stored locally and retrieval involves looking at a specified memory location. As we shall see, our model stores information globally and is content addressable.

The problem we will now discuss is that of the storage of memory traces (engrams) which we will define as large patterns of individual neuron activities which tend to act as units in operations of the system. We will argue that a simple rule of synaptic plasticity is sufficient for the storage of information and is in fact an optimum way to do so. (Anderson, 1972, p. 203) The idea that synaptic change with use could serve as a mechanism for memory goes back at least as far as the nineteenth century. (Tanzi, 1893). More recently, Eccles has reviewed the evidence for synaptic change with use and considered several specific mechanisms. A great deal of experimental effort has gone into this research with as yet no conclusive results. Various mechanisms have been proposed, and many of these are compatible with our model.

In line with the evidence presented, let us proceed to construct an idealized system and to make certain reasonable assumptions.

1. We consider a system, common in cortex, where one large group of neurons, α , projects to another large group of neurons, β . α and β do not have to be distinct systems. Often a group of neurons projects to itself via recurrent collaterals having a long conduction time. Examples of such projection systems are, the projections of thalamic nuclei to cortex, and the intracortical projections.

2. The trace (or engram of memory) is the simultaneous activities shown by a large group of neurons.

3. Synaptic interactions add linearly.

4. Synaptic weights are coded so that change in synaptic weight is proportional to the product of pre- and post-synaptic activities at a given time.

With these assumptions we shall see that the system is capable of behavior suggestive of a biological memory. Our system can, among other things,

1. Recognize a previously presented (and incorporated) trace.

2. Store associations in the sense that if a trace f is associated with another trace g by making the proper synaptic adjustments according to our rules, then presentation of f gives rise to g plus noise.

In order to have a concrete physiological system in mind we imagine our model to be of cerebral cortex and we identify a trace as the simultaneous pattern of individual activities of cortical cells. As is well known each neuron has a resting rate of firing and upon stimulation this rate can change. At any time then, we can represent the state of a neuron by a number representing the level of its activity. For reasons that we discuss later, we can just as well represent the state of the cell by the algebraic value of the neuron's instantaneous firing rate above or below spontaneous rate. (Thus if a neuron unstimulated has a firing rate of 17 spikes per second, and we observe it firing at 13 spikes per second, we would represent its state as -3). If we have N neurons, we can represent the state of the system by making N entries into a column; in other words, by a vector whose components are the states of the N neurons. We identify the vector with the trace. An input can then be characterized in terms of its effect on the neurons of cerebral cortex by the vector $|f\rangle$, where we have borrowed the Dirac Bra-Ket notation from physics. (In this notation, a vector \mathbf{f} is written as $|f\rangle$ and the inner product of the two vectors \mathbf{f} and \mathbf{g} as $\langle f|g\rangle$). The most useful property of this notation derives from the fact that if we have a complete set of vectors, say three non-coplanar vectors in three dimensional space, we can write $\sum |f\rangle\langle f| = 1$ with proper ORTHO orthonormalization, where the sum is over vectors of the complete set.)

Some of the assumptions we have made require further justification. For one, we have assumed a linear system for the response characteristics of the neurons to stimuli. Certainly we do not expect strictly linear behavior in so complex a system. From a practical point of view, assumption of linearity allows immense simplification of the mathematics. Fortunately, assumption of linearity is quite close to reality as far as the central nervous system is concerned provided we define precisely what system characteristics we are interested in. Mountcastle (1967) has proposed as a general rule that there is a linear relation between the output of first order afferent fibers and the sensory response of the nervous system. He has considerable data on the tactile responses of monkeys indicating preservation of linear transduction of first order afferent output up to units in cortex. In the visual system, Maffei *et al.* (1967) have shown, the firing rate of lateral geniculate body cells follow in a linear fashion the sinusoidal intensity modulation of a light stimulus. Further, Maffei (1968) has shown that LGB cells apparently use spatial averaging in order to preserve linearity of cell response. As in many physical systems, the linearity assumption has a strict domain of validity for small changes in stimulus intensity. Neurons, though having highly nonlinear portions of their individual responses, may respond, on the average in group, in quite linear fashion.

We have implicitly assumed that the behaviour of neurons is a direct reflection of the stimuli applied. In connection with this we would like to mention a very significant set of experiments conducted by Hirsh and Spinelli (1971). Hirsh and Spinelli raised kittens in an environment where they received as their sole visual

input three horizontal stripes to one eye and three vertical stripes to the other. They then studied the organization of receptive fields in visual cortex and found, first, that a given cell in visual cortex was driven by one eye or the other, but not both, in contrast to normal cats where around 80% of cells are binocular. Second, the elongated receptive fields of the visual cortical cells now conform to the direction of the input driving them, that is, receptive field orientations are horizontal or vertical depending on whether they were connected to the eye receiving horizontal or vertical input. These experiment provide clear evidence that cell activity may mirror stimulus form in a very simple way. Let us try to reflect this in our model.

In order to proceed we must make some simplifying assumptions in order to facilitate calculations. The inner product of a trace is taken to be the ‘power’ of the trace. In some intuitive sense, the power stands for the strength of the trace and for convenience we assume all traces have equal power. We also assume all traces have a mean value of zero.

We can now state the central assumption of our model. To form the memory for a group of traces, we simply form the vector sum of all the traces of the group. Thus, if there are K traces to be stored, then we form the memory vector $|s\rangle$ as

$$|s\rangle = \sum_{k=1}^k |f^k\rangle.$$

We further assume that all the traces are statistically independent and the statistics of all elements in the vectors are the same. (That is, neurons are similar statistically to each other.)

As an aside, Noda and Adey (1970) found that when two cells in parietal cortex were recorded simultaneously with the same microelectrode, the two cells although physically very close together were not correlated in their discharge when the cat was awake or in REM sleep. Thus adjacent cells tended to act as ‘individuals’, as if each cell sampled the environment independently.

By the central limit theorem, the sum of many uncorrelated traces should closely approximate a vector whose elements are the values taken by a normally distributed random variable. (Thus the noise-like atmosphere of the nervous system may be a simple consequence of this kind of organization.) Since we know nothing of the details of the traces in the memory in general we will calculate average values over many sets of allowable traces. In order to get a quantitative measure of how good the system is, we will ask how close is the reconstructed trace to the desired trace. In the language of the electrical engineer we will ask for the signal to noise ratio.

ACKNOWLEDGEMENTS

We have been greatly assisted by many people both at Brown University and at the Rockefeller University. We would like to thank and acknowledge the assistance

of Professor C. Elbaum and Professor H. Kucera at Brown University and Bruce W. Knight at the Rockefeller University.

REFERENCES

- [1] J. A. Anderson, *Math. Biosci.* **8**, 137 (1970).
- [2] J. A. Anderson, *Math. Biosci.* **14**, 197 (1972).
- [3] O. D. Creutzfeldt and M. Ito, *Exp. Brain Res.* **6**, 324 (1968).
- [4] J. C. Eccles, in *Brain and Human Behavior*, edited by A. G. Karczmar and J. C. Eccles, Springer, Berlin (1972).
- [5] H. V. B. Hirsh and D. N. Spinelli, *Exp. Brain Res.* **13**, 509 (1971).
- [6] L. Maffei and G. Rizzolatti, *J. Physiol.* **195**, 215 (1968).
- [7] V. B. Mountcastle, in *The Neurosciences*, edited by G. C. Quarton, T. Melnechuk, and F. O. Schmitt, Rockefeller University Press, New York (1967).
- [8] H. Noda and W. R. Adey, *J. Neurophysiol.* **33**, 572 (1970).
- [9] D. H. Perkel and T. H. Bullock, *Neurosciences Research Program Bulletin* **6**, 221 (1968).
- [10] E. Tanzi, *Riv. Sep. Trensia* **19**, 149 (1893).
- [11] W. Wickelgren, Multitrace Strength Theory, in *Models of Human Memory*, edited by D. A. Norman, Academic Press, New York (1972).